Oct1 recruits the histone lysine demethylase Utx to canalize lineage specification

Jelena Perovanovic, Zuolian Shen, Yifan Wu, Dean Tantin*

Department of Pathology and Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112 U.S.A.

*Corresponding author: dean.tantin@path.utah.edu

Running title: Oct1 mediates gene bivalency resolution

[*Keywords:* early development; embryonic stem cells; single-cell RNA sequencing; Oct1/Pou2f1; Utx]

Characters in title (including spaces): 84

Words in abstract: 170

Character count: 62,848

Total main figures and tables: 6

Abstract

The pathways used to transition from an undifferentiated, pluripotent state into cell type-specific states are incompletely understood. Here we show that the transcription factor Oct1/Pou2f1 allows for correct induction of silent, developmental lineage-specific genes to "canalize" developmental progression. Using inducible-conditional *Oct1* knockout embryonic stem cells and single-cell gene expression profiling, we show that mesodermal differentiation is impaired in the Oct1 deficient condition. Oct1 deficient cells fail to form late paraxial mesoderm and early somite stage populations, and show "leaky" developmental trajectories with inappropriate lineage branching and accumulation of poorly differentiated cells that retain gene expression and metabolic hallmarks of pluripotency. Oct1 directly binds and regulates genes critical for developmental regulation, including genes encoding mesoderm-specific master regulators and components of chromatin regulatory complexes. Oct1 positively resolves gene bivalency by recruiting Utx to mesoderm-specific, bivalent/poised targets, thereby removing inhibitory H3K27me3 chromatin marks to activate their expression. Oct1 deficient cells fail to remove H3K27me3 and positively resolve bivalency. Ectopic Oct1 expression improves the ability of cells to differentiate accurately under mesoderm lineage-inducing conditions.

Introduction

Lineage specification is a key process in the development of multicellular organisms by which undifferentiated cells progressively acquire tissue- and cell type-specific features (Seydoux and Braun 2006). It is dynamically regulated, requiring extensive transcriptional and epigenetic remodeling to selectively activate lineage-appropriate gene expression programs while stably repressing the expression of genes specific for alternative lineages. Embryonic stem cells (ESCs) represent a pluripotent cell type capable of both self-renewal and differentiation into all three germ layers (Beddington and Robertson 1989).The three germ layers are established during gastrulation – a massive spatial reorganization of the embryo from a single-layer epiblast into a multilayered organism. One of the germ layers – mesoderm (MD) – gives rise to dermomyotome (muscle and dermis), sclerotome (axial skeleton) and lateral MD (cardiac) among other tissue types.

The dynamic transcriptional changes underlying lineage specification require extensive chromatin remodeling and spatiotemporal activation of master transcription factors. Remodeling is potentiated by a unique chromatin landscape in pluripotent cells. Chromatin in ESCs is largely accessible and lacks stable heterochromatin domains (Meshorer and Misteli 2006; Schlesinger and Meshorer 2019). A large number of genes encoding lineage-specific developmental regulators are marked at promoters and gene bodies by nucleosomes that simultaneously convey activating (H3K4me3) and repressing (H3K27me3) potential (Bernstein et al. 2006; Ku et al. 2008). In ESCs, these “bivalent” genes are silent or expressed at low levels, but poised for activation. During development, gene bivalency resolves via either removal of activating marks and gene silencing or removal of repressive marks and gene activation. Which bivalent genes resolve in which direction is lineage-specific, resulting in distinct, cell fate-appropriate transcription programs and durable repression of lineage-inappropriate genes. The programs that instruct correct bivalency resolution are incompletely understood.

POU transcription factors play varied and central roles in the regulation of development (Tantin 2013). The POU factor Oct4 (encoded by *Pou5f1*) is a master regulator of the induction and maintenance of pluripotency (Takahashi and Yamanaka 2006; Palmieri et al. 1994; Nichols et al. 1998). Oct4 associates with bivalent genes in ESCs (Bernstein et al. 2006), but is rapidly silenced in their differentiated progeny before bivalency resolution and the induction of tissue- and cell type-specific gene expression. A second POU protein, Oct1, is co-expressed with Oct4, but unlike Oct4 is expressed beyond pluripotency (Shen et al. 2017). Oct1 is widely expressed and required for placental and embryonic development (Wang et al. 2004; Sebastiano et al.

2010). Circumventing the placental defects via tetraploid complementation results in developmental arrest at E8.25 with no more than five somites (Sebastiano et al. 2010). Oct1-deficient ESCs are morphologically normal and show appropriate self-renewal and growth characteristics, but upon differentiation have multiple phenotypic and gene expression defects. These include decreases in developmental lineage-specific gene expression coupled with elevated expression of developmentally incorrect genes (Shen et al. 2017). The mechanism however is not known.

Here, using MD differentiation of *Oct1*/*Pou2f1* inducible-conditional ESCs, we show that Oct1 “canalizes” lineage specification by directly de-repressing lineage-specific bivalent genes. Oct1 is needed for consequent early MD development and adoption of terminal MD phenotypes. In the absence of Oct1, differentiating cells show “fuzzy” differentiation programs and increased predilection to proceed down incorrect developmental trajectories. We show that Oct1 interacts with components of the UTX H3K27me3 demethylase complex, and recruits UTX to developmentally appropriate target genes to facilitate bivalency resolution. Binding sites for Oct1 and Oct4 in pluripotent cells in which Oct1 binding is carried forward during MD differentiation are also enriched in binding sites for Smad transcription factors, key mediators of MD specification. Oct1 and Smad3 interact in differentiating cells, providing a means of restricting Oct1’s bivalency-resolving potential. Finally, we show that ectopic Oct1 expression increases MD developmental specification, suggesting a possible application in improving the specificity of developmental outcomes in differentiating ESCs or iPSCs.

Results

Loss of Oct1 Causes Aberrant Mesodermal Differentiation

To identify how Oct1 loss perturbs cells during differentiation, we performed a single-cell RNA-seq (scRNA-seq) analysis using an MD differentiation protocol that results in terminally-differentiated myotubes after 3 weeks of culture. The differentiation process is initiated by Bmp4 treatment to activate TGFβ signaling, and generates dynamic changes in metabolic programs (Oginuma et al. 2017), epithelial-to-mesenchymal transition (EMT) (Diaz-Cuadros et al. 2020) and induction of MD-specific gene expression programs during the differentiation timecourse (Chal et al. 2015). Cells were collected at day 0 (D0, pluripotent ESCs), D3 and D6 of differentiation to identify changes in populations and gene expression patterns early in the differentiation timecourse. Data from the different conditions were subjected to graph-based clustering approach (Macosko et al. 2015) and uniform manifold approximation and projection (UMAP) to visualize clusters of cells with similar gene expression. Integrated analysis of Oct1-sufficient *Pou2f1*$^{fl/fl}$;Rosa26-CreER^{T2};LSL-YFP conditional ESCs (hereafter, "parental" cells) at the three timepoints identifies a range of cell populations, from ESC-like to paraxial MD to early somite (Fig. 1A). These populations recapitulate developmental stages associated with induction of early MD markers (*Fgf17* and *T*/*Tbxt*/*Brachyury*; cluster 2 and 7), paraxial MD (*Meox1*, *Twist1*, *Pax3*) and early somite (*Cxcr4*) (Fig. 1B, Supplemental Fig. S1). These results provide a single-cell resolution map of the earliest stages of MD differentiation.

Next, we compared the parental controls to tamoxifen-treated, Oct1-deficient cells (hereafter, "cKO") to identify changes in cell populations and gene expression (Table S1). cKO cells have poor mesodermal differentiation capacity with reductions in key mesodermal clusters, in particular at differentiation D6. For example, parental cell populations characterized by paraxial MD gene expression represent the largest two clusters, accounting for 14.7 and 12.9% of cells, respectively, while a population associated with somites accounted for 5.6% of cells (Fig. 1C left panel). In contrast, cKO D6 cells showed increases in cells that retain epithelial characteristics and decreased percentages of appropriately differentiated paraxial MD (2.3% and 7.5%), and somites (0.5%, Fig. 1C, right panel). Comparing gene expression between parental and cKO paraxial mesoderm clusters, cKO cells fail to appropriately induce lineage-appropriate genes such as *Pax3* and *Pax7*, and inappropriately upregulate lineage-inappropriate markers such as the epithelial-specific genes *Krt8* and *Krt18* (Fig. 1D).

"Pseudotime" analysis of scRNA-seq data allows multiple differentiation timepoints to be overlaid with defined starting (ESC) and endpoint (somite) populations. Parental control cells

progress through a largely linear pathway with one minor branch retaining an inappropriate epithelial fate (Fig. 2A, top panel). In contrast, differentiating cKO ESCs show a much larger proportion of cells inappropriately branching and proceeding further down an alternative developmental trajectory (bottom panel), consistent with the diminished developmental progression of paraxial MD to the somite stage, and consistent with enrichment of cells that inappropriately maintain an epithelial state (Fig. 1C, Supplemental Fig. S1). We also examined pseudotemporal gene expression using specific genes associated with pluripotency and MD development. In this analysis, the position of each cell in pseudotime is shown on the X-axis and gene expression on the Y-axis (Fig. 2B). Parental cells show robust early expression of genes associated with pluripotency such *as Klf4* that lose expression roughly mid-way through pseudotime. *T* is transiently induced, while the early somite markers *Pax7* and *Cxcr4* are efficiently and coordinately induced later (Fig. 2B, top panel). In contrast, cKO cells exhibit inappropriately prolonged expression of *Klf4* and *Tbxt*, and largely fail to upregulate *Pax7* and *Cxcr4* (bottom panel). We used pseudotemporal ordering to visualize expression of the 2000 most variable (dynamic) genes in parental cells, and compared expression of the same genes in the same order in cKO cells (Fig. 2C). This analysis revealed clusters associated with pluripotency (*Klf4*, *Nanog*), early MD differentiation (*T*), paraxial mesoderm (*Rspo3, Meox1, Mesp1*), and early somites (*Dll1, Pax3*). Although the pseudotime gene expression pattern for pluripotency-associated genes was largely normal in cKO cells, other clusters revealed “fuzzy” patterns of gene expression early in differentiation, with prolonged expression of genes such as *Tbxt* and poor/blurred expression of genes in the clusters associated with paraxial mesoderm. In addition, the induction of somitic genes normally expressed later in development such as *Dll1*, *Pax3* and *Pax7* was poor in cKO cells (Fig. 2C).

“Velocity” analysis allows developmental directionality to be inferred based on the ratio of unspliced, newly-synthesized pre-mRNA to mature spliced mRNA. A vector is then assigned to each cell that indicates developmental direction and magnitude relative to the other cells. We found that parental cells at D6 of differentiation form discrete clusters such as anterior MD and somites (Fig. 2D, marked by long arrows). cKO cells by contrast progress primarily towards a poorly differentiated state characterized by multiple lineage markers. Moreover, the cells show stationary profiles indicative of failed differentiation potential (Fig. 2D). Cumulatively, the data demonstrate that Oct1 is required for differentiation progression and appropriate transcription program activation, with Oct1 deficient cells unable to canalize appropriate lineage specification programs.

