Targeted DamID reveals differential binding of mammalian pluripotency factors

Running title: Mammalian Targeted DamID

Seth W. Cheetham[1,2†], Wolfram H. Gruhn[1†], Jelle van den Ameele[1†], Robert Krautz[1†], Tony D Southall[1, 3], Toshihiro Kobayashi[1,4], M. Azim Surani[1] and Andrea H. Brand[1]*

[1.]The Gurdon Institute and Department of Physiology, Development and Neuroscience, University of Cambridge, Tennis Court Road, Cambridge UK.

[2.]Present address: Mater Research Institute-University of Queensland, Woolloongabba QLD 4102, Australia.

[3.]Present address: Department of Life Sciences, Imperial College London, Sir Ernst Chain Building, London, UK.

[4.]Present address: National Institute for Physiological Sciences, Center for Genetic Analysis of Behavior, Japan.

*Corresponding author, email: a.brand@gurdon.cam.ac.uk

[†]These authors contributed equally to this work.

Abstract

The precise control of gene expression by transcription factor networks is critical to organismal development. The predominant approach for mapping transcription factor-chromatin interactions has been chromatin immunoprecipitation (ChIP). However, ChIP requires a large number of homogeneous cells and antisera with high specificity. A second approach, DamID, has the drawback that high levels of Dam methylase are toxic. Here we modify our Targeted DamID approach (TaDa) to enable cell type-specific expression in mammalian systems, generating an inducible system (mammalian TaDa or MaTaDa) to identify protein/DNA interactions in 100 to 1000 times fewer cells than ChIP. We mapped the binding sites of key pluripotency factors, OCT4 and PRDM14, in mouse embryonic stem cells, epiblast-like cells and primordial germ cell-like cells (PGCLCs). PGCLCs are an important system to elucidate primordial germ cell development in mice. We monitored PRDM14 binding during the specification of PGCLCs, identifying direct targets of PRDM14 that are key to understanding its critical role in PGCLC development. We show that MaTaDa is a sensitive and accurate method to assess cell type specific transcription factor binding in limited numbers of cells.

Introduction

Chromatin immunoprecipitation (ChIP) has been widely used to characterise transcription factor-chromatin interactions (Furey, 2012), but this approach is limited by the requirement for large numbers of homogeneous cell populations and specific antibodies (Tsankov et al., 2015). As a result, mapping transcription-factor occupancy *in vivo* in rare cell types, such as stem cells, is technically challenging. DNA adenine methylation identification (DamID) has recently emerged as an alternative approach for genome-wide profiling (Aughey and Southall, 2016; Marshall and Brand, 2015; Marshall et al., 2016; Otsuki et al., 2014; Southall et al., 2013; van Steensel and Henikoff, 2000). In DamID, a DNA- or chromatin-binding protein is fused to an *E. coli* Dam methylase. Wherever the Dam-fusion protein binds to DNA or chromatin, it methylates adenine within the sequence GATC. Endogenous adenine methylation is extremely rare in eukaryotes (Koziol et al., 2015; Wu et al., 2016; Zhang et al., 2015) such that the tagged sequences can be detected easily by digestion with DpnI, which only cuts at methylated GATC sites. In this way binding sites can be identified genome-wide without cell isolation, fixation or immunoprecipitation.

Although DamID is particularly well suited for *in vivo* analysis, a major caveat is cytotoxicity resulting from high levels of expression of the Dam methylase (Southall et al., 2013; Catherine Davidson and A.H.B., data not shown). As a result, DamID in mammalian cells has generally relied on low level, ubiquitous expression from an uninduced heat-shock promoter (van Steensel and Henikoff, 2000; Vogel et al., 2007). However, this precludes the identification of cell-type-specific binding or detection of dynamic changes in DNA- or chromatin-interactions. To overcome these limitations, we modified Targeted DamID (TaDa) (Southall et al., 2013) for use in mammalian cells, enabling the rapid, accurate and sensitive identification of transcription factor binding sites. Mammalian TaDa (MaTaDa) enables genome-wide protein-DNA interaction profiling in a temporally regulated, cell type-specific

fashion.

First, we validated MaTaDa in murine embryonic stem cells by reanalysing OCT4 occupancy during the transition from the naïve to primed pluripotent cell state (Buecker et al., 2014) Next, we mapped the binding sites of the transcription factor PRDM14, which controls embryonic stem cell (ESC) pluripotency and is of pivotal importance for acquisition of primordial germ cell (PGC) fate in mice . Making use of *in vitro* specification of PGC-like cells (PGCLCs), we identified a set of novel cis-regulatory elements bound by PRDM14 specifically in PGCLCs. Analysis of these loci suggests that PRDM14 is involved in the suppression of EGFR/MAPK signalling and the regulation of genes associated with cell migration.

A mammalian system for Targeted DamID

We engineered a construct for conditional expression of Dam-fusion proteins in mammalian cells, comprising a ubiquitous promoter (PGK) driving expression of a transcript encoding a primary open reading frame encoding mCherry (ORF1 246 amino acids). The primary ORF is followed by two TAA stop codons and a single nucleotide frameshift upstream of a secondary open reading frame encoding the Dam fusion protein (ORF2; **Fig 1A**). We showed previously that translation of this bicistronic mRNA results in expression of ORF1 followed by rare ribosomal re-entry and translational re-initiation, resulting in extremely low levels of expression of ORF2, the Dam fusion protein (Southall et al., 2013).

For spatial and temporal control of MaTaDa, we inserted a GFP or puromycin coding sequence and SV40 terminator, flanked by loxP sites, between the PGK promoter and the TaDa construct (**Fig 1A, Fig S2A**). To excise the floxed cassette, we used a Cre-estrogen-receptor fusion (Cre-ER) **(Fig 1A, Fig S2A)**. Cre-ER is constitutively expressed, but only translocates to the nucleus and induces Dam expression upon tamoxifen administration **(Fig**

1B, C). In the absence of Cre, GFP or puromycin are expressed and transcription is terminated upstream of ORF1. The expression of GFP or puromycin can be used to assess transfection efficiency or select transfected cells. Excision of the stop-cassette results in loss of GFP or puromycin expression and induction of TaDa. In this way, MaTaDa enables both spatial and temporal control, directed by targeted expression of Cre, in either cell lines or transgenic animals. The low levels of Dam methylase expression are non-toxic and preclude dominant effects that might result from the overexpression of transcription factors. To lessen the potential for steric effects between the Dam methylase and fused proteins (Ramialison et al., 2017) we inserted a myc tag as a spacer.

Mapping binding sites of pluripotency factors with MaTaDa

During early mammalian development pluripotent stem cells have the potential to form all cell types of the embryo. Naïve and primed pluripotent states have been characterised in mouse based on the functional, transcriptional and epigenetic characteristics of the pre- and post-implantation epiblast (Nichols and Smith, 2009). Mouse embryonic stem cells (mESCs) are used extensively as an *in vitro* model to study the molecular mechanisms of pluripotency. In the presence of small molecules that promote Wnt/β-catenin and inhibit FGF/MAPK signalling mESCs remain in a naïve pluripotent state similar to the pre-implantation epiblast (Ying et al., 2008) **(Fig S1A)**. Stimulation of the FGF signalling pathway promotes differentiation of ESCs into epiblast like cells (EpiLCs), a primed pluripotency state. The unifying mechanistic features of pluripotency are key transcription factors like OCT4 (POU5F1) and PRDM14. OCT4 is a master regulator of both primed and naïve pluripotency in ESCs (Zeineddine et al., 2014). During the transition from naïve to primed pluripotency, the OCT4 binding pattern changes dynamically due to the availability of cofactors (Buecker et al., 2014). PRDM14, which can function as a transcriptional repressor (Nady et al., 2015;

Yamaji et al., 2013) promotes naïve pluripotency in mESCs (Ma et al., 2011; Yamaji et al., 2013). PRDM14 is also critical for the development of primordial germ cells (Yamaji et al., 2008).

To assess whether MaTaDa can detect differential binding of transcription factors we generated stable mESC lines carrying MaTaDa constructs encoding either Dam alone or a Dam-transcription factor fusion protein, together with the CreER-expression construct with a Zeocine resistance gene (**Fig 1A, Fig S2A**). Cells were selected for GFP expression or puromycin resistance and co-selected with Zeocine and expanded. Induction of CreER with tamoxifen for 24 hours resulted in induction of Dam expression and the loss of GFP fluorescence from virtually all cells (97.7%) 48 hours later (**Fig 1D,E**). Robust methylation of genomic DNA was detected in treated cells as compared to untransfected or EtOH-treated cells (**Fig 1F)**. Some faint DNA amplification was observed in cells treated with EtOH (**Fig 1F**), possibly due to low level expression before tamoxifen treatment and high sensitivity of the DamID technique. However, the considerably higher induction upon tamoxifen treatment allowed us to identify changes in transcription factor binding patterns during differentiation. Importantly, tamoxifen treatment of OCT4-Dam or PRDM14-DAM expressing ESCs did not lead to a large increase in OCT4 or PRDM14 expression and consequently there was no effect on pluripotency or differentiation (**Fig S1B, Fig S2C)**.

MaTaDa identifies genome-wide transcription factor occupancy with high accuracy and sensitivity

We sequenced DamID libraries from 150,000 mESCs expressing either Dam alone (control), Dam-OCT4 or Dam-PRDM14. OCT4 bound to 18,103 sites, and PRDM14 to 8784 sites when a cut-off of $q<10^{-25}$ was used (**Fig S3**). MaTaDa peaks were consistently detected between biological replicates (**Fig S3A-D)**, and genome-wide correlations between replicate

experiments are shown in **Fig S3E,F**. We detected binding to key OCT4 targets, such as *Nanog*, *Sox2*, *Klf4* and *Myc* **(Fig 2A)**. Similarly, PRDM14 was found to bind key genes involved in pluripotency, including *OCT4* and *Nanog*, and differentiation, such as *Fgfr1* and *Xist* **(Fig S4A)**.