Oct1 Occupies Developmental Genes in ESCs

One model that explains the above data is that Oct1 occupies developmental-specific targets occupied by Oct4 in ESCs, to mediate induction of developmentally appropriate genes and repression of genes specific to alternative lineages. To test this hypothesis, we performed Oct1 ChIP-seq. Prior work showed that Oct1 occupancy of Oct4 targets increases during retinoic acid (RA)-mediated differentiation as Oct4 is lost (Shen et al. 2017). RA ultimately differentiates cells into neuroectodermal lineages (Bain et al. 1995). In undifferentiated ESCs, strong Oct1 binding (with Oct4) was only observed in a group of ~100 genes containing Oct protein variant binding sites termed MOREs (Shen et al. 2017). We used a different Oct1 antibody with superior enrichment properties (see methods) to perform ChIP-seq with undifferentiated parental Oct1-sufficient ESCs, as well as cells differentiated with RA for 6 d or towards MD for 3 and 6 d. Oct4 ChIP-seq in undifferentiated ESCs was performed as a parallel control.

In pluripotent cells, ~22,000 Oct4 peaks were identified, corresponding to ~6,000 genes with transcription start sites (TSS) within 20 kb (Table S2). ~45% of Oct4 targets directly overlap with Oct1 peaks and conversely ~60% of Oct1 targets overlap with Oct4 peaks (Fig. 3A). These shared targets in ESCs are on average more strongly enriched for Oct4 than Oct1 as shown by tag density (Fig. 3B). Targets shared by Oct1 and Oct4 in ESCs include *Polr2a*, which encodes the largest subunit of RNA polymerase II, *Pou5f1*, which encodes Oct4, and *Dll1*, which encodes a developmentally-inducible mediator of Notch signaling expressed in the MD lineage. Tracks are shown in Supplemental Fig. S2A. We confirmed Oct1 and Oct4 binding to *Polr2a*, *Pou5f1* and *Dll1* using ChIP-qPCR (Fig. S2B). Oct1 binding to *Polr2a* , which contains two adjacent MOREs that can bind four Oct proteins, is exceptional in that it is stronger than Oct4, however for the other genes the Oct4 signal is more robust. Re-ChIP (sequential ChIP) indicates that Oct1 and Oct4 co-bind these sites, rather than bind in an alternating fashion (Fig. S2C). Cumulatively, these data indicate that Oct1 co-binds with Oct4 to a broad array of targets, including developmental-specific targets, in ESCs.

During RA-mediated differentiation, Oct1 occupies many of the Oct4 peaks present in pluripotent cells (Fig. S2D). 6929 peaks (~2600 genes) shared between Oct1 and Oct4 in pluripotent cells continue to be bound by Oct1 at RA differentiation D6. Nearly as many peaks (5471, ~2200 genes) are exclusively bound by Oct4 in pluripotent cells but become occupied by Oct1 as cells differentiate and Oct4 is lost.

RA differentiation proceeds through pathways that model the natural unfolding of development poorly. Particularly early in RA-mediated differentiation, multiple superimposed developmental lineages are present (Kawaguchi et al. 2010; Walker et al. 2007). The complex populations can complicate analyses of lineage-specific development. In contrast, protocols for MD differentiation mimic normal development (Chal et al. 2015). We therefore performed ChIP-seq using D3 and D6 MD-differentiated cells. In contrast to RA differentiation, only ~200 Oct4-bound peaks not occupied by Oct1 in pluripotent cells become occupied by Oct1 at D6 of MD differentiation (Fig. 3A). Another ~800 peaks shared by Oct4 and Oct1 in pluripotent cells continue to be bound by Oct1 at D6. Analysis of these peaks using GREAT (McLean et al. 2010) identifies enrichment for oxidative stress, ribosomal and mitochondrial organization, Notch signaling and post-implantation development including somite formation (Fig. S3A). Also in contrast to RA differentiation, >6000 peaks become uniquely bound by Oct1 at MD D6 (Fig. 3A). ~2300 peaks are uniquely bound at D3 (Supplemental Fig. S3B). The positions of these peaks shift with differentiation. In ESCs, both Oct1 and Oct4 primarily bind to distal regulatory regions 5 to 500 kb from the TSS. Upon differentiation, binding becomes enriched at more proximal regions <5 kb from the TSS (Fig. S3C). These data suggest that in the pluripotent state, Oct1 and Oct4 bind a broad range of targets primarily at distal enhancers to maintain cellular plasticity and permissiveness for different lineages. Upon differentiation, the majority of distal peaks are lost and binding becomes restricted to proximal sites in key lineage-specific genes.

To pattern the Oct1 and Oct4 occupation during MD differentiation, we applied hierarchical clustering. Four major clusters of genes were identified, three of which show static or decreased Oct1 binding over time (Fig. 3C, clusters 1, 2, 4). Gene ontogeny (GO) analysis indicates involvement in signaling, mitochondrial function, stem cell differentiation and the regulation of Notch and TGF-β signaling. In contrast, Oct1 binding increases with differentiation in cluster 3. Genes near these peaks are associated with chromatin-modifying activities including acetyltransferase and polycomb group complexes. Oct1 and Oct4 enrichment at example cluster 4 genes encoding polycomb complex members is shown in Fig. 3D. These genes (*Ezh2*, *Suz12*, *Ring1* and *Ezh1*) show a gain in Oct1 occupancy as cells progress from pluripotency through D3 to D6 of MD differentiation. In the scRNA-seq UMAP projection, *Ezh2* expression is downregulated in cKO cells at D6, in particular in paraxial mesoderm (Fig. 3E, F). The pseudotime induction of *Ezh2* expression observed in parental cells also fails in cKO (Fig.

3G). These data indicate that during MD differentiation Oct1 directly binds to, and mediates the induction of, genes encoding epigenetic regulators.

Oct1 is Necessary for Appropriate Bivalency Resolution

To investigate pertubations in gene expression caused by Oct1 loss, we performed RNA-seq using parental and cKO ESCs, and cells differentiated using RA for 2 or 6 d. ~2,200, ~3,900 and 7,000 significantly differentially expressed genes were identified at D0, D2 and D6, respectively (Table S3). Hierarchical clustering reveals groups of genes regulated similarly between parental and cKO cells, and others with differential expression. One group was marked by strong induction in differentiating cKO but not parental D6 cells (Supplemental Fig. S4). GO analysis of these differentially expressed genes reveals association with visceral endodem and extraembryonic tissue, suggesting the inappropriate adoption of gene expression programs for alternative lineages (Supplemental Fig. S4). This is consistent with prior findings showing lineage-inappropriate gene expression in RA-differentiating cells lacking Oct1 (Shen et al. 2017). Another group of genes are not induced in the cKO condition and are enriched in mesenchyme and MD specification. Examples include *Eya4*, *Pax6*, and *Rspo3* (Fig. 4A, box). GO analysis of this group identifies. Although RA ultimately results in neuronal cells, these programs are transiently induced during differentiation (Cunningham and Duester 2015). To identify potential regulators of these genes, we queried the ChIP Enrichment Analysis (ChEA) database (Lachmann et al. 2010). Genes with failed induction in differentiating cKO cells tend to be bound by the polycomb repressor compex 2 (PRC2) components Suz12 and Mtf2 (Fig. 4B). These data indicate that loss of Oct1 results in defective induction of developmental genes that are also regulated by PRC2 and H3K27me3.

To test the hypothesis that Oct1 resolves bivalent, lineage-specific genes to their normal activation state, we studied their expression in parental and cKO ESCs and RA-differentiated cells. To identify bivalent genes, we intersected ESC H3K27me3- and H3K4me3-enriched ChIP-seq peaks from the ENCODE database. This procedure identifies 3861 bivalent genes (Table S4). This gene set was intersected with the RNA-seq set from 0, 2 and 6D RA-differentiated parental and cKO ECSs, identifying 1465 bivalent genes expressed in one or more conditions (Table S5). We limited our analysis to the 364 genes with bivalent peaks in the core promoter (±200bp from the TSS, Table S5). These bivalent genes are induced poorly in cKO cells, maintaining a low expression mean similar to the undifferentiated state (Fig. 4C top panel). Two-way ANOVA shows that the differences are significant comparing either parental vs cKO, or

differentiation across timepoints ($p < 5e-16$) (Fig. 4C). The interaction between genotype and differentiation is also significant ($p < 2e-6$), indicating that the bivalent gene induction defects in cKO cells accumulate over time.

To query Oct1 occupancy on bivalent genes during differentiation, we intersected the high-quality bivalent gene dataset with Oct1 ChIP-seq in pluripotent and MD-differentiated parental cells, observing an increase in binding over time (Fig. 4D, green and blue lines). A similar analysis at D6 using just MD-specific bivalent genes (generated by intersecting the bivalent gene dataset with MD development GO:0007498, Table S6) shows even stronger Oct1 enrichment (Fig. 4D, purple line). These findings indicate that 1) Oct1 binds lineage-specific bivalent genes during differentiation, and 2) differentiating Oct1 cKO ESCs fail to correctly induce these genes as part of normal lineage-specific gene expression programs.

Bivalency is positively resolved via the removal of repressive H3K27me3 marks to activate lineage-appropriate genes, or negatively resolved by the removal of activating H3K4me3 mark to repress lineage-inappropriate genes (Dhar et al. 2016; Bernstein et al. 2006). To test if cKO cells inappropriately retain H3K27me3 at lineage-appropriate genes during MD differentiation, we performed ChIP-qPCR using primers flanking a strong Oct1 peak on the *Pax3* promoter (Fig. 5A). *Pax3* is induced by D6 of MD differentiation (Fig. 2C). H3K27me3 was robust and equivalent in D0 undifferentiated cells, and D6 parental cells show reduced H3K27me3. In contrast, cKO retain elevated H3K27me3. Consequently, ~3-fold higher H3K27me3 enrichment is observed in differentiated cKO relative to parental cells (Fig. 5B). Global levels of H3K27me3 are unchanged at the same timepoint as evidenced by immunofluorescence (Fig. 5C), indicating that the effects are gene-specific.

We hypothesized that Oct1 recruits H3K27-specific demethylases to lineage-specific bivalent genes to mediate their induction during differentiation. One such demethylase is Utx/Kdm6a (Cloos et al. 2008). In ChIP-qPCR experiments at the same *Pax3* promoter, Uts was enriched at D6 of MD differentiation (Fig. 5D). In scRNA-seq of MD-differentiated cells, *Pax3* mRNA expression is down-regulated at D6 (Fig. 1D). In addition, fewer Pax3-expressing cells are present in paraxial MD population at D6 (Fig. 5E).