Next we compared our data for OCT4 and PRDM14 binding in 150,000 mESCs to published ChIP-seq data from 50 and 20 million mESCs, respectively (Buecker et al., 2014; Ma et al., 2011). Importantly, due to the nature of DamID in acquiring counts only at GATC-motifs in the genome, the resolution of signal across the genome is different from ChIP-seq. In addition, random methylation requires stringent normalization to the Dam-only control, and a ratio is calculated between Dam-fusion and Dam only. This prevents direct comparisons between MaTaDa and ChIP-seq in a quantitative manner. We therefore used a peak-centered analysis for genome-wide comparisons between both techniques. Both the OCT4 and PRDM14 MaTaDa signals were highly enriched over ChIP peaks (**Fig 2B, Fig S4B**). Conversely ChIP-seq signal is highly enriched over MaTaDa peaks (**Fig 2C, Fig S4C)**. Overlap between peaks was dependent on the stringency of peak calling (**Fig 2D-G, Fig S4D-F**), but at $q<10^{-25}$, 1901 of 3880 (49%) OCT4 ChIP peaks overlapped with MaTaDa peaks (18096 peaks at $q<10^{-25}$) **(Fig 2B-G, Fig S5A-C)**, as did 1824 of 5681(32%) of PRDM14 ChIP-seq peaks (8784 MaTaDa peaks at $q<10^{-25}$) (**Fig S4B-J**) (Ma et al., 2011). Nevertheless, at any given q-value, a subset of peaks was always specific to either technique (**Fig 2F-G, Fig S4E-F, Fig S5A-E**). Interestingly, while both for ChIP-seq and for DamID it is generally thought that peak intensity grossly correlates with binding strength, the correlation (R^2) for peak intensity of peaks common to both techniques (at $q<10^{-25}$) was only 0.07 for OCT4 and 0.12 for PRDM14 (**Fig 2H, Fig S4H**).

We next conducted motif and genomic feature enrichment analysis (Imrichová et al., 2015) on the OCT4 MaTaDa peaks ($q<10^{-25}$). The three most highly ranked transcription factor motifs for which a position weight matrix was available all corresponded to OCT4-related motifs (**Fig 2I**). Interestingly, presence of OCT4-motifs at any given q-value was higher under common and ChIP-specific than under MaTaDa-specific peaks (**Fig S5F**). This could suggest that MaTaDa captures more indirect OCT4 binding events than ChIP-seq in these conditions. Peaks were also enriched for enhancers and active chromatin, as illustrated by the enrichment for ESC DNAse-accessible sites and H3K27ac and H3K4me1 histone marks (**Fig 2J**). The enrichment for genomic features did not change considerably when either MaTaDa or ChIP-seq specific peaks were analyzed (**Fig S5G,H**).

We conclude that MaTaDa was able to profile genome-wide transcription factor occupancy accurately and with high sensitivity, and can function as an alternative or complementary approach to ChIP-seq.

MaTaDa is sufficiently sensitive to profile rare populations of cells

A major advantage of TaDa over ChIP is the ability to profile rare populations of cells *in vivo* (Otsuki et al., 2014; Southall et al., 2013). To test the sensitivity of MaTaDa, we tested whether MaTaDa could profile binding in only 10,000 cells. Remarkably, the binding profiles for PRDM14 were strikingly similar to those obtained from 150,000 cells (**Fig 3A-C**). However, peak-calling at any given q-value always resulted in more peaks being called from 10,000 than 150,000 cells (**Fig 3B-E,G**), due to a lower signal-to-noise ratio, as expected from the low cell number. Nevertheless, the sensitivity of this low-cell number MaTaDa was very high, with nearly all peaks from 150,000 cells ($q<10^{-100}$) eventually being recovered by PRDM14 MaTaDa on 10,000 cells (**Fig 3F**). The ability to profile transcription factor

183 binding in small numbers of cells suggests that MaTaDa has sufficient sensitivity to uncover
184 transcription factor-genome interactions in rare cell types *in vivo*.

185 **MaTaDa captures cell-type specific transcription factor binding**

186 We tested whether MaTaDa could profile differential transcription factor occupancy between
187 different but related cell types by generating OCT4 binding profiles during ESC
188 differentiation. Removal of 2i and LIF and addition of FGF2 and Activin drives the transition
189 of naïve ESCs to epiblast-like stem cells (EpiLCs) (Hayashi et al., 2011). EpiLCs are
190 analogous to the cells of the post-implantation epiblast and are in a primed pluripotent state.
191 During the transition from ESCs to EpiLCs, OCT4 interacts with OTX2, and together they
192 bind a distinct set of enhancers to promote the activation of pro-differentiation genes
193 (Buecker et al., 2014).

194 First, we generated genome-wide MaTaDa profiles of OCT4 occupancy in naïve ESCs and
195 EpiLCs **(Fig 4)**. Next, we compared our data to a previously defined set of binding sites that
196 are bound by OCT4 predominantly in either naïve ESCs or EpiLCs (Buecker et al., 2014).
197 We found that ground-state pluripotency genes, such as *Klf4* **(Fig 4A),** were bound primarily
198 in ESCs while pro-differentiation genes, such as *Fgf5* **(Fig 4B),** were bound exclusively in
199 EpiLCs, but not in the ESCs from which they were derived. Critically, ESC-specific
200 enhancers were not strongly bound following differentiation **(Fig 4C)**, demonstrating that
201 MaTaDa is able to capture differential transcription factor occupancy, enabling the detection
202 of spatially and temporally restricted protein-chromatin interactions.

203 **PRDM14 binding in ESCs and PGCLCs**

204 In addition to controlling ESC pluripotency, PRDM14 is essential for specifying PGCs from
205 post-implantation epiblast cells (Yamaji et al., 2013). Paralleling mouse embryonic
206 development, the establishment of a primed pluripotent state in EpiLCs is required for

specification of PGCLCs *in vitro* (Hayashi et al., 2011; Ohinata et al., 2009). Notably, PGCLCs are functionally equivalent to the PGCs found in the early mouse embryo and hence represent an important system for studying mammalian germ line development (Hayashi et al., 2011). PGC identity is controlled by a transcription factor network consisting of BLIMP1, PRDM14 and TFAp2c, which suppresses expression of somatic genes and promotes transcription of germ cell genes (Magnúsdóttir et al., 2013; Nakaki et al., 2013). Expression of PRDM14 was shown to be sufficient to drive differentiation of EpiLCs into PGCLCs (Nakaki et al., 2013). However, deriving sufficient quantities of PGCLCs is difficult and has limited the mechanistic understanding of this pivotal developmental process. In particular, it has not been possible to determine how PRDM14 binding changes during PGCLC development and the key PRDM14 targets that drive the EpiLC-PGCLC transformation have not been identified previously.

We induced differentiation of EpiLCs into PGCLs using the growth factors BMP4, BMP8, SCF, EGF and LIF. PGCLCs (~10-15% of all cells, up to 150,000 per replicate) were then FACS-purified using endogenous Dppa3-GFP and Esg1-tdTomato reporters (**Materials and Methods, Fig S6).** Using MaTaDa we monitored PRDM14 binding in the first 3 days of PGCLC specification (**Fig S6A-D**) and found genome-wide changes in PRDM14 binding (**Figure 5A-C**). While 77% of binding sites ($q<10^{-100}$) were shared between ESCs and PGCLCs, the level of PRDM14 occupancy was strikingly different. 2450 sites were preferentially bound by PRDM14 in ESCs (>2 fold higher in ESCs) and 698 in PGCLCs (>2 fold higher in PGCLC), suggesting that PRDM14 regulates a functionally distinct set of genes in ESCs and PGCLCs (**Fig 5C**). Most PRDM14 peaks occur >2kb from transcriptional start sites (**Fig S7A-C)** at distinct sites in ESCs and PGCLCs (**Fig S7C).** Chen et al. (2012) defined 1277 putative enhancers in ESCs, based upon chromatin marks and transcription factor occupancy. Interestingly, we found that ESC-enriched PRDM14 sites coincided with

these presumptive regulatory regions, but PGCLC-enriched sites did not (**Fig 5D)**. The PGCLC-enriched sites were enriched for the PRDM14 core motif (GGTCTCTAA; p=4.39e-6). Interestingly, by *de novo* analysis we discovered another motif (**Fig S7E**; E=2.1e-033) that is similar to motifs recognised by RXRG (E=4.80e-05), the pluripotency factors NR5A2 (E=1.52e-4) and NR6A1 (E=4.16e-02) indicating that PGCLC-enriched sites may be regulatory regions bound by several pluripotency associated factors. A second motif (**Fig S7E**; E=2.0e-002) similar to the SOX motif is present at all 698 binding sites. Although SOX proteins have been shown to recognize similar motifs, they become restricted in their binding patterns through interactions with specific co-factors (Hou et al., 2017; Kondoh and Kamachi, 2010). Our data suggest that PRDM14 binds at novel, previously unidentified, PGCLC-specific enhancers.

Next we analysed genes associated with PRDM14 binding in ESCs and PGCLCs. We found that ESC-enriched PRDM14 target genes are implicated in the regulation of embryonic development and negative regulation of cell differentiation (**Fig S7D**) which includes genes differentially expressed in PRDM14 mutants, like *Fgfr1*, *Fgfr2* and *Dnmt3b* (Costello et al., 2011; Grabole et al., 2013).

By comparing the transcriptomes of EpiLCs and PGCLCs (Sasaki et al., 2015), we found that 445/2889 differentially expressed genes (15%) are direct targets of PRDM14 (p<1e-16, **Fig 5E-F**). Key PGC specification genes including *Tfap2c, Dppa3, Nr5a2 and Esrrb* were both bound by PRDM14 and upregulated in PGCLCs, while EpiLC genes including *Wnt8a, Otx2, Pou3f1 and Dnmt3a* were downregulated. PRDM14 may thus play a key role in PGC specification by upregulating key reprogramming genes and repressing EpiLC genes.

PGCLC-enriched PRDM14 targets also included genes involved in EGFR and MAPK signalling **(Fig S7D**). Notably, inhibition of the MAPK pathway is sufficient to upregulate key PGC markers context dependently (Kimura et al., 2014).

Most intriguingly, in both ESCs and PGCLCs PRDM14 binding is enriched in the vicinity of genes that function in neuronal development, cell migration and cell morphology **(Fig S7D)**. Consistent with this finding, genes involved in neurogenesis become induced upon depletion of PRDM14 in ESCs (Yamaji et al., 2013). In the developing mouse embryo PGCs migrate from the area of specification into the endoderm epithelium of hindgut and colonize the developing genital ridge at E10.5 (Anderson et al., 2000; Clark and Eddy, 1975). To investigate a potential function of PRDM14 in PGC migration we focused on *Wnt5a* and *Tnc*, which are significantly bound by PRDM14 in PGCLCs (**Fig 5A, Fig S8A**) and have been implicated in cell migration (Chawengsaksophak et al., 2012; Nishio et al., 2005). Expression of these genes is extremely low in ESCs and EpiLCs (**Fig S8B, C**). Upon PGCLC induction, Wnt5a and Tnc expression become induced only in those cells that fail to acquire the PGC identity, while PRDM14-expressing PGCLCs (**Fig S7D**) repress both migration associated genes. In contrast, the Wnt5a receptors Ror2 and Fzd5 are not repressed in specified PGCLCs (**Fig S8E-F**) (Ishikawa et al., 2001; Niehrs, 2012). Together, our results suggest that PRDM14 plays a direct role in controlling PGCLC migration towards the genital ridge.

Discussion

The study of genome-wide interactions of transcription factors and chromatin has been largely dominated by a single biochemical technique, ChIP. ChIP-seq experiments typically require millions of cells, precluding the identification of transcription factor targets in small populations of cells, including stem cells *in vivo*. In addition, the accuracy in detection of binding sites, as determined by ChIP, is difficult to assess due to the paucity of alternative

techniques. Here, we have developed mammalian Targeted DamID (MaTaDa), which enables transcription factor occupancy to be profiled with high sensitivity in a temporally and spatially controlled manner. MaTaDa overcomes the potential toxicity associated with expression of high levels of Dam methylase and avoids potential artefacts caused by overexpression of a transcription factor. Notably, expression of Dam-PRDM14 in PGCLCs did not result in adverse effects on cell growth or PGCLC specification efficiency compared with parental cells or cell lines expressing Dam alone (**Fig S2B-D**). A key finding is that MaTaDa can reveal the binding sites of master regulators of pluripotency in as few as 10,000 cells and potentially even fewer. While we used pure cell populations in this study experiment we have previously shown that given specific transgene expression TaDa profiles can be generated from a tiny proportion of cells of interest in complex tissues (Southall et al., 2013). This will allow MaTaDa to be applied *in vivo* to assess chromatin occupancy in rare and previously inaccessible cell populations.

Despite the overall similarity of binding profiles, MaTaDa and ChIP-seq results differed in notable ways. Many binding sites were detected using one method but not the other, despite the use of identical cell lines and culture conditions (**Fig 2, FigS4, FigS5**). In this respect it was striking that the degree of overlap between ChIP-seq peaks and peaks obtained by MaTaDa was similar for OCT4 and PRDM14. This indicates that the incongruities between MaTaDa and ChIP might derive from fundamental differences between the two techniques. Distinguishing which approach is a more accurate reflection of the binding of a transcription factor is not straightforward. Indeed, no studies have thus far systematically analysed the similarities between alternative methods for determining the genome-wide occupancy of transcription factors. Although ChIP signals were strongly enriched over MaTaDa peaks **(Fig 2 and FigS4)**, the intensity of peaks did not always correlate (**Fig 2H, Fig S4H**). An underlying assumption of ChIP is that regions that are most strongly detected are most

commonly or strongly bound and are thus “key targets”. However, additional factors including binding kinetics, steric effects of fusion proteins (Ramialison et al., 2017), accessibility of antibody-targeted epitopes and amplification, cross-linking or sonication bias could all affect the observed MaTaDa and ChIP signal intensities (Meyer and Liu, 2014). MaTaDa and ChIP may therefore represent complementary approaches for understanding transcription factor-genome interactions.

We designed MaTaDa to take advantage of the large collection of Cre-expressing constructs, cell-lines and model organisms for targeted expression *in vivo*. As has been demonstrated in *Drosophila* (Cheetham and Brand, 2018; Marshall and Brand, 2017; Southall et al., 2013), the generation of MaTaDa transgenic animals for key transcription factors, chromatin complexes and RNA polymerase II will permit the characterisation of the molecular landscape of gene regulation in almost any cell-type. Analysing the interactions between transcription factors and enhancers in small and pure populations of stem cells *in vivo* will be of vital importance for an increased understanding of the transcriptional control of development. While we observe some leakiness *in vitro* (**Fig 1F**), our system can clearly identify rearrangements in transcription factor occupancy.

Taking advantage of the high sensitivity of MaTaDa we were able for the first time to monitor PRDM14-chromatin association during the course of PGCLC specification. PRDM14 DNA binding in PGCLCs and ESCs differs significantly in location and intensity, which may be a consequence of cofactor availability or the distinct epigenetic state of each cell type. While 2i ESCs reside in an epigenetic state characterised by low abundance of repressive chromatin marks, such as DNA methylation or H3K9me2, PGCLCs are specified from EpiLCs, which are associated with elevated levels of repressive chromatin modifications (Leitch et al., 2013; Zylicz et al., 2015). Hence, the epigenetic environment

established in EpiLCs could potentially restrict PRDM14 binding during early PGCLC specification. We found that 15% of genes differentially expressed in the transition between EpiLCs and PGCLCs are direct PRDM14 targets. Interestingly, PRDM14 may function not only as a repressor but also as an activator, as it was bound at both up- and down-regulated genes. Interestingly, PGCLC-enriched PRDM14 binding sites did not correspond to predicted ESC enhancers (Chen et al., 2012) and may identify novel PGCLC-specific enhancers. Several of these presumptive enhancers regulate components of the MAPK pathway. In mice, PGCs are specified in the proximity of cells undergoing mesodermal differentiation. Consequently, inductive signals like Wnt and BMP that initiate PGC specification in the postimplantation embryo are also involved in defining the mesoderm linage (Behringer et al., 1999; Winnier et al., 1995). Further, inhibition of the MAPK pathway during mesodermal differentiation results in the upregulation of PGC marker genes (Kimura et al., 2014). This suggests that PRDM14 functions during early murine PGC specification by inhibiting the MAPK signalling pathway and thereby prevents establishment of the mesodermal cell fate.

Interestingly, we find that PRDM14 binds in the vicinity of genes associated with cell migration (**Fig S7D**) such as *Wnt5a* and *Tnc*, which are most significantly bound by PRDM14 in PGCLCs (**Fig 5A and S8A**). Wnt5a and its receptor Ror2 function in PGC migration. Loss of Wnt5a signalling strongly impairs PGC migration to the genital ridge (Chawengsaksophak et al., 2012; Laird et al., 2011). Here we find that Wnt5a and Tnc but not Wnt5a receptors are repressed in PRDM14-expressing PGCLCs, which suggests that Wnt5a is secreted from somatic cells to promote directed PGC migration, while Wnt5a repression in PGCs may prevent autocrine stimulation

Ectopic expression of PRDM14 has been linked to several types of cancer, such as lymphatic leukaemia, lung carcinoma and most prominently breast cancer (Dettman et al., 2011;

Nishikawa et al., 2007; Taniguchi et al., 2017; Zhang et al., 2013). A comprehensive understanding of Wnt5a function in breast cancer remains elusive, however there is evidence that decreased Wnt5a expression in these tumours is associated with a poorer prognosis (Zeng et al., 2016). Hence, a link between PRMD14, Wnt5a and cell migration might be of clinical relevance. We conclude that MaTaDa holds great promise for the *in vivo* analysis of transcription factor and chromatin protein interactions during development and disease.

Materials and Methods

Embryonic stem (ES) cell culture

E14tg2a ES cells were cultured in Serum/ LIF medium for maintenance (Ser/LIF medium: GMEM (Invitrogen; 21710-025), 10% FBS (Invitrogen 10270-106), 1% Non Essential Amino Acid (Invitrogen; 11140), 1mM Sodium Pyruvate (Invitrogen; 1130-070), 2mM L-Glutamine (Invitrogen; 25030-024), 1% Penicillin-Streptomycin(Invitrogen; 15140-22) 0.2% 2-mercaptoethanol (Invitrogen; 21985-023) and 0.1% LIF (obtained from CSCR Cambridge). Cells were grown on gelatine-coated cell culture dishes (Thermo-Fisher) and passaged by dissociating to ESC colonies with TrypLE (Invitrogen 12604-021). For experiments ESCs were grown in 2i/LIF medium (N2B27 medium (R&D), 1 μM PD0325901, 3 μM CHIR99021 (Stemgent) and 0.1% LIF (obtained from CSCR Cambridge)) on fibronectin-coated dishes (17 μg ml^{-1}; Millipore) for at least 4 passages.

Induction of Epiblast-like cells (EpiLCs) and primordial germ cell-like cells (PGCLCs)

EpiLCs were induced as described previously (Hayashi et al., 2011). In brief, 2i ESCs were differentiated into EpilCs by treatment with FGF2 and ActivinA for 40h. Subsequently, PGCLCs specification was induced by a cytokine cocktail consisting of BMP4 (0.5μg/ml), BMP8(0.5μg/ml), SCF (0.1μg/ml), EGF (0.05μg/ml) and LIF (1x; made by CSCR Cambridge) (**Fig S5A**). The induction of PGCLC specification by cytokines is inefficient and hence, FACS purification of successfully specified cells is required. Here, we made use of a dual reporter ESC line harbouring stable integrations of GFP and tdTomato in the endogenous Dppa3 and Esg1 loci, respectively (described in Hackett et al. 2017, manuscript under revision). While expression of Dppa3/GFP is used to identify specifying PGCLCs, high Esg1/tdTomato expression marks undesired cell types such as ESCs or EpiLCs.

piggyBac transposition

2.5ug of PBase, 0.5µg CreER plasmid, 2.5ug of MaTaDa plasmid was diluted in total 50ul of Gibco® Opti-MEM™ Media. 5 µl of Lipofectamine 2000 (Invitrogen) was diluted in 45µl of OptiMem and mixed with the plasmid solution and vortexed. The solution was incubated for 10 mins at room temperature and then added to E14 ESCs in culture. The cells were incubated for 4-6 hrs at 37^0C, 5%CO_2. The media was then changed. Transfection efficiency usually ranges between 3-30%. Cells were selected with 25µg/ml Zeocine for 7 days after transfection.

Flow cytometry

Embryoids were washed with PBS (Gibco), dissociated by incubation in 10 mM tissue culture grade EDTA (Invitrogen) for 3-5 min at 37°C and subjected to FACS using the Sony SH800S Cell Sorter. FACS data were analysed using the Flowjo software.

qRT-PCR

Total RNA was isolated from 20k to 200k cells using the RNeasy Mini Kit (Qiagen) including on column DNase digest. cDNA was generated using the SuperScript III Reverse Transcriptase kit (Thermo Fisher Scientific) and 250 ng random primer (Invitrogen) per reaction. The cDNA was quantified using the SYBR Select Master Mix (Applied Biosystems) and the QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). PCR reaction mix and qPCR program were prepared according to manufacturer's instructions.

MaTaDa constructs

To clone the maTaDa construct PGK-LGL-Dam, LT3-Dam-Myc was amplified and inserted into a vector with the PGK promoted driving expression of a floxed GFP cassette using Gibson assembly (**Fig S9A**). The OCT4 CDS was amplified from mESC cDNA and inserted into PGK-LGL-Dam using the restriction enzymes BglII and NotI (**Fig S9B**). PGK-LPL-Dam was constructed by replacing the floxed GFP cassette with a puromycin resistance cassette (**Fig S9C**). PRDM14 was amplified from mESC cDNA and inserted into PGK-LPL-Dam with BglII and NotI (**Fig S9D)**.