We tested Oct1-Utx interaction by co-immunoprecipitation. We immunoprecipitated Utx from extracts from undifferentiated parental ESCs, as well as cells at D6 of MD- and RA-mediated differentiation, immunoblotting for Oct1. The Oct1-UTX interaction is established in ESCs and maintained during differentiation along multiple developmental lineages (Fig. 5F). Utx

recruitment by Oct1 potentially explains the failure of bivalent, lineage-specific genes such as *Pax3* to positively resolve bivalency in cKO cells (Fig. 5B).

Smad2 motifs are enriched in Oct1-bound peaks in cells differentiating towards MD

The broad expression of Oct1 raises the question of how Oct1 recruits Utx specifically at lineage-specific genes to mediate gene activation. Chromatin accessibility, lineage-specific signaling and co-bound transcription regulators may provide this specificity. We performed motif analysis (Heinz et al. 2010) using DNA sequences ±100 bp from the center of the 807 peaks co-bound by Oct1 and Oct4 in pluripotent cells that remain Oct1-bound during MD differentiation (Fig. 3A). This procedure identifies not only Oct1 sites, but also binding sites for Smads, the terminal transcription regulators of TGFβ/Nodal signaling (Fig. 5F). Smad transcription factors are not identified in an equivalent analysis using RA-mediated differentiation (Fig. 5G). TGFβ signals are critical for MD induction (Zhou et al. 1993; Conlon et al. 1994). A study in zebrafish prominently identified Oct1 transcription factor motifs nearby binding events for Smad2/3 (Liu et al. 2011). This study further showed that ectopic expression of the zebrafish Oct1 ortholog enhances mesoderm induction, that Smad2 and Oct1 physically interact, and that the two factors cooperate to enhance transcription. Smad3 also interacts with Oct4 and co-occupies target genes in pluripotent cells (Mullen et al. 2011). Consistent with these findings, Oct1 immunoprecipitation in both pluripotent and D6 MD-differentiated cells efficienty detects co-precipitated Smad3 (Fig. 5H). *Smad3* is expressed in both populations, but induced by MD differentiation (Fig. 5I).

Loss of Oct1 results in phenotypically abnormal terminal MD differentiation

To study the consequences of Oct1 deficiency for MD development, we differentiated parental and cKO cells for 19 days, and queried expression of the myogenic genes *Myod* and *Myog* by RT-qPCR. These factors are typically detected by D8. Relative to Rps2 (which encodes a ribosome 40S subunit), both genes are strongly expressed in parental but not cKO cells (Fig. 6A). Consistently, after 23 days of differentiation, immunostaining for myosin heavy chain (MHC) reveals robust expression in parental cells, with cKO expression close to background (Fig. 6B). These results demonstrate that cKO ESCs differentiate into muscle poorly, consistent with their defective early gene expression programs and developmental trajectories.

Metabolic changes during development are necessary to complete the MD differentiation program (Oginuma et al. 2017; Cliff et al. 2017). Among these changes are the induction of high glycolytic flux during formation of paraxial MD, which is required for posterior elongation of embryonic axis (Oginuma et al. 2017). To determine changes in steady-state metabolites in parental and cKO ESCs at day 0 and 6 of MD differentiation, we performed GC-MS. Principle component analysis of the metabolites showed few differences in metabolite levels (Supplemental Fig, S5A). Differentiation of parental cells significantly changed metabolite composition, in particular in the PC1 axis (Supplemental Fig. S5A). Metabolites markedly changed by differentiation include the amino acids leucine and threonine (up), and TCA intermediates such as malate and fumarate (down, Supplemental Fig. S5B). Compared to parental cells, cKO cells showed a significant downregulation ($P<0.007$) of both D-glucose and pyruvate, and an increase ($P<0.01$) in TCA intermediates such as citrate, isocitrate, malate and succinate (Supplemental Fig. S5C). These results suggest that cKO cells induce glycolysis poorly during MD differentiation.

Ectopic Oct1 expression improves mesodermal specification and differentiation

We transduced parental and cKO ESCs with retroviral vectors encoding murine Oct1, or empty vector (EV) controls. The vectors encode a puromycin resistance cassette, allowing for selection of transduced cells. Immediately after selection, populations of transduced cells were subjected to MD differentiation for 11 d, and probed for *Myod*, *Myog* and *Pax3* by RT-qPCR. EV-transduced cKO cells showed defects in *Myod* and *Myog* expression (Fig. 6C), as observed in untransduced cells. Strikingly, Oct1-transduced ESCs strongly express both genes. In contrast *Pax3*, which is important for early muscle lineage specification, is decreased (Fig. 6C). The combination of elevated *Myod* and *Myog* expression and decreased *Pax3* at late differentiation timepoints suggests that ectopic Oct1 expression enables cells to more efficiently transit through a *Pax3*-expressing intermediate state, such that more cells enter into a terminal differentiation state expressing the terminal myogenic program but not *Pax3*. Immunoblotting confirmed ectopic Oct1 expression (Fig. 6D).

Discussion

During development, chromatin transitions from a pluripotent state permissive for different lineages to a lineage-restricted state. Pluripotent cells maintain genes encoding developmental-specific mediators in a poised “bivalent” configuration that allows for either induction or stable repression later, depending on the developmental lineage (Bernstein et al. 2006). Here we show that the transcription factor Oct1, an Oct4 paralog, directly binds to, and mediates the induction of, developmental lineage-specific bivalent genes. Oct1 mediates the induction of these genes by recruiting the UTX histone demethylase to remove repressive H3K27me3 marks. As a consequence of these molecular defects, Oct1 deficient ESCs manifest defective MD differentiation including failure to reprogram metabolism and failure to express *Myod* and *Myog* mRNA, and myosin heavy chain protein.

The central role of Oct1 in “canalizing” differentiating pluripotent cells was shown using single-cell RNA-seq with Oct1 inducible-conditional (cKO) ESCs in early steps of mesodermal specification. In the absence of Oct1, cells undergoing differentiation retain pluripotency characteristics such as an epithelial state, and fail to achieve somite-stage gene expression patterns. These cells undergo inappropriate developmental branching with mis-expression of developmentally inappropriate genes. Velocity analysis indicates that cKO cells progress towards poorly differentiated state marked by epithelial gene expression and oxidative stress. The induction of genes important early in MD differentiation such as *Tbxt* (*T*, *Brachyury*), becomes weaker and asynchronous, losing temporal coherence. Later, there is failed induction of genes such as *Pax7* and *Cxcr4*.

In differentiating cells, Oct1 occupies subsets of genes bound by Oct4 in pluripotent cells. During RA-mediated differentiation, the spectrum of Oct1-bound targets largely overlap Oct4 binding events. This robust occupancy may reflect a superposition of many developmental states early in RA-mediated differentiation, which does not recapitulate standard steps associated with embryonic development. In contrast, during MD differentiation Oct1 occupies a small subset of Oct4-bound genes. These include MD-specific bivalent developmental mediators and genes encoding chromatin-modifying enzymes.

Bivalent, developmentally poised genes are poorly induced in cKO cells. Oct1 occupancy on these genes normally increases with differentiation and shows the stongest enrichment with lineage-specific genes, suggesting a critical role in canalization of cell fate. Oct1 also binds and induces the expression of genes encoding PRC complex members such as *Ezh2* in MD-specific clusters in parental but not cKO cells. In addition, Oct1 recruits Utx to

lineage-specific, Oct1-bound targets such as *Pax3.* The lack of Utx recruitment to lineage-specific genes in cKO cells is consistent with the abnormal retention of H3K27me3 at the promoters of these genes and their failed trascritional upregulation. The mechanism(s) that allow Oct1 to use Utx as a cofactor specifically at developmental-appropriate targets are incompletely understood but may involve other collaborating transcription factors. For example, Smad transcription factor binding sites are enriched near sites of Oct1 binding during MD- but not RA-mediated differentiation. TGFβ signals drive Smad transcription factor binding and are important for MD specification (Mullen et al. 2011; Liu et al. 2011). Co-IP experiments in both pluripotent and D6 MD-differentiated cells show an interaction between Oct1 and Smad3, consistent with prior findings (Liu et al. 2011).

We show that ectopic Oct1 expression can improve canalization and lineage-specific gene expression. Under MD differentiation conditions, exogenous Oct1 increases expression of the terminal differentiation markers *Myod* and *Myog*, while decreasing the lineage-specification marker *Pax3*, which is transiently expressed during MD differentiation and is necessary for later expression of myogenic genes. Because *Pax3* is no longer expressed in terminally differentiating cells, these results suggest that ectopic Oct1 enables transit through a *Pax3*-expressing intermediate to potentiate productive terminal differentiation. Understanding how to improved canalization of differentiating cells could lead to improvements in regenerative medicine, and diagnostic and therapeutic approaches for various diseases.

Fig. Legends

Figure 1. Differentiating Oct1 deficient ESCs fail to canalize into mesodermal lineages. (*A*) UMAP projection of scRNA-seq data from superimposed parental (Oct1 sufficient) undifferentiated ESCs, and parental cells at days 3 and 6 of MD differentiation. Cells were clustered based on gene expression and labeled based the expression of developmental markers as in (*B*). Three combined replicate plates were used for the analysis. (*B*) Violin plots showing gene expression levels of key developmental markers by cluster. Data were log-normalized for each cell using the natural logarithm, scaled and averaged using mean(expm1(x)). (*C*) Comparative UMAP projections of integrated D6 parental and Oct1 deficient (cKO) scRNA-seq populations. Clusters were labeled computationally and identified based on gene expression. Relative frequencies are shown. (*D*) Differential analysis of anterior MD shown as a scatter plot. Red dots label significantly differentially expression genes based on FDR corrected $p < 0.05$ and a fold change >1.2. Example genes are labeled.

Figure 2. Oct1 deficient cells show perturbed developmental trajectories. (*A*) Pseudotime analysis of pluripotent and differentiating parental (top panel) and cKO (bottom panel). Colors of the dots correspond to the time point at which cells were collected (red: D0, green: D3, blue: D6). (*B*) Expression levels of *Klf4*, *Tbxt*, *Pax7* and *Cxrc4* across pseudotime in parental (top panels) and cKO (bottom panels). Black trendline represents an average expression across all populations. (*C*) Heatmap depicting expression of the 2000 most dynamic genes (based on Seurat's FindVariableFeatures function) in parental D6 cells. Gene expression was plotted as a heat map across pseudotime in parental (left panel) and cKO (right panel). Dynamic genes were first hierarchically clustered in parental cells to cluster groups of genes that behave similarly in pseudotime, then plotted in the same order in cKO cells. (*D*) Velocity gene expression analysis of parental (left panel) and cKO (right panel) cells at differentiation day 6. Direction of arrows point toward clusters with gene expression closest to the future state of each cell based on unspliced/spliced transcripts. Arrow length represents magnitude.