DamID-seq

DamID-seq was performed as described previously (Marshall et al., 2016). Briefly, cells were dissociated with TrypLE, washed and counted. gDNA was extracted using the Qiagen QIAamp® DNA Micro Kit. The DNA was then digested overnight at 37^0C with DpnI to cut methylated GATC sites (New England Biolabs) and purified with a QIAquick® PCR Purification Kit. Adaptors were blunt-end ligated for two hours at 16^0C using T4 DNA ligase (New England Biolabs) and heat inactivated at 65^0C for 20 minutes. The ligated DNA was then digested with DpnII to cleave any unmethylated GATC sites (New England Biolabs) and purified with a 1:1 ratio of Seramag beads. Adaptor-ligated fragments were then amplified with MyTaq^{TM} (Bioline) and PCR purified. The amplified DNA was then sonicated and digested with AlwI (New England Biolabs) to remove the adaptors, generating diverse DNA ends. The fragments were then prepared for Illumina sequencing according to the modified TruSeq protocol described in Marshall et al., 2017. All sequencing was performed as single end 50 bp reads generated by the Gurdon Institute NGS Core using an Illumina HiSeq 1500.

Published data acquisition

sra files were acquired from the Gene Expression Omnibus (Clough and Barrett, 2016) via wget (v1.17.1). sra-files were converted with fastq-dump (v2.3.5) to fastq-files or abi-dump (v2.3.5) to csfasta- & qual-files for colorspace data.

Quality check (QC)

Quality check was performed for all files individually with FastQC (v0.11.4). Residual adapter sequences and low quality bases were removed with cutadapt (v1.9.1) or TrimGalore (v0.4.5) when needed. Total and unique reads were summed to assess library size. The lengths of the reads was determined as additional quality check with awk (v4.1.3).

damidseq_pipeline

A file of GATC-sites for GRCm38 genome was generated as gff-file with gatc.track.maker.pl (see: https://github.com/AHBrand-Lab). Analysis of fastq-files from DamID-experiments was performed with the damidseq pipeline script (Marshall and Brand, 2015) that maps reads to an indexed bowtie2 genome (i.e., GRCm38), bins into GATC-fragments according to GATC-sites and normalises reads against a Dam-only control. Binding intensities were quantile normalised across all replicates (i.e., across all bedgraph-files) for the same experiment and subsequently averaged. Pearson correlation coefficients and R^2 values for comparisons of individual normalized replicates were calculated between pairs of bedgraph files in the RStudio environment with base functions (base, v3.4.3; RStudio, v1.1.423).

ChIP-seq mapping

Reads were mapped to the indexed mouse genome (mm10) with bowtie2 (v2.2.9) (Langmead and Salzberg, 2012) or optionally to the corresponding masked version including only major chromosomes to improve data quality. Resultant sam-files were converted to bam-files,

sorted and indexed with samtools (v1.3.1) (Li et al., 2009). Duplicates were removed with the MarkDuplicates picard tool (v.1.95) when needed. Total and unique mapped reads were counted with awk (v4.1.3) and bedtools (v2.25.0) (Quinlan and Hall, 2010).

Browser tracks and data visualisation

Reads were extended as well as binned and resulting tracks converted with the bamCoverage deeptools command (v3.0.2). Files were converted into bw-files with awk (v4.1.3) and bedGraphToBigWig (v4) or to tdf-files with the Integrative Genomic Viewer (IGV) (Robinson et al., 2011). Data was visualised using IGV, with the midline for MaTaDa ratio tracks set at 1, and for ChIP-seq at 0. Heatmaps were generated using Seqplots in the RStudio IDE (v1.12.0; Stempor and Ahringer, 2016).

ChIP-peak calling and quantification

Sorted bam-files for input (Buecker, C. et al., 2015) or HA-/EGFP-flag samples (Ma, Z. et al., 2011; Yamaji, M. et al., 2013) were merged with samtools (v1.3.1) to serve as combined control sample. Broad peaks were called with MACS2 (v2.1.0) for the individual bam-files in comparison to the combined control sample with the following specifications: --keep-dup all --bw 300 --qvalue 0.05 --mfold 5 50 --broad --broad-cutoff 0.1. Peaks were called on the individual bam-files for the experimental samples in comparison to the combined control sample. The number of significant peaks were read out at sequentially decreasing q-value (i.e., represented as “-log10(qvalue)” in line with MACS2), peaks in accessory contigs and mitochondrial genome were filtered out. Residual peaks sorted according to coordinates and data converted into bed-format. Reads were accumulated over peaks by intersecting bam-file derived read coordinates with peaks using bedtools (v2.25.0), summing the reads with awk (v4.1.3) and normalising them for library size and peak length. Data was prepared and plotted

with tidy tools in the RStudio IDE as violin and scatterplots (i.e., ggplot2, v2.2.1; tibble, v1.4.2; tidyr, v0.8.0; readr 1.1.1; purr, v0.2.4; dplyr, v0.7.4; stringr, v1.3.0).

DamID-peak calling

bam-files with extended reads for Dam only generated by the pipeline for every sample were merged and used as combined control for each individual Dam fusion sample during MACS2 (v2.1.0) peak calling. Additionally, peaks were called for a merged bam-file consisting of all Dam fusion samples in comparison to the merged bam-file for Dam only in line with peak calling for ChIP-samples (i.e., merge-vs-merge).

Consensus peaks

Peaks were defined as reproducible across all replicates at a given q-value, when overlapping peaks from these biological replicates were consistently identified in more than 50% (Yang et al., 2014) of all cases (including the merge-vs-merge, e.g., 2 of 3, 3 of 4). Coordinates of consensus peaks were defined as the maximum area covered by all overlapping peaks which prevents peak duplication.

DamID-peak quantification

DamID-binding intensities for identified peaks were aggregated by identifying all GATC-fragments overlapping with the area of the peak, trimming the first and the last fragment to peak coordinates and summing the weighted scores associated with the fragments. Data was analysed and visualized in accordance with the corresponding ChIP datasets in RStudio (see ‘ChIP-peak calling and quantification’).

Common peaks

Peaks shared amongst ChIP and DamID or between experiments done with the same technique were identified by intersecting the corresponding peak collections with bedtools (v2.25.0) intersectBed. Coordinates of common peaks were defined, deduplicated and sorted similar to the generation of consensus peaks. The extent of overlap between common peaks was evaluated depending on the q-value, which was either gradually changed for the sets of peaks from both techniques or by keeping the q-value of one set and changing the other. The latter allows to evaluate the recovery of peaks in the compared set despite differing significance. Similarly, the distributions of common and individual peaks were determined by identifying the closest peak from the compared dataset to the summits of the investigated peak set dependent on the q-value with the closest tool of the bedtools suite (v2.25.0). These distributions were plotted as densities of peak numbers dependent on the distance to the reference peak summits with ggridges (v0.5.0) in RStudio.

Annotation

Peaks were annotated to overlapping genomic features or nearest gene, respectively, (e.g., for intergenic/distal peaks) with the ChIPseeker-package (v1.10.3) in the R-environment using annotations from TxDb.Mmusculus.UCSC.mm10.knownGene (v3.4.0) and gene IDs from org.Mm.eg.db (v3.4.0).

RNASeq

cutadapt (v1.9.1) trimmed reads for RNA-seq from EpiLCs and PGCLCs (Sasaki et al., 2015) were reconverted from fastq- to csfasta- and qual-file formats and Phred+33-scores translated into numeric Q scores. Reads were aligned and assembled with tophat (v2.0.14) (Kim et al., 2013) which used bowtie1 (v1.1.2) to map colorspace data using --no-coverage-search. Differentially expressed genes were identified with cufflinks (v2.2.0; i.e., --compatible-hits-

norm --no-length-correction --library-type fr-secondstrand --max-mle-iterations 50000), cuffmerge (v1.0.0) and cuffdiff (v2.2.0; i.e., --compatible-hits-norm --no-length-correction --library-type fr-secondstrand --max-mle-iterations 50000 --frag-bias-correction --multi-read-correct) (Trapnell et al., 2012).

Significant genes were filtered by q-value ≤ 0.05 and a fold-change of ≥ 2 (i.e., EpiLC-vs-d4BVSC) and overlapped with the list of annotated peaks associated with intergenic regions to identify differentially expressed genes associated with putative enhancer peaks. Similar numbers of genes were randomly sampled without replacement 10000 times for empirically testing the enrichment of putative enhancer peaks with the associated genes by approximating a normal distribution.

Enrichment

Coordinates for promoters, exons, introns, 5' and 3'UTRs, as well as intergenic regions were deduced for transcript and exon annotations from biomaRt (v2.30.0). A list of high probability enhancers were derived from (Chen et al., 2012). Enrichment analysis for ESC- and PGCLC-specific as well as common peaks sets (i.e., -log10(q-value)≥100, FC≥2) with gene features and enhancers was performed with gat (v1.2.2) (Heger et al., 2013) using 100.000 sampling iterations and mm10 chromosomes as workspace. Fold enrichment of genomic features (i.e., enhancers) or log2 fold enrichment (i.e., gene features) for the associations were displayed together with their respective q-values.

Motif detection

Motifs were detected *de novo* using the MEME suite programs MEME and compared to known motifs using TOMTOM (Bailey et al., 2015). Enrichment of the PRDM14 motif was

556 detected using AME (Bailey et al., 2015). Transcription factor motifs and overlap with
557 chromatin marks as well as DNaseI Hypersensitivity sites in Oct4 peak sets were screened
558 with i-cisTarget (Imrichova et al., 2015).
559
560 **Data availability**
561 Generated datasets were deposited under the GEO accession GSE101971.
562

Acknowledgements

We would like to thank Kay Harnish for help with sequencing, and Catherine Davidson for technical assistance. This work was funded by Wellcome Trust Senior Investigator Award 103792, Wellcome Trust Program Grant 092545, and Royal Society Darwin Trust Research Professorship to AHB. S.W.C. was funded by a Herchel Smith Research Studentship. JvdA was supported by EMBO Long-term Fellowship ALTF 1600_2014 and Wellcome Trust Postdoctoral Training Fellowship for Clinicians 105839. T.K. was supported by Postdoctoral Fellowships for Research Abroad, The Uehara Memorial Foundation Research Fellowship and Kanae Foundation for the Promotion of Medical Science. MAS was supported by HFSP RGP0020/2012 and Wellcome Trust Senior Investigator Award 096738. A.H.B acknowledges core funding to the Gurdon Institute from the Wellcome Trust (092096) and CRUK (C6946/A14492).

Author Contributions

S.W.C, W.H.G, T.D.S, M.A.S and A.H.B designed the experiments. S.W.C, W.H.G, T.K and J.v.d.A performed the experiments. S.W.C, W.H.G, R.K and A.H.B analysed the data. S.W.C, W.H.G, J.v.d.A and A.H.B wrote the manuscript. All authors reviewed the manuscript prior to submission.

Conflicts of interest

The authors declare no conflicts of interest.

References

Anderson, R., Copeland, T. K., Schöler, H., Heasman, J. and Wylie, C. (2000). The onset of germ cell migration in the mouse embryo. *Mech. Dev.* **91**, 61–8.