Figure 3. Oct1 co-occupies target sites with Oct4 in ESCs, and regulates their expression during differentiation. (*A*) Venn diagram showing common and unique binding sites based on Oct4 and Oct1 ChIP-Seq in ESCs, d3 and d6 of mesodermal differentiation shows shared peaks between different conditions. (*B*) Tag density around peak center in Oct4 (red) and Oct1 (blue) ChIPseq

data sets. (*C*) A matrix of ChIPseq enrichment 2 kb around peak centers was computed for the merged peak list and shown as heatmap. Loci in red show high enrichment and white loci show no enrichment. GO terms corresponding to each cluster are shown on the right. (*D*) Example tracks of representative loci are shown. Y-axes were scaled to the same value for each gene. (*E*) *Ezh2* gene expression is shown in UMAP projections for parental and cKO ESCs at D6 of MD differentiation. The paraxial MD cluster is outlined in red. (*F*) Violin plots showing *Ezh2* gene expression in cells within the paraxial MD cluster shown in (*E*). (*G*) *Ezh2* expression in pseudotime in parental (top panel) and cKO (bottom panel) cells. Black trendline represents average expression across all the cells at any given position in pseudotime.

Figure 4. Bivalent genes fail to upregulated in differentiating Oct1 conditional knockout ESCs. (*A*) Heatmap of D6 differential genes (parental vs. cKO) in all time points is shown. Genes in red show high expression and blue genes show low expression/no expression. Representative genes that correspond to the middle cluster are shown on the right. (*B*) ChIP-X Enrichment Analysis (ChEA) query results for middle cluster in (*A*) are shown. (*C*) Expression of annotated bivalent genes is shown using violin plots for parental (blue) and cKO (red) ESCs at RA differentiation days 0, 2 and 6. Table below shows results of two-way ANOVA test comparing the two genotypes during differentiation. (*D*) Oct1 enrichment based on tag density around peak center at annotated bivalent genes in ESCs (red), and at MD differentiation D2 (green) and D6 (blue). An additional analysis was performed for MD-specific bivalent genes at MD differentiation D6 (purple) are shown.

Figure 5. Oct1 interacts with Utx to demethylate H3K27me3 at bivalent genes. (*A*) Signal tracks (*Mm10* v. D191020) showing Oct1 and Oct4 enrichment at the *Pax3* 5' region. Positions of identified HOMER Smad2/3/4 motifs are shown below. (*B*) H3K27me3 ChIP-qPCR enrichment at the *Pax3* promoter in parental and cKO ESCs and at MD differentiation D0 and D6. Enrichment is normalized to a nonspecific genomic region encoding the 28S ribosomal subunit that is set to 1. (*C*) H3K27me3 immunofluorescence images from parental and cKO cultures differentiated for 6 d. Images were collected a 40× magnification. (*D*) Utx enrichment at the *Pax3* promoter in parental and cKO cells on D6 of differentiation. (*E*) UMAP projection of scRNA-seq data using parental and cKO cells at MD D6, highlighting *Pax3* expression. (*F*) Immunoprecipitation of Utx followed by Oct1 immunoblot using parental ESCs, ESCs differentiated for 6 d under mesoderm-promoting conditions and ESCs differentiated using RA for 6 d. (*G*) HOMER motif analysis of Oct1

peaks that shared with Oct4 and maintained after 6 d of differentiation towards the mesoderm lineage (left panel) or using RA differentiation (right panel). (*H*) Smad3 immunoblot using cell lysates immunoprecipitated with Oct1 antibodies, or rabbit IgG controls. Parental ESCs (lanes 1-3) and MD D6-differentiated cells (lanes 4-6) were used. 5% input is shown, though Smad3 was undetectable with these amounts. (*I*) *Smad3* gene expression is shown in violin plots for parental ESCs at D6 of MD differentiation.

Figure 6. Myogenic lineage induction is defective in the absence of Oct1 and augmented with Oct1 ectopic expression. (*A*) Relative mRNA expression levels of myogenic genes *Myod1* and *Myog* in parental and cKO differentiated for 19 d. (*B*) MF20 staining showing Myh1e expression in parental (left panel) and cKO (right panel) cells at day 23 of MD differentiation. (*C*) RT-qPCR for the myogenic genes *Myod1*, *Myog* and *Pax3* in parental and cKO cells transduced with retroviruses encoding Oct1, or empty vector controls. Transduced cells were selected using puromycin for 48 hr prior to differentiation for 11 d. (*D*) Immunoblot showing successful ectopic Oct1 expression in cKO cells.

Materials and methods

Cell culture

ESCs were cultured as previously described (Shakya et al., 2015a) with 2i conditions: ERK inhibitor PD0325901 (1 μM, LC Laboratories) and GSK3 inhibitor CHIR99021 (3 μM, LC Laboratories). Cultures were maintained on irradiated feeders (ThermoFisher). Prior to all experiments ESCs were plated on gelatin to deplete the feeders. For MD differentiation, ESCs were plated on gelatin and cultured as previously described (Chal et al. 2015). Briefly, parental and cKO cells were cultured in N2B27 medium supplemented with recombinant Bmp4 (Peprotech) for 2 d. After 48 hr, media was changed to RDL (Rspo3, DMSO, LDN) medium. Cells were harvested 24 hr (day 3) or 96 hr (day 6) later. For muscle differentiation, cells were switched to HIFL (Hgf, Igf, Fgf, Ldn) medium and cultured for 48 hr (day 8) after which medium was switched to 2% horse serum (ThermoFisher). Cells were harvested on day 11 (overexpression experiments) or 19 (RT-qPCR).

Single cell RNA-seq

Single cell transcriptomes were analyzed as described previously (Dell'Orso et al. 2019). The 10X Genomics Chromium Single Cell Gene Expression Solution with 3' chemistry, version 3 (PN-1000075) was used to tag individual cells with 16 bp barcodes and specific transcripts with 10 bp Unique Molecular Identifiers (UMIs) according to manufacturer instructions. Briefly, single-cell suspensions were isolated using trypsinization and resuspension in PBS with 0.04% BSA (ThermoFisher). Suspensions were filtered through 40 μm cell strainers. Viability and cell count were assessed using a Countess I (ThermoFisher). Equilibrium to targeted cell recovery of 6,000 cells along with Gel Beads and reverse transcription reagents were loaded to Chromium Single Cell A to form Gel-bead-in Emulsions (GEMs). Within individual GEMs, cDNA generated from captured and barcoded mRNA was synthesized by reverse transcription at 53°C for 45 min. Samples were then heated to 85°C for 5 min. Subsequent A tailing, end repair, adaptor ligation and sample indexing were performed in bulk according to manufacturer instructions. The resulting barcoding libraries were qualified on Agilent Technology 2200 TapeStation system and subjected to qPCR using a KAPA Biosystems Library Quantification Kit for Illumina Platforms (KK4842). The multiple libraries were then normalized and sequenced on an Illumina NovaSeq 6000 using the 2×150 PE mode.

Data Processing and clustering

Sequences from the Chromium platform were de-multiplexed and aligned using CellRanger ver. 3.1.0 (10X Genomics) with default parameters mm10-3.0.0. Clustering, filtering, variable gene selection and dimensionality reduction were performed using Seurat ver.3.1.5 (Stuart et al. 2019) according to the following workflow: 1, Cells with <300 detected genes and >10000 genes were excluded further analysis. 2, Cells with <12% UMIs mapping to mitochondrial genes were retained for downstream analysis. 3, The UMI counts per ten thousand were log-normalized for each cell using the natural logarithm. 4, Variable genes (2000 features) were selected using the FindVariableFeatures function. 5, Common anchors between the three parental timepoints (Fig.1A) or parental and cKO D6 (Fig.1C) were identified using FindIntegrationAnchors function that were further used to integrate these sets. 6, Gene expression levels in the integrated set were scaled along each gene and linear dimensional reduction was performed. The number of principal components was decided through the assessment of statistical plots (JackStrawPlot and ElbowPlot). 7, Cells were clustered using a by a shared nearest neighbor (SNN) modularity optimization-based clustering algorithm and visualized using two-dimensional uniform manifold approximation and projection (UMAP). 8, Cluster identities were defined based on the distribution of the specific markers. Differentiational gene expression analysis between the parental and cKO clusters was performed using FindMarkers. Genes with adjusted $p<0.01$ were marked red on scatter plots.

Pseudotime and Velocity analysis

Trajectory analysis of scRNA-seq was performed using Monocle v.2.16.0 (Trapnell et al. 2014). WT and KO sets were filtered using same parameter as above and merged to generate WT and KO set. Cells were ordered based on gene lists for the ESC (beginning) and somite (end) clusters in parental UMAP (Fig.1A). Next, we performed dimensional reduction using the DDRTree method to visualize the dataset, ordered the cells by global gene expression levels, and visualized the trajectory of the dataset. Veolocity analysis was performed using the velocyto package (La Manno et al. 2018). Loom files were produced using following paramters: velocyto run10x -m mm10. rmsk.gtf genes.gtf. Gtf files were produced from the Cell Ranger pipeline. Velocity embeddings were produced using the velocyto.r and SeuratWrappers packages. Matrices were filtered using following parameters: nFeature_spliced > 300, nFeature_spliced < 10000, nFeature_unspliced > 200, nFeature_unspliced < 6000, percent.mt < 12. Velocity was calculated using RunVelocity using following paremeters: deltaT = 1, kCells = 25, fit.quantile =

0.02. Velocity embedding were projected on T-SNE maps using the show.velocity.on.embedding.cor function.

RT-qPCR

RNA was isolated using TRIzol (ThermoFisher). cDNA was synthesized using a SuperScript Vilo cDNA Synthesis Kit (ThermoFisher). RT-qPCR oligonucleotide primers are listed in Table S7 and were confirmed to generate a single product of the correct size. To ensure specific PCR amplification, every RT-qPCR run was followed by a dissociation phase analysis (denaturation curve) to confirm the presence of single amplified peaks.

ChIP

ChIP-qPCR and ChIP-seq were performed as previously described (Perovanovic et al. 2016). Briefly, WT and cKO cells were crosslinked with 1% formaldehyde for 10 min and quenched for 5 min using 2.5M glycine. Culture plates were washed using ice cold PBS and cells were harvested by cell scaping. Cells were lysed in Farnham buffer (5 mM Pipes pH 8.0, 85 mM KCl, 0.5% NP-40) and subsequently in RIPA buffer (phosphate-buffered saline, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS). Chromatin was sheared using a Covaris sonicator for 5 min (30 sec on/30 sec off) with amplitude=40. Correct chromatin fragmentation was confirmed using 1% agarose gels. 50 µg of chromatin was subjected to IP overnight at 4°C with 4 µg of anti-Oct1 (Novus Biological), Oct4 (Santa Cruz) or H3K27me3 (Milipore) antibodies. As a control, we used 5 µg of sheared, non-precipitated input chromatin. Samples were incubated with protein G magnetic beads (ThermoFisher) for 5 hr and washed in Low Salt buffer (20 mM Tris-Cl pH 8.0, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100), High Salt buffer (identical but with 500 mM NaCl), LiCl buffer, and Tris-EDTA pH 8.0 plus 1 mM EDTA (TE buffer). Washes were performed at 4°C for 10 min with rotation. For re-ChIP, 2% fragmented chromatin was saved as input and the rest used for IP with Oct4 antibody at 4°C overnight on a rotator. Samples were then incubated with magnetic beads for 5h at 4°C. The beads were washed for 10 min with Low Salt buffer, High Salt buffer, Low Salt buffer, LiCl buffer, and TE buffer sequentially at 4°C. Chromatin was eluted with 300 μL IP Elution buffer at RT, then diluted 10-fold in RIPA buffer. Diluted samples were then subjected to a second IP with 4 μg of Oct1 antibody (Novus Biological) at 4°C overnight, and incubated with magnetic beads for 5 hr at 4°C. The beads were washed again as described above, then eluted with 300 μL IP Elution Buffer at RT. Re-ChIP samples, together with the 2% input, were incubated at 65°C overnight to reverse

crosslinking. DNA was purified using phenol-chloroform-isoamyl alcohol extraction followed by PCR clean up. qPCR primers can be found in Table S7 and were confirmed to generate a single product of the correct size. The results were reported as qPCR values normalized to input chromatin (gDNA) and non-specific region and presented as fold enrichment.

ChIP-seq analysis

After chromatin was precipitated as described above, and libraries were sequenced using Illumina NovaSeq. Between 22 and 26 million paired-end Illumina sequencing reads were aligned to the mouse *Mm10* reference genome using Novocraft novoalign v3.8.2, allowing for one random alignment of multi-mapping reads, and providing the adapter sequences for automatic trimming concordant with alignment. ChIP was analyzed using the MultiRepMacsChIPSeq pipeline v12.2, using options "--pe --optdist 10000 --dupfrac 0.15 --mapq 10 --cutoff 2 --tdep 10 --peaksize 500 --peakgap 100 --species mouse --chrskip 'chrMIPhiX' --blacklist mm10.blacklist.bed".

Immunoprecipitation

Cells were lysed with Cell Lysis Buffer (Life Technologies) in the presence of protease inhibitors (EDTA-free tablet, Roche). IP was performed using 500 µg of extract. Extracts were incubated with 4 µg of anti-Utx (Cell Signaling, D3Q1I) or Oct1 (Novus Biologicals, NBP2-21584) antibodies, or rabbit IgG control overnight at 4°C. Protein-antibody complexes were precipitated with protein-G magnetic beads (Thremo Fisher) for 3 hr at 4°C with rotation and washed 3 times with Low Salt buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100) plus protease inhibitors. Precipitated proteins were analyzed by immunoblot.

Bulk RNA-seq

RNA was prepared from three independent cultures of undifferentiated or 2 d and 6 d RA-differentiated parental or cKO ESCs. Poly(A) RNA was purified from total RNA samples (100–500 ng) with oligo(dT) magnetic beads, and stranded mRNA sequencing libraries were prepared as described using the Illumina TruSeq mRNA library preparation kit and RNA UD Indexes. Molarity of adapter-modified molecules was defined by qPCR using the Kapa Biosystems Library Quant Kit. Individual libraries were normalized to 1.3 nM. Sequencing libraries were chemically denatured and applied to an Illumina NovaSeq flow cell using the NovaSeq XP chemistry workflow. Following transfer of the flowcell to an Illumina NovaSeq instrument, 2×51 cycle paired-end sequence was performed using a NovaSeq S1 reagent kit. Between 13 and 18

million paired-end reads were generated for each condition. More than 99% of aligned reads mapping to the correct strand.

Bulk RNAseq analysis

The Illumina adapter sequence was trimmed using cutadapt version 1.16. Fastq data quality were checked using Fastqc verision 0.11.5. Reads were aligned to the mouse *Mm10* genome using STAR version 2.7.3a in two-pass mode. Aligned reads were checked for quality using the Picard tools' CollectRnaSeqMetrics command to count the number of read-matching exons, UTRs, introns and intergenic regions, and to calculate normalized gene coverage from the top 1000 expressed transcripts. Between 13 and 18 million paired-end reads were generated for each condition, with >99% of aligned reads mapping to the correct strand. Differentially expressed genes were identified using a 5% FDR with DESeq2 version 1.24.0 (Love et al. 2014). Genes with a count of at least 50 in one or more samples were tested. Genes showing at least 2.5-fold change of expression at an adjusted $p<0.01$ were selected as differentially expressed. Figures were generated in R version 4.0.0 using functions from ggplots libraries and pheatmap.

Immunofluorescence

Immunofluorescence was performed as described previously (Gnocchi et al. 2009) with modifications, using rabbit anti-H3K27me3 (Milipore) and mouse anti-MHC (MF20, Developmental Hybridoma bank) antibodies. Secondary antibodies were goat anti-rabbit-Alexa568 and anti-mouse-Alexa568 (ThermoFisher).

Retroviral Oct1 overexpression

Oct1 was ectopically expressed in ESC cells using a previously described retroviral vector (pBabePuro-Oct1) (Kang et al. 2009). pBabePuro was used as an empty vector control. The vector was co-transfected together with the pCL-Eco packaging plasmid into HEK293T cells to generate retroviral particles. Retroviral supernatants were collected 48 hr later, filtered through 0.45 μm filters and applied to ESCs cultures maintained on puromycin-resistant feeder fibroblasts (ThermoFisher). The mixed population of cells was subjected to selection with puromycin for 48 hr.

Metabolic profiling

Cold 90% methanol (MeOH) was added to each sample to give a final concentration of 80%. Samples were then incubated at -20˚C for 1 hr. After incubation the samples were centrifuged at 20,000×*g* for 10 min at 4˚C. The supernatant was then transferred from each sample tube into a labeled, fresh micro centrifuge tube. Pooled quality control samples were made by removing a fraction of collected supernatant from each sample and process blanks were made using only extraction solvent and no cell culture. The samples were then dried en vacuo. GC-MS was performed with an Agilent 5977b GC-MS MSD-HES and an Agilent 7693A automatic liquid sampler. Data were analyzed using in-house software to prepare for analysis by the "MetaboAnalyst" software tool (Chong et al. 2018). Statistical analysis was performed using MetaboAnalystR.

Acknowledgements

We thank G. Kardon and C. Kikani for critical reading of the manuscript. We thank B. Dalley and O. Allen at the HCI High-Throughput Genomics facility and T. Parnell and B. Lohman at the HCI Bioinformatic Analysis Shared Resource for assistance with ChIP-seq and scRNA-seq. We thank James E. Cox and Tyler Van Ry from the Metabolomics Core for their assistance with metabolomic profiling. We thank Olivier Pourquié from the Harvard Medical School for the assistance with the mesodermal differentiation protocols. This work was supported by a grant from the National Institutes of Health/National Institute of General Medical Sciences (R01GM122778) to DT.

Author contributions: DT conceived and supervised experiments, and provided administrative and material support. JP conceived the study, designed experiments and acquired and interpreted data. ZS and YW acquired and interpreted data. All authors were involved in writing, reviewing and revising the manuscript.

References

Bain G, Kitchens D, Yao M, Huettner JE, Gottlieb DI. 1995. Embryonic stem cells express neuronal properties in vitro. *Dev Biol* **168**: 342–357. https://pubmed.ncbi.nlm.nih.gov/7729574/ (Accessed September 30, 2020).

Beddington RSP, Robertson EJ. 1989. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. *Development* **105**: 733–737. https://pubmed.ncbi.nlm.nih.gov/2598811/ (Accessed July 8, 2020).

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–26. http://www.ncbi.nlm.nih.gov/pubmed/16630819 (Accessed January 21, 2014).

Chal J, Oginuma M, Tanoury Z Al, Gobert B, Sumara O, Hick A, Bousson F, Zidouni Y, Mursch C, Moncuquet P, et al. 2015. Differentiation of pluripotent stem cells to muscle fiber to model Duchenne muscular dystrophy. *Nat Biotechnol* **33**: 962–9.

Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J. 2018. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* **46**: W486–W494. https://pubmed.ncbi.nlm.nih.gov/29762782/ (Accessed November 4, 2020).

Cliff TS, Wu T, Boward BR, Yin A, Yin H, Glushka JN, Prestegaard JH, Dalton S. 2017. MYC Controls Human Pluripotent Stem Cell Fate Decisions through Regulation of Metabolic Flux. *Cell Stem Cell*.

Cloos PAC, Christensen J, Agger K, Helin K. 2008. Erasing the methyl mark: Histone demethylases at the center of cellular differentiation and disease. *Genes Dev* **22**: 1115–1140. https://pubmed.ncbi.nlm.nih.gov/18451103/ (Accessed September 30, 2020).

Conlon FL, Lyons KM, Takaesu N, Barth KS, Kispert A, Herrmann B, Robertson EJ. 1994. A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development*.

Cunningham TJ, Duester G. 2015. Mechanisms of retinoic acid signalling and its roles in organ and limb development. *Nat Rev Mol Cell Biol* **16**: 110–123. https://pubmed.ncbi.nlm.nih.gov/25560970/ (Accessed November 3, 2020).

Dell'Orso S, Juan AH, Ko K-D, Naz F, Perovanovic J, Gutierrez-Cruz G, Feng X, Sartorelli V. 2019. Correction: Single cell analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions (doi: 10.1242/dev.174177). *Development*.

Dhar SS, Lee SH, Chen K, Zhu G, Oh WK, Allton K, Gafni O, Kim YZ, Tomoiga AS, Barton MC, et al. 2016. An essential role for UTX in resolution and activation of bivalent promoters. *Nucleic Acids Res* **44**: 3659–3674. https://pubmed.ncbi.nlm.nih.gov/26762983/ (Accessed October 9, 2020).

Diaz-Cuadros M, Wagner DE, Budjan C, Hubaud A, Tarazona OA, Donelly S, Michaut A, Al Tanoury Z, Yoshioka-Kobayashi K, Niino Y, et al. 2020. In vitro characterization of the human segmentation clock. *Nature* **580**: 113–118. https://doi.org/10.1038/s41586-019-1885-9 (Accessed September 22, 2020).

Gnocchi VF, White RB, Ono Y, Ellis JA, Zammit PS. 2009. Further characterisation of the molecular signature of quiescent and activated mouse muscle satellite cells. *PLoS One* **4**: e5205. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2666265&tool=pmcentrez&rendertype=abstract (Accessed March 3, 2014).

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell*.