Aughey, G. N. and Southall, T. D. (2016). Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 25–37.

Bailey, T. L., Johnson, J., Grant, C. E. and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* **43**, W39–W49.

Behringer, R. R., Bradley, A., Liu, P., Wakamiya, M., Shea, M. J. and Albrecht, U. (1999). Requirement for Wnt3 in vertebrate axis formation. *Nat. Genet.* **22**, 361–365.

Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T. and Wysocka, J. (2014). Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–53.

Chawengsaksophak, K., Svingen, T., Ng, E. T., Epp, T., Spiller, C. M., Clark, C., Cooper, H. and Koopman, P. (2012). Loss of Wnt5a disrupts primordial germ cell migration and male sexual development in mice. *Biol. Reprod.* **86**, 1–12.

Cheetham, S. W. and Brand, A. H. (2018). RNA-DamID reveals cell-type-specific binding of roX RNAs at chromatin-entry sites. *Nat. Struct. Mol. Biol.* **25**, 109–114.

Chen, C., Morris, Q. and Mitchell, J. A. (2012). Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. *BMC Genomics* **13**, 152.

Clark, J. M. and Eddy, E. M. (1975). Fine structural observations on the origin and associations of primordial germ cells of the mouse. *Dev. Biol.* **47**, 136–55.

Clough, E. and Barrett, T. (2016). The Gene Expression Omnibus Database. In *Methods in molecular biology (Clifton, N.J.)*, pp. 93–110.

Costello, I., Pimeisl, I.-M., Dräger, S., Bikoff, E. K., Robertson, E. J. and Arnold, S. J. (2011). The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. *Nat. Cell Biol.* **13**, 1084–91.

Dettman, E. J., Simko, S. J., Ayanga, B., Carofino, B. L., Margolin, J. F., Morse, H. C. and Justice, M. J. (2011). Prdm14 initiates lymphoblastic leukemia after expanding a population of cells resembling common lymphoid progenitors. *Oncogene* **30**, 2859–73.

Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat. Rev. Genet.* **13**, 840–852.

Grabole, N., Tischler, J., Hackett, J. A., Kim, S., Tang, F., Leitch, H. G., Magnúsdóttir, E. and Surani, M. A. (2013). Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. *EMBO Rep.* **14**, 629–37.

Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. and Saitou, M. (2011). Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–32.

Heger, A., Webber, C., Goodson, M., Ponting, C. P. and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics*

29, 2046–2048.

Hou, L., Srivastava, Y. and Jauch, R. (2017). Molecular basis for the genome engagement by Sox proteins. *Semin. Cell Dev. Biol.* **63**, 2–12.

Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D. and Aerts, S. (2015). I-cisTarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* **43**, W57–W64.

Ishikawa, T., Tamai, Y., Zorn, A. M., Yoshida, H., Seldin, M. F., Nishikawa, S. and Taketo, M. M. (2001). Mouse Wnt receptor gene Fzd5 is essential for yolk sac and placental angiogenesis. *Development* **128**, 25–33.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Kimura, T., Kaga, Y., Ohta, H., Odamoto, M., Sekita, Y., Li, K., Yamano, N., Fujikawa, K., Isotani, A., Sasaki, N., et al. (2014). Induction of primordial germ cell-like cells from mouse embryonic stem cells by ERK signal inhibition. *Stem Cells* **32**, 2668–78.

Kondoh, H. and Kamachi, Y. (2010). SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. *Int. J. Biochem. Cell Biol.* **42**, 391–9.

Koziol, M. J., Bradshaw, C. R., Allen, G. E., Costa, A. S. H., Frezza, C. and Gurdon, J. B. (2015). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat. Struct. Mol. Biol.* **23**, 24–30.

Laird, D. J., Altshuler-Keylin, S., Kissner, M. D., Zhou, X. and Anderson, K. V (2011). Ror2 enhances polarity and directional migration of primordial germ cells. *PLoS Genet.* **7**, e1002428.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359.

Leitch, H. G., McEwen, K. R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J. G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.* **20**, 311–316.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A. and Wysocka, J. (2011). Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.* **18**, 120–7.

Magnúsdóttir, E., Dietmann, S., Murakami, K., Günesdogan, U., Tang, F., Bao, S., Diamanti, E., Lao, K., Gottgens, B. and Azim Surani, M. (2013). A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat. Cell Biol.* **15**, 905–15.

Marshall, O. J. and Brand, A. H. (2015). Damidseq-pipeline: An automated pipeline for

processing DamID sequencing datasets. *Bioinformatics* **31**, 3371–3373.

Marshall, O. J. and Brand, A. H. (2017). Chromatin state changes during neural development revealed by in vivo cell-type specific profiling. *Nat. Commun.* **8**,.

Marshall, O. J., Southall, T. D., Cheetham, S. W. and Brand, A. H. (2016). Cell-type-specific profiling of protein–DNA interactions without cell isolation using targeted DamID with next-generation sequencing. *Nat. Protoc.* **11**, 1586–1598.

Meyer, C. A. and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **15**, 709–721.

Nady, N., Gupta, A., Ma, Z., Swigut, T., Koide, A., Koide, S. and Wysocka, J. (2015). ETO family protein Mtgr1 mediates Prdm14 functions in stem cell maintenance and primordial germ cell formation. *Elife* **4**,.

Nakaki, F., Hayashi, K., Ohta, H., Kurimoto, K., Yabuta, Y. and Saitou, M. (2013). Induction of mouse germ-cell fate by transcription factors in vitro. *Nature* **501**, 222–226.

Nichols, J. and Smith, A. (2009). Naive and Primed Pluripotent States. *Cell Stem Cell* **4**, 487–492.

Niehrs, C. (2012). The complex world of WNT receptor signalling. *Nat. Rev. Mol. Cell Biol.* **13**, 767–779.

Nishikawa, N., Toyota, M., Suzuki, H., Honma, T., Fujikane, T., Ohmura, T., Nishidate, T., Ohe-Toyota, M., Maruyama, R., Sonoda, T., et al. (2007). Gene amplification and overexpression of PRDM14 in breast cancers. *Cancer Res.* **67**, 9649–57.

Nishio, T., Kawaguchi, S., Yamamoto, M., Iseda, T., Kawasaki, T. and Hase, T. (2005). Tenascin-C regulates proliferation and migration of cultured astrocytes in a scratch wound assay. *Neuroscience* **132**, 87–102.

Ohinata, Y., Ohta, H., Shigeta, M., Yamanaka, K., Wakayama, T. and Saitou, M. (2009). A Signaling Principle for the Specification of the Germ Cell Lineage in Mice. *Cell* **137**, 571–584.

Otsuki, L., Cheetham, S. W. and Brand, A. H. (2014). Freedom of expression: cell-type-specific gene profiling. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 429–43.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Ramialison, M., Waardenberg, A. J., Schonrock, N., Doan, T., de Jong, D., Bouveret, R. and Harvey, R. P. (2017). Analysis of steric effects in DamID profiling of transcription factor target genes. *Genomics* **109**, 75–82.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011). Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–6.

Sasaki, K., Yokobayashi, S., Nakamura, T., Okamoto, I., Yabuta, Y., Kurimoto, K., Ohta, H., Moritoki, Y., Iwatani, C., Tsuchiya, H., et al. (2015). Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178–194.

Southall, T. D., Gold, K. S., Egger, B., Davidson, C. M., Caygill, E. E., Marshall, O. J. and Brand, A. H. (2013). Cell-type-specific profiling of gene expression and chromatin

binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. *Dev. Cell* **26**, 101–12.

Stempor, P. and Ahringer, J. (2016). SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Res.* **1**, 14.

Taniguchi, H., Hoshino, D., Moriya, C., Zembutsu, H., Nishiyama, N., Yamamoto, H., Kataoka, K. and Imai, K. (2017). Silencing PRDM14 expression by an innovative RNAi therapy inhibits stemness, tumorigenicity, and metastasis of breast cancer. *Oncotarget* **8**, 46856–46874.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., Gnirke, A. and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–9.

van Steensel, B. and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424–8.

Vogel, M. J., Peric-Hupkes, D. and van Steensel, B. (2007). Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nat. Protoc.* **2**, 1467–1478.

Winnier, G., Blessing, M., Labosky, P. A. and Hogan, B. L. (1995). Bone morphogenetic protein-4 is required for mesoderm formation and patterning in the mouse. *Genes Dev.* **9**, 2105–16.

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S. D., Mackintosh, S. G., Zhong, M., et al. (2016). DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333.

Yamaji, M., Seki, Y., Kurimoto, K., Yabuta, Y., Yuasa, M., Shigeta, M., Yamanaka, K., Ohinata, Y. and Saitou, M. (2008). Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.* **40**, 1016–22.

Yamaji, M., Ueda, J., Hayashi, K., Ohta, H., Yabuta, Y., Kurimoto, K., Nakato, R., Yamada, Y., Shirahige, K., Saitou, M., et al. (2013). PRDM14 Ensures Naive Pluripotency through Dual Regulation of Signaling and Epigenetic Pathways in Mouse Embryonic Stem Cells. *Cell Stem Cell* **12**, 368–382.

Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D. and McIntyre, L. M. (2014). Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput. Struct. Biotechnol. J.* **9**, e201401002.

Zeineddine, D., Hammoud, A. A., Mortada, M. and Boeuf, H. (2014). The Oct4 protein: more than a magic stemness marker. *Am. J. Stem Cells* **3**, 74–82.

Zeng, R., Huang, J., Zhong, M.-Z., Li, L., Yang, G., Liu, L., Wu, Y., Yao, X., Shi, J. and Wu, Z. (2016). Multiple Roles of WNT5A in Breast Cancer. *Med. Sci. Monit.* **22**, 5058–5067.

Zhang, T., Meng, L., Dong, W., Shen, H., Zhang, S., Liu, Q. and Du, J. (2013). High expression of PRDM14 correlates with cell differentiation and is a novel prognostic

marker in resected non-small cell lung cancer. *Med. Oncol.* **30**, 605.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., et al. (2015). N6-Methyladenine DNA Modification in Drosophila. *Cell* **161**, 893–906.

Zylicz, J. J., Dietmann, S., Günesdogan, U., Hackett, J. A., Cougot, D., Lee, C. and Surani, M. A. (2015). Chromatin dynamics and the role of G9a in gene regulation and enhancer silencing during early mouse development. *Elife* **4**,.