Kang J, Gemberling M, Nakamura M, Whitby FG, Handa H, Fairbrother WG, Tantin D. 2009. A general mechanism for transcription regulation by Octl and Oct4 in response to genotoxic

and oxidative stress. *Genes Dev*.

Kawaguchi J, Nichols J, Gierl MS, Faial T, Smith A. 2010. Isolation and propagation of enteric neural crest progenitor cells from mouse embryonic stem cells and embryos. *Development* **137**: 693–704. https://pubmed.ncbi.nlm.nih.gov/20147374/ (Accessed October 15, 2020).

Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**. https://pubmed.ncbi.nlm.nih.gov/18974828/ (Accessed July 8, 2020).

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560**: 494–498. http://www.nature.com/articles/s41586-018-0414-6 (Accessed April 24, 2019).

Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. 2010. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**: 2438–2444. https://pubmed.ncbi.nlm.nih.gov/20709693/ (Accessed September 30, 2020).

Liu Z, Lin X, Cai Z, Zhang Z, Han C, Jia S, Meng A, Wang Q. 2011. Global identification of SMAD2 target genes reveals a role for multiple co-regulatory factors in zebrafish early gastrulas. *J Biol Chem* **286**: 28520–28532. /pmc/articles/PMC3151094/?report=abstract (Accessed October 14, 2020).

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (Accessed October 19, 2020).

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**: 1202–1214. http://dx.doi.org/10.1016/j.cell.2015.05.002 (Accessed September 22, 2020).

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501. http://www.ncbi.nlm.nih.gov/pubmed/20436461 (Accessed November 6, 2013).

Meshorer E, Misteli T. 2006. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol* **7**: 540–6. http://www.ncbi.nlm.nih.gov/pubmed/16723974.

Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. 2011. Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**: 565–76. http://www.ncbi.nlm.nih.gov/pubmed/22036565 (Accessed April 25, 2018).

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Schöler H, Smith A. 1998. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**: 379–391. https://pubmed.ncbi.nlm.nih.gov/9814708/ (Accessed July 8, 2020).

Oginuma M, Moncuquet P, Xiong F, Karoly E, Chal J, Guevorkian K, Pourquié O. 2017. A Gradient of Glycolytic Activity Coordinates FGF and Wnt Signaling during Elongation of the Body Axis in Amniote Embryos. *Dev Cell* **40**: 342–353.e10. http://dx.doi.org/10.1016/j.devcel.2017.02.001 (Accessed August 13, 2020).

Palmieri SL, Peter W, Hess H, Schöler HR. 1994. Oct-4 transcription factor is differentially expressed in the mouse embryo during establishment of the firsttwo extraembryonic cell lineages involved in implantation. *Dev Biol* **166**: 259–267. https://pubmed.ncbi.nlm.nih.gov/7958450/ (Accessed July 8, 2020).

Perovanovic J, DellOrso S, Gnochi VF, Jaiswal JK, Sartorelli V, Vigouroux C, Mamchaoui K, Mouly V, Bonne G, Hoffman EP. 2016. Laminopathies disrupt epigenomic developmental programs and cell fate. *Sci Transl Med* **8**: 335ra58-335ra58. http://stm.sciencemag.org/content/8/335/335ra58.full-text.pdf+html.

Schlesinger S, Meshorer E. 2019. Open Chromatin, Epigenetic Plasticity, and Nuclear Organization in Pluripotency. *Dev Cell* **48**: 135–150.

Sebastiano V, Dalvai M, Gentile L, Schubart K, Sutter J, Wu GM, Tapia N, Esch D, Ju JY, Hübner K, et al. 2010. Oct1 regulates trophoblast development during early mouse embryogenesis. *Development* **137**: 3551–3560. https://pubmed.ncbi.nlm.nih.gov/20876643/ (Accessed July 8, 2020).

Seydoux G, Braun RE. 2006. Pathway to Totipotency: Lessons from Germ Cells. *Cell* **127**: 891–904.

Shen Z, Kang J, Shakya A, Tabaka M, Jarboe EA, Regev A, Tantin D. 2017. Enforcement of developmental lineage specificity by transcription factor Oct1. *Elife* **6**. https://elifesciences.org/articles/20937 (Accessed July 8, 2019).

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888–

1902.e21. https://pubmed.ncbi.nlm.nih.gov/31178118/ (Accessed July 27, 2020).

Takahashi K, Yamanaka S. 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*.

Tantin D. 2013. Oct transcription factors in development and stem cells: insights and mechanisms. *Development*.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**: 381–386. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4122333/ (Accessed July 27, 2020).

Walker E, Ohishi M, Davey RE, Zhang W, Cassar PA, Tanaka TS, Der SD, Morris Q, Hughes TR, Zandstra PW, et al. 2007. Prediction and Testing of Novel Transcriptional Networks Regulating Embryonic Stem Cell Self-Renewal and Commitment. *Cell Stem Cell*.

Wang VEH, Schmidt T, Chen J, Sharp PA, Tantin D. 2004. Embryonic Lethality, Decreased Erythropoiesis, and Defective Octamer-Dependent Promoter Activation in Oct-1-Deficient Mice. *Mol Cell Biol* **24**: 1022–1032.

Zhou X, Sasaki H, Lowe L, Hogan BLM, Kuehn MR. 1993. Nodal is a novel TGF-β-like gene expressed in the mouse node during gastrulation. *Nature* **361**: 543–547. https://pubmed.ncbi.nlm.nih.gov/8429908/ (Accessed October 14, 2020).

Perovanovic *et al.* Figure 1

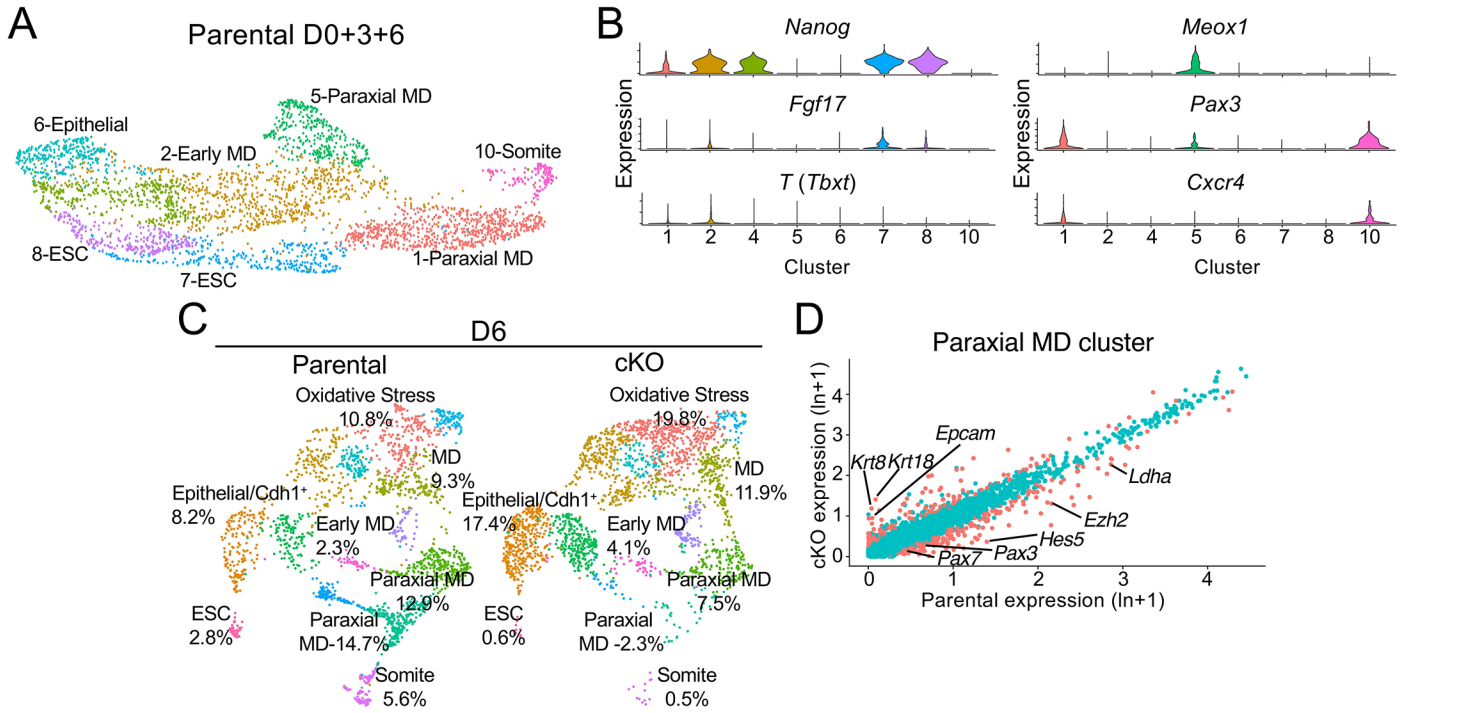

Figure 1. Differentiating Oct1 deficient ESCs fail to canalize into mesodermal lineages. (*A*) UMAP projection of scRNA-seq data from superimposed parental (Oct1 sufficient) undifferentiated ESCs, and parental cells at days 3 and 6 of MD differentiation. Cells were clustered based on gene expression and labeled based the expression of developmental markers as in (*B*). Three combined replicate plates were used for the analysis. (*B*) Violin plots showing gene expression levels of key developmental markers by cluster. Data were log-normalized for each cell using the natural logarithm, scaled and averaged using mean(expm1(x)). (*C*) Comparative UMAP projections of integrated D6 parental and Oct1 deficient (cKO) scRNA-seq populations. Clusters were labeled computationally and identified based on gene expression. Relative frequencies are shown. (*D*) Differential analysis of anterior MD shown as a scatter plot. Red dots label significantly differentially expression genes based on FDR corrected $p < 0.05$ and a fold change >1.2. Example genes are labeled.