Figure legends

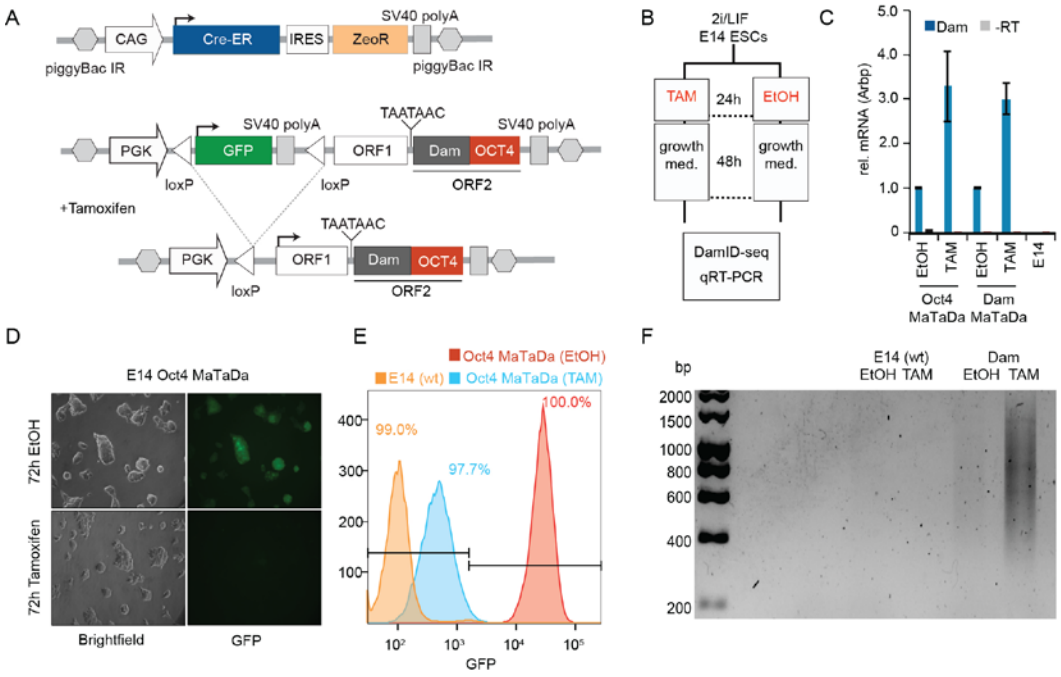


Figure 1. Conditional expression of Dam-OCT4 in ESCs

(**A**) The PGK promoter drives expression of a floxed GFP cassette (green). Upon Cre induction the floxed cassette is excised allowing expression of ORF1 (white; 246 amino acids). Rare translational re-initiation results in low-level expression of the Dam-OCT4 fusion protein (ORF2; grey/blue). Cre-ER is constitutively expressed and translocates to the nucleus upon tamoxifen treatment. **(B)** E14 mESCs transformed with Dam-alone or Dam-OCT4 MaTaDa constructs were treated with tamoxifen or EtOH for 24 hrs and then allowed to grow for 48hrs. The cells were then processed for DamID-seq and qPCR. **(C)** Induction of Dam transcription after tamoxifen treatment compared to EtOH control measured by qPCR. Error bars represent mean +/- sem. **(D)** Following tamoxifen treatment MaTaDa containing ESCs rapidly lose GFP fluorescence. **(E)** FACS analysis demonstrates the efficient loss of GFP fluorescence in MaTaDa expressing cells following Cre induction, with some perdurance of GFP protein at 72 hours resulting in higher levels of fluorescence compared to

774 the parental cell line (E14). **(F)** Amplification of mouse genomic DNA methylated by

775 MaTaDa after tamoxifen treatment.

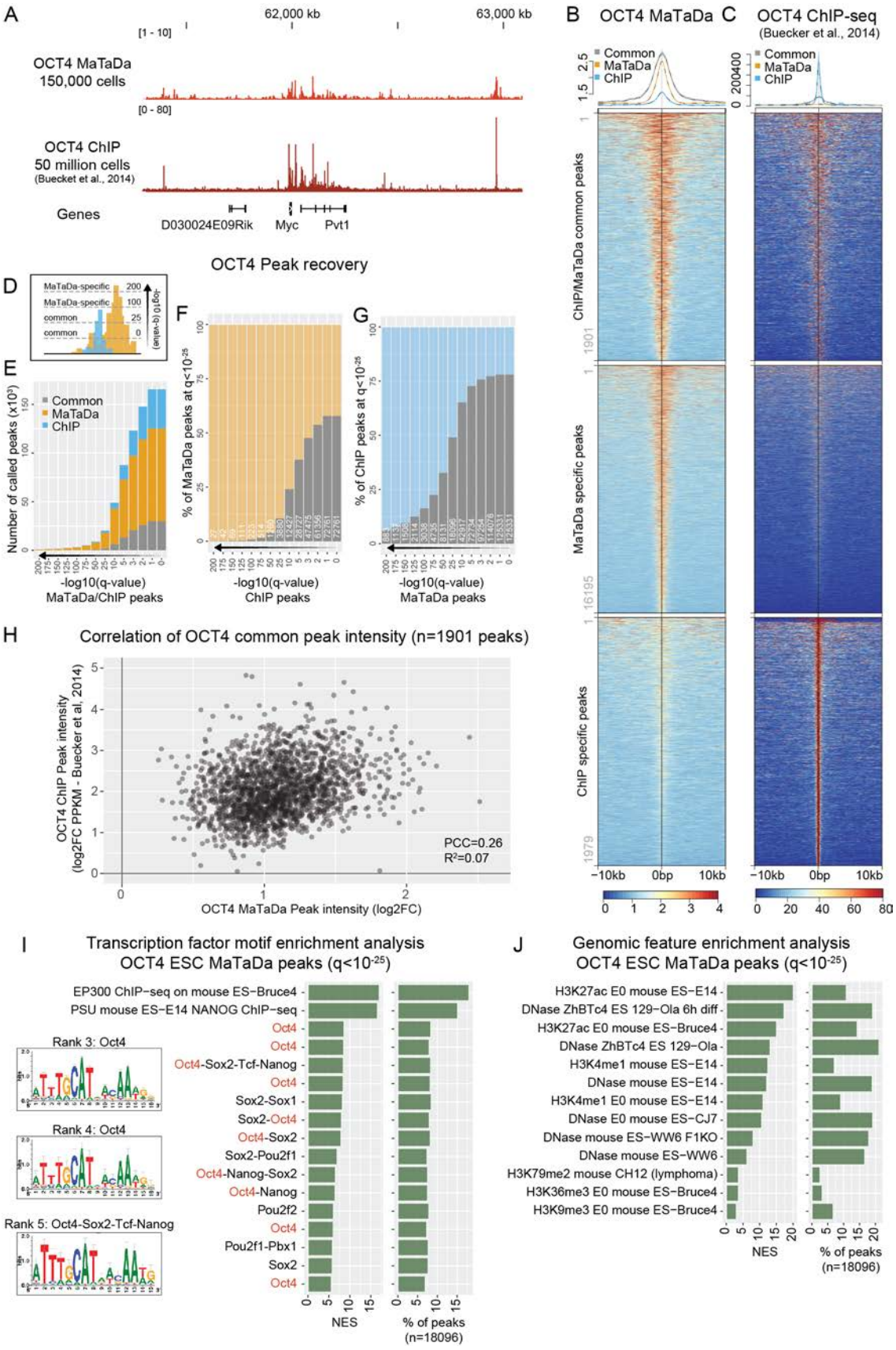
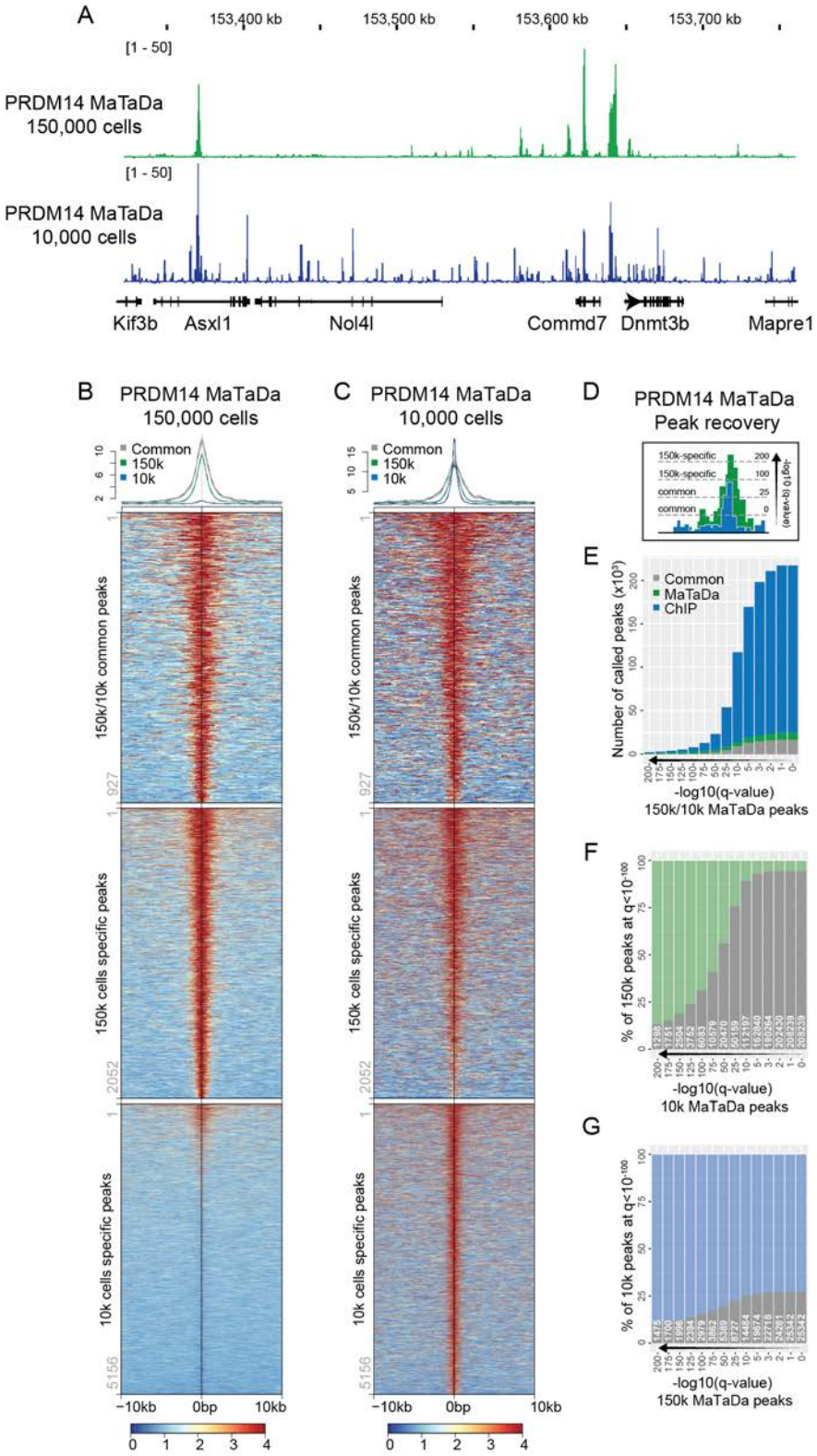


Figure 2. MaTaDa accurately profiles genome-wide transcription factor occupancy

(A) Genome browser view of OCT4 binding at the Myc locus (MaTaDa; average of three replicates) compared to ChIP. MaTaDa data are represented as fold enrichment of Dam-fusion over Dam only; ChIP-seq data are represented as aligned reads. (**B,C**) OCT4 MaTaDa (**B**) and ChIP-seq (**C**) ESC signal is plotted over a 10kb window either side of the peak midpoint, for peaks common to MaTaDa and ChIP-seq (top), specific to MaTaDa only (middle) and specific to ChIP-seq only (bottom) at $q<10^{-25}$. Above are metaplots of the MaTaDa (**B**) and ChIP-seq (**C**) signal. (**D**) Schematic to illustrate how peak recovery between two different datasets can vary depending on the q-value. (**E-G**) Number (**E**) or percentage (**F,G**) of peaks called upon changing the q-value for peak-detection, either for MaTaDa and ChIP-seq in parallel (**E**), or compared to a fixed q-value $<10^{-25}$ for MaTaDa (**F**) or ChIP-seq (**G**). Common peaks are grey, MaTaDa-specific peaks are orange, ChIP-seq specific peaks are blue. (**H**) Scatterplot of peak intensity for peaks ($q<10^{-25}$) common to OCT4 MaTaDa and ChIP-seq. PCC, Pearson correlation coefficient. (**I,J**) Transcription factor motif (**I**) and genomic feature (**J**) enrichment analysis of OCT4 MaTaDa peaks ($q<10^{-25}$, 18096 peaks). Position weight matrices (PWM) are shown for the top three enriched motifs for which a PWM was available. Normalized enrichment score (NES) and percentage of peaks containing the feature are indicated.



795

796 **Figure 3. MaTaDa has sufficient sensitivity to profile rare cell populations**

797 **(A)** Genome browser view of PRDM14 occupancy at the Dnmt3b locus from 10,000 (10k)

798 cells (average of 5 replicates) and 150,000 (150k) cells (average of 3 replicates). Data are

represented as fold enrichment of Dam-fusion over Dam only. (**B,C**) PRDM14 MaTaDa signal from 150,000 ESC (**B**) and 10,000 ESC (**C**) is plotted over a 10 kb window either side of the peak midpoint, for peaks common to 150k and 10k (top), specific to 150k only (middle) and specific to 10k only (bottom) at $q<10^{-100}$. Above are metaplots of the 150k (**B**) and 10k (**C**) signal. (**D**) Schematic to illustrate how peak recovery can vary depending on the q-value. (**E-G**) Number (**E**) or percentage (**F,G**) of peaks called upon changing the q-value for peak-detection, either for 150k and 10k ESC in parallel (**E**), or compared to a fixed q-value $<10^{-100}$ for 150k (**F**) or 10k (**G**). Common peaks are grey, 150k-specific peaks are green, 10k-specific peaks are blue.

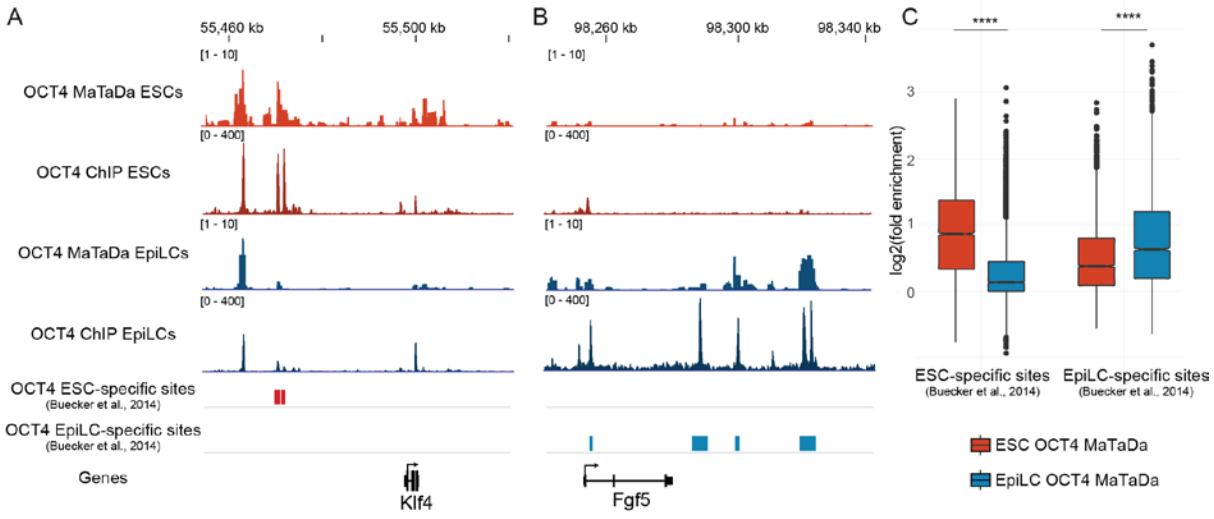
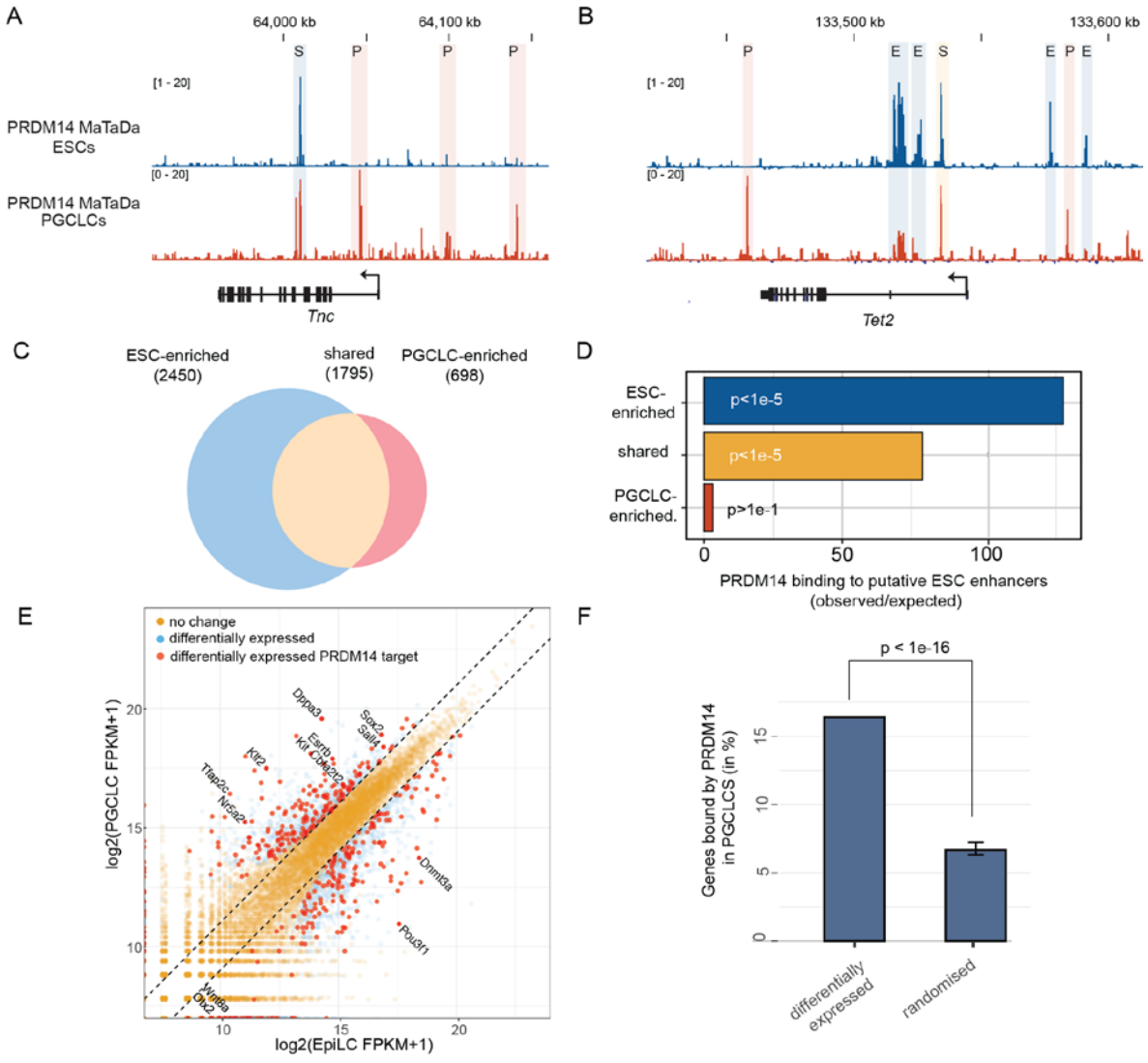


Figure 4. Differential binding of OCT4 in ESCs and EpiLCS as identified by MaTaDa

(A) Genome browser view of the ground state pluripotency gene Klf4, showing OCT4 binding to nearby enhancers in ESCs (average of three replicates) but not EpiLCs (average of two replicates). **(B)** Genome browser view of the pro-differentiation gene Fgf5, showing OCT4 binding to nearby enhancers in EpiLCs (average of two replicates) but not ESCs (average of three replicates). **(C)** Box plot demonstrating that MaTaDa recapitulates the dynamic binding of OCT4 to EpiLC and ESC specific sites. P-values <1e-5 unpaired,

817 unequal variance t test.