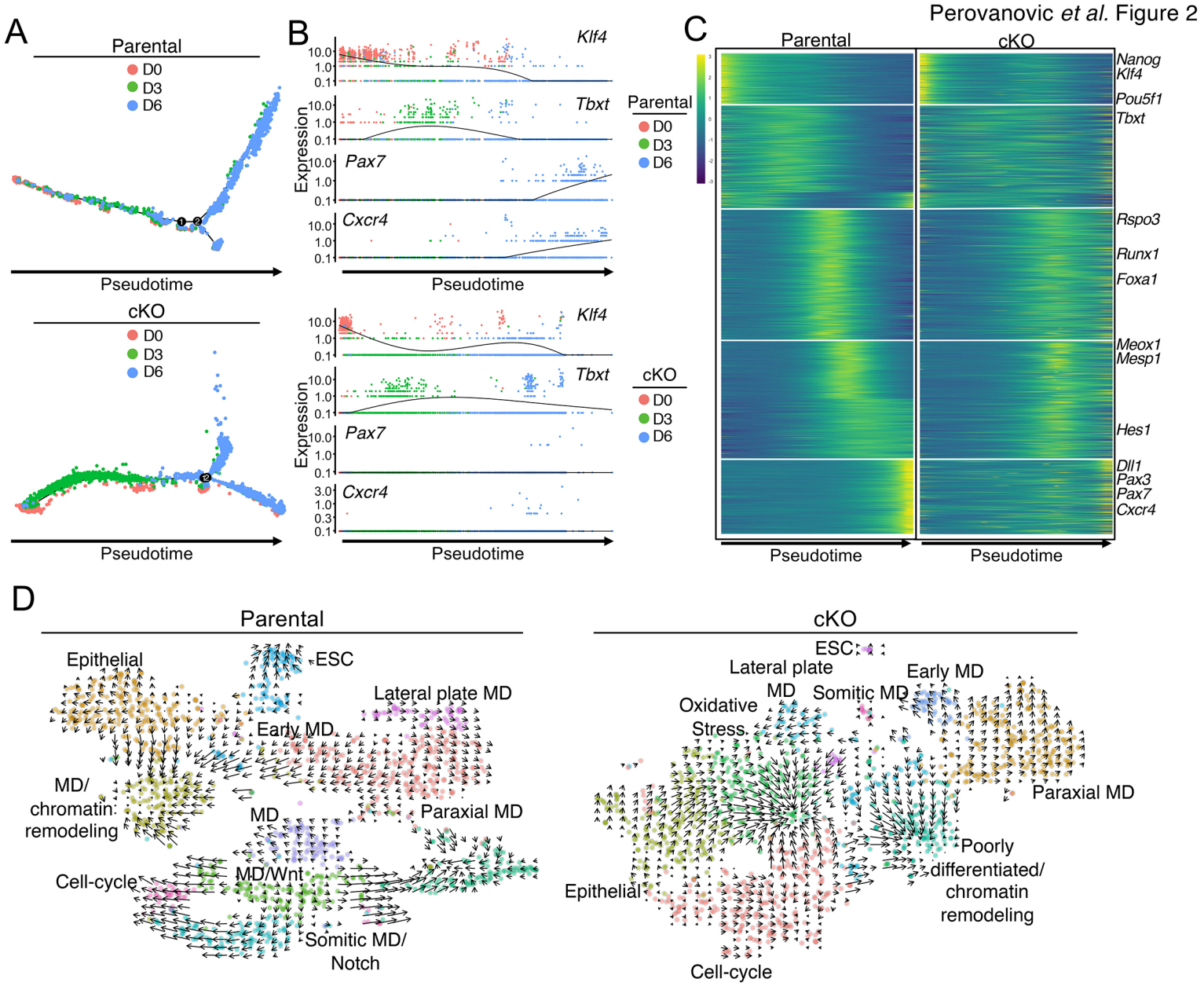

Figure 2. Oct1 deficient cells show perturbed developmental trajectories. (*A*) Pseudotime analysis of pluripotent and differentiating parental (top panel) and cKO (bottom panel). Colors of the dots correspond to the time point at which cells were collected (red: D0, green: D3, blue: D6). (*B*) Expression levels of *Klf4*, *Tbxt*, *Pax7* and *Cxrc4* across pseudotime in parental (top panels) and cKO (bottom panels). Black trendline represents an average expression across all populations. (*C*) Heatmap depicting expression of the 2000 most dynamic genes (based on Seurat's FindVariableFeatures function) in parental D6 cells. Gene expression was plotted as a heat map across pseudotime in parental (left panel) and cKO (right panel). Dynamic genes were first hierarchically clustered in parental cells to cluster groups of genes that behave similarly in pseudotime, then plotted in the same order in cKO cells. (*D*) Velocity gene expression analysis of parental (left panel) and cKO (right panel) cells at differentiation day 6. Direction of arrows point toward clusters with gene expression closest to the future state of each cell based on unspliced/spliced transcripts. Arrow length represents magnitude.

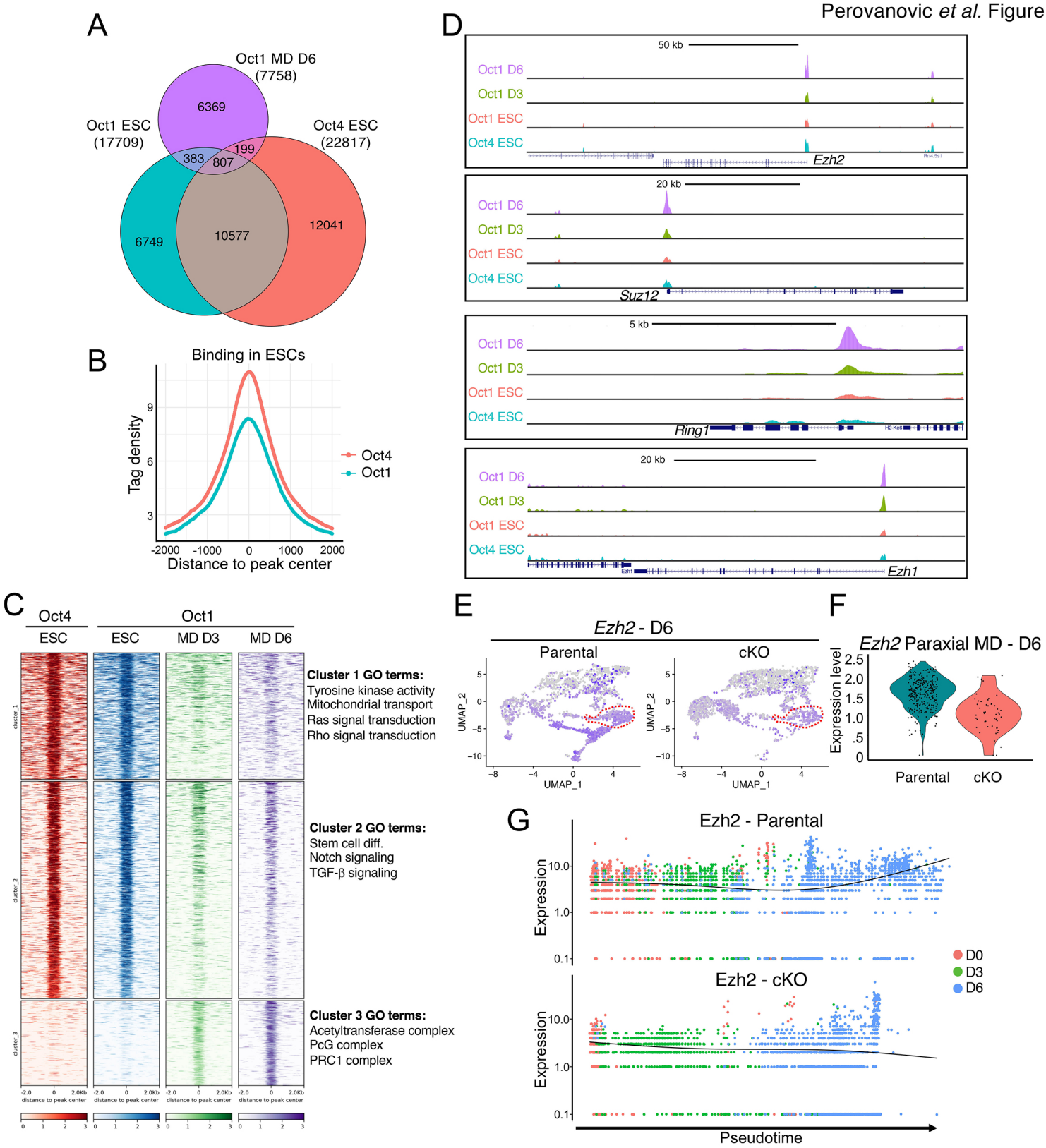

Figure 3. Oct1 co-occupies target sites with Oct4 in ESCs, and regulates their expression during differentiation. (*A*) Venn diagram showing common and unique binding sites based on Oct4 and Oct1 ChIP-Seq in ESCs, d3 and d6 of mesodermal differentiation shows shared peaks between different conditions. (*B*) Tag density around peak center in Oct4 (red) and Oct1 (blue) ChIPseq data sets. (*C*) A matrix of ChIPseq enrichment 2 kb around peak centers was computed for the merged peak list and shown as heatmap. Loci in red show high enrichment and white loci show no enrichment. GO terms corresponding to each cluster are shown on the right. (*D*) Example tracks of representative loci are shown. Y-axes were scaled to the same value for each gene. (*E*) *Ezh2* gene expression is shown in UMAP projections for parental and cKO ESCs at D6 of MD differentiation. The paraxial MD cluster is outlined in red. (*F*) Violin plots showing *Ezh2* gene expression in cells within the paraxial MD cluster shown in (*E*). (*G*) *Ezh2* expression in pseudotime in parental (top panel) and cKO (bottom panel) cells. Black trendline represents average expression across all the cells at any given position in pseudotime.

Perovanovic *et al.* Figure 4

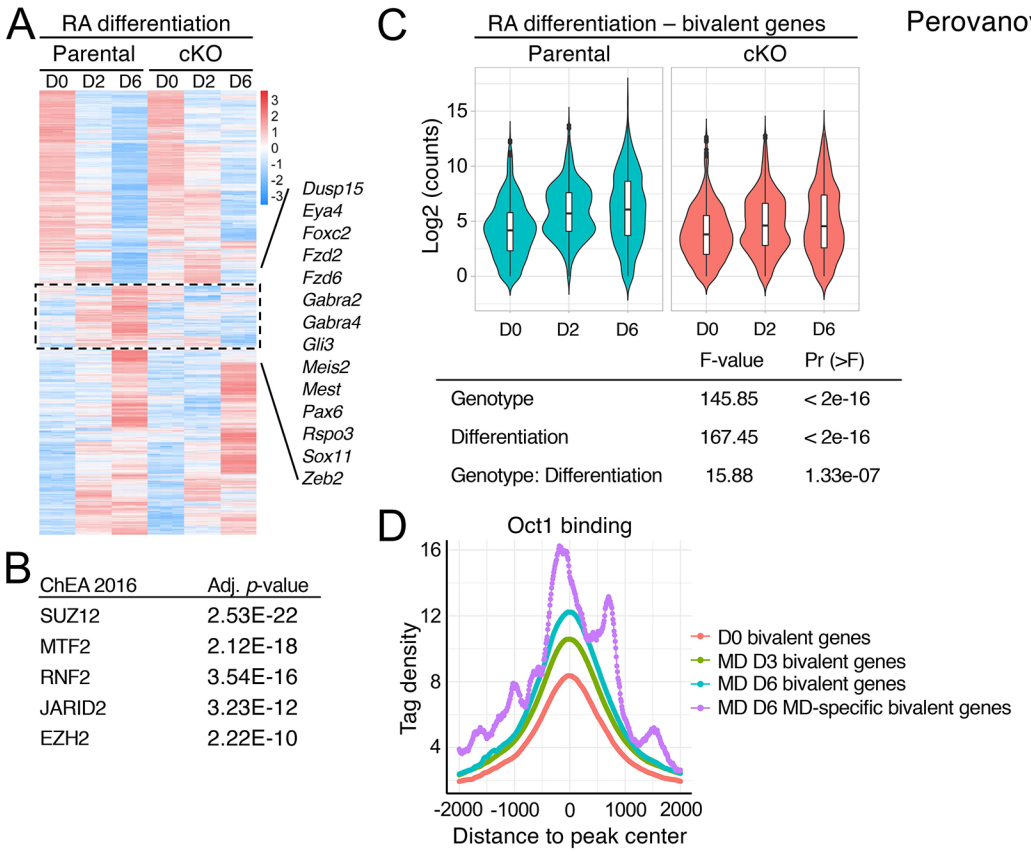

Figure 4. Bivalent genes fail to upregulated in differentiating Oct1 conditional knockout ESCs. (*A*) Heatmap of D6 differential genes (parental vs. cKO) in all time points is shown. Genes in red show high expression and blue genes show low expression/no expression. Representative genes that correspond to the middle cluster are shown on the right. (*B*) ChIP-X Enrichment Analysis (ChEA) query results for middle cluster in (*A*) are shown. (*C*) Expression of annotated bivalent genes is shown using violin plots for parental (blue) and cKO (red) ESCs at RA differentiation days 0, 2 and 6. Table below shows results of two-way ANOVA test comparing the two genotypes during differentiation. (*D*) Oct1 enrichment based on tag density around peak center at annotated bivalent genes in ESCs (red), and at MD differentiation D2 (green) and D6 (blue). An additional analysis was performed for MD-specific bivalent genes at MD differentiation D6 (purple) are shown.

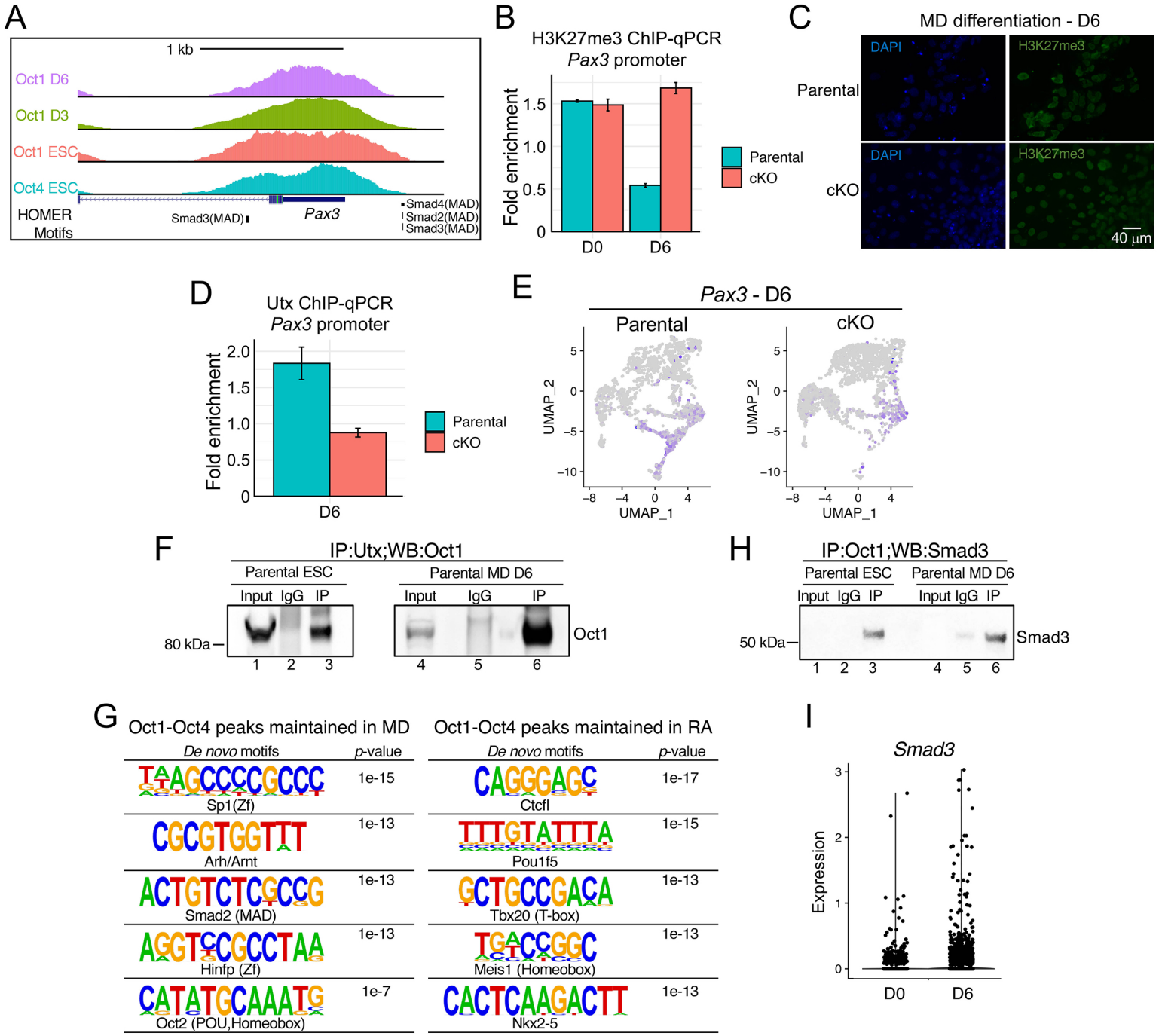

Figure 5. Oct1 interacts with Utx to demethylate H3K27me3 at bivalent genes. (*A*) Signal tracks (*Mm10* v. D191020) showing Oct1 and Oct4 enrichment at the *Pax3* 5' region. Positions of identified HOMER Smad2/3/4 motifs are shown below. (*B*) H3K27me3 ChIP-qPCR enrichment at the *Pax3* promoter in parental and cKO ESCs and at MD differentiation D0 and D6. Enrichment is normalized to a nonspecific genomic region encoding the 28S ribosomal subunit that is set to 1. (*C*) H3K27me3 immunofluorescence images from parental and cKO cultures differentiated for 6 d. Images were collected a 40× magnification. (*D*) Utx enrichment at the *Pax3* promoter in parental and cKO cells on D6 of differentiation. (*E*) UMAP projection of scRNA-seq data using parental and cKO cells at MD D6, highlighting *Pax3* expression. (*F*) Immunoprecipitation of Utx followed by Oct1 Western blot using parental ESCs, ESCs differentiated for 6 d under mesoderm-promoting conditions and ESCs differentiated using RA for 6 d. (*G*) HOMER motif analysis of Oct1 peaks that shared with Oct4 and maintained after 6 d of differentiation towards the mesoderm lineage (left panel) or using RA differentiation (right panel). (*H*) Smad3 immunoblot using cell lysates immunoprecipitated with Oct1 antibodies, or rabbit IgG controls. Parental ESCs (lanes 1-3) and MD D6-differentiated cells (lanes 4-6) were used. 5% input is shown, though Smad3 was undetectable with these amounts. (*I*) *Smad3* gene expression is shown in violin plots for parental ESCs at D6 of MD differentiation.

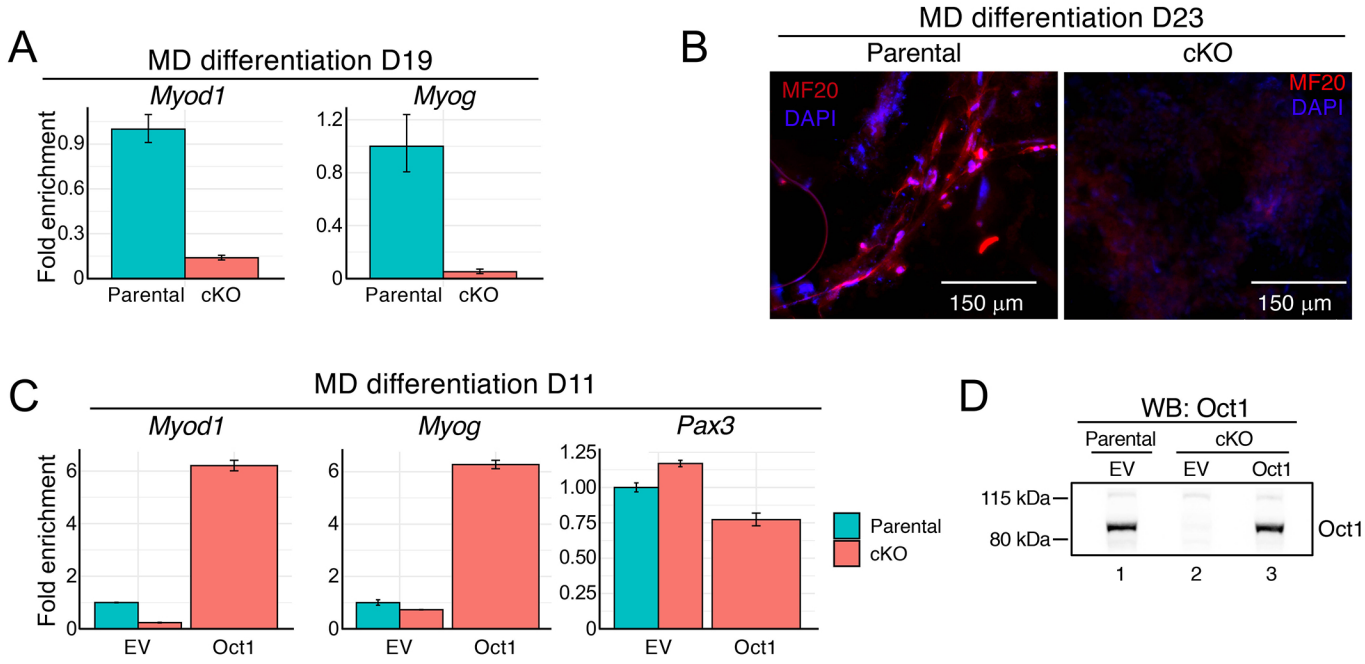

Figure 6. Myogenic lineage induction is defective in the absence of Oct1 and augmented with Oct1 ectopic expression. (*A*) Relative mRNA expression levels of myogenic genes *Myod1* and *Myog* in parental and cKO differentiated for 19 d. (*B*) MF20 staining showing Myh1e expression in parental (left panel) and cKO (right panel) cells at day 23 of MD differentiation. (*C*) RT-qPCR for the myogenic genes *Myod1*, *Myog* and *Pax3* in parental and cKO cells transduced with retroviruses encoding Oct1, or empty vector controls. Transduced cells were selected using puromycin for 48 hr prior to differentiation for 11 d. (*D*) Western blot showing successful ectopic Oct1 expression in cKO cells.