818



819

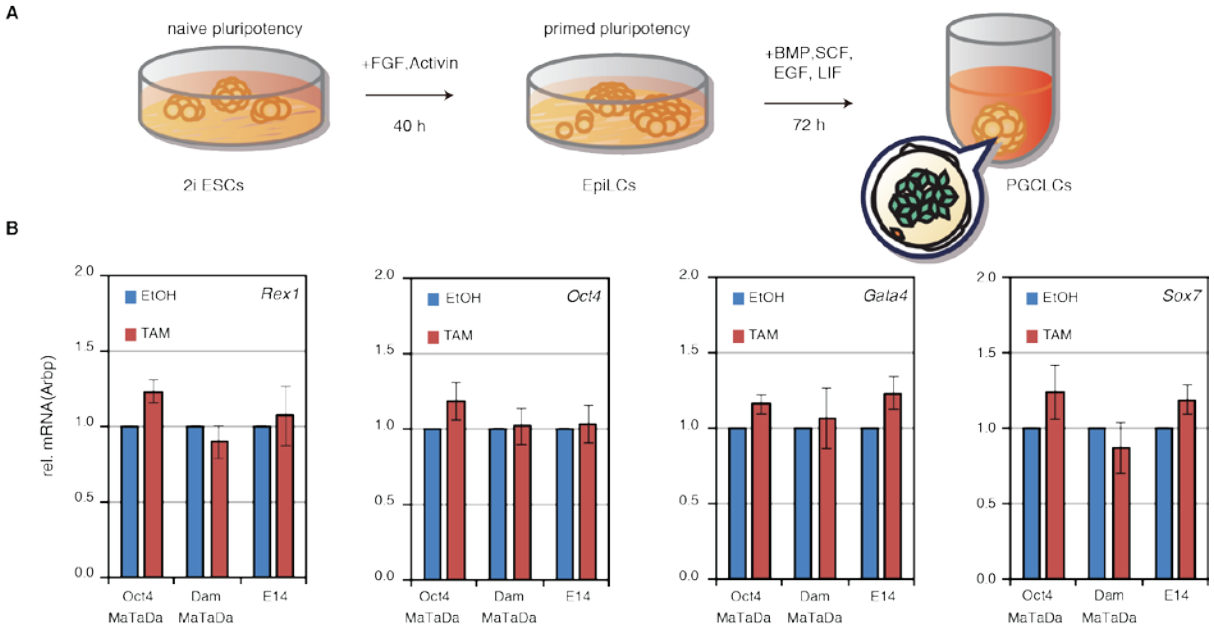
820 **Figure 5. Distinct chromatin association of PRDM14 in ESCs and PGCLCs**

821 (**A,B**) Genome browser view of PGCLC-enriched PRDM14 occupancy in ESCs and PGCLCs
822 (both average of three replicates) at the Tnc (A) and the Tet2 (B) loci. Shared (S), PGCLC-
823 enriched (Pe) and ES-enriched (Ee) sites are shown. (**C**) PRDM14 bound regions are
824 subdivided into ESC-enriched (>2 fold), shared in ESCs/PGCLCs, and PGCLC-enriched (>2
825 fold). (**D**) Overlap between ESC-defined enhancers and genomic loci occupied by PRDM14
826 in ESCs (ESC-enriched), PGCLCs (PGCLC-enriched) or both (shared). P-values are

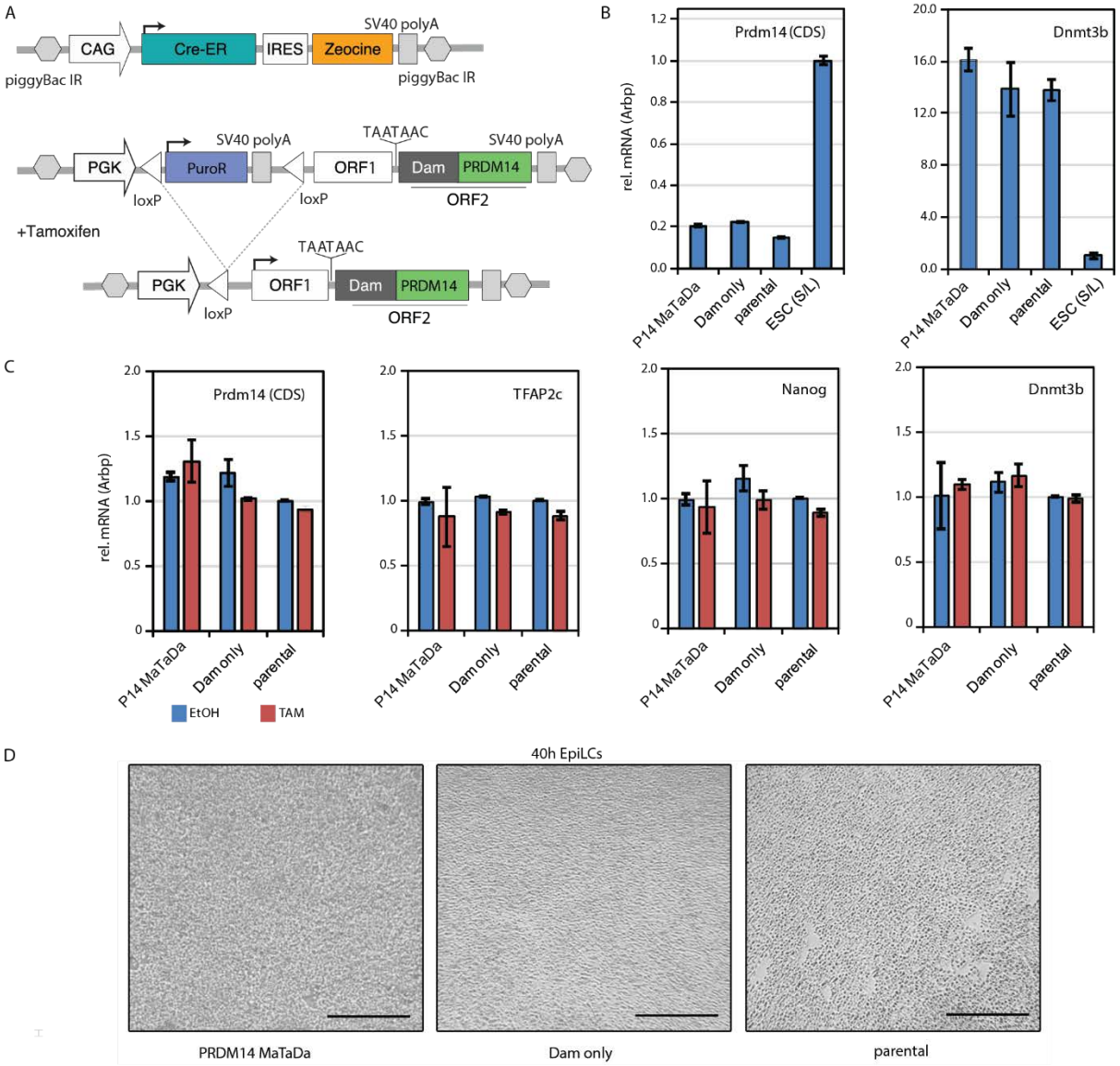
827 calculated by Genomic Association Test (Heger et al., 2013). (**E**) Comparison between genes
828 differentially expressed during PGCLC specification and genes bound by PRDM14 in
829 PGCLCs. Gene expression is plotted in log2 FPKM (fragments per kilobase mapped reads).
830 (**F)** A large proportion of PGCLC-PRDM14 targets are differentially expressed between
831 EpiLCs and PGCLCs. P values are calculated by an empirical test based on a normal
832 distribution.

833

Supplementary Figures

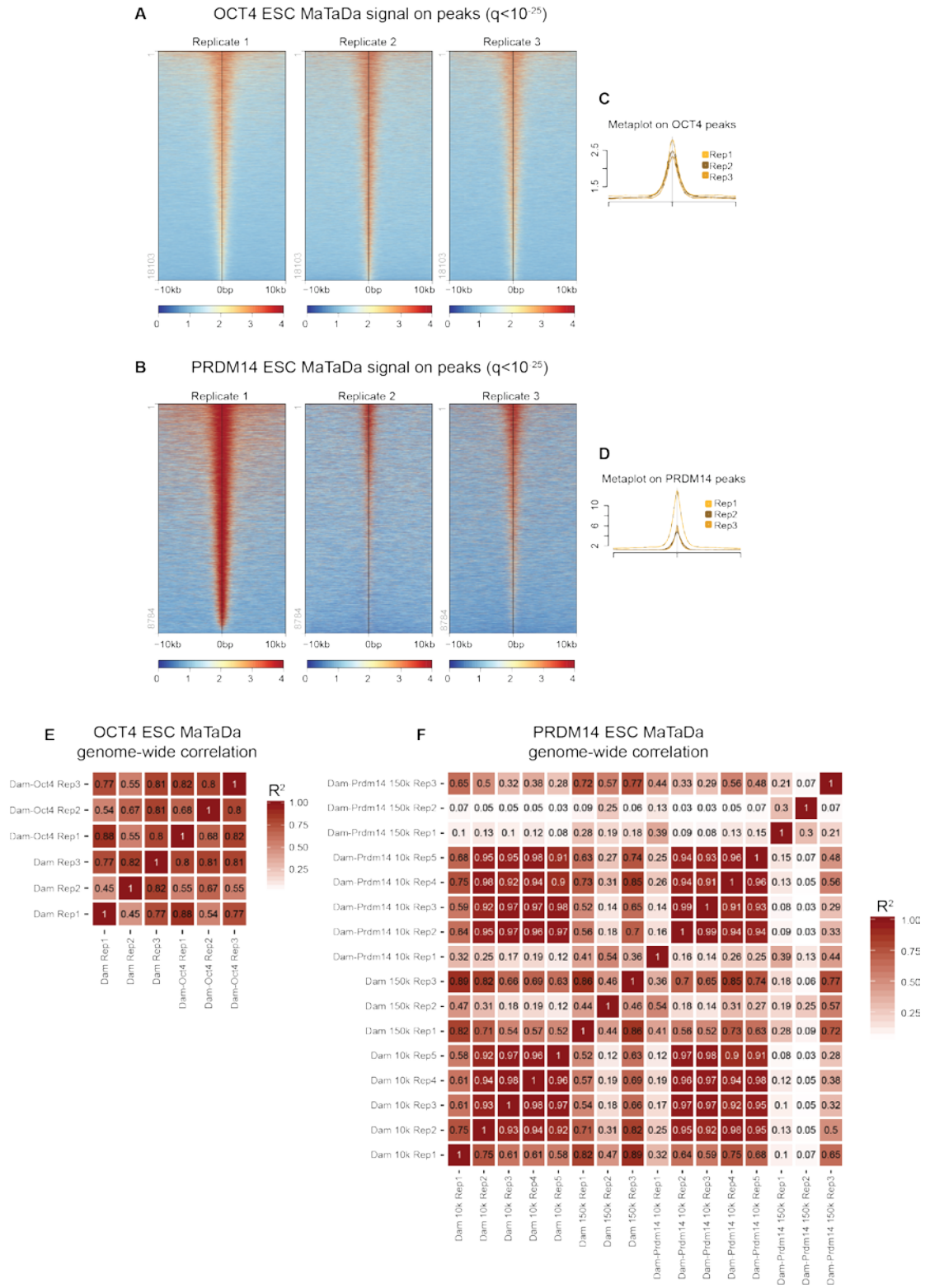


Supplementary Figure 1. MaTaDa does not affect cell fate
(**A**) ESCs cultures in 2i media remain in a naïve pluripotency. The addition of FGF and Activin drives differentiation into EpiLCs, a primed pluripotent state. Supplementation with BMP, SCF, EGF and LIF promotes the transition into primordial germ cells. **(B)** qPCR of key marker genes of pluripotency and endodermal fate reveals that OCT4 MaTaDa has no effect upon cell fate.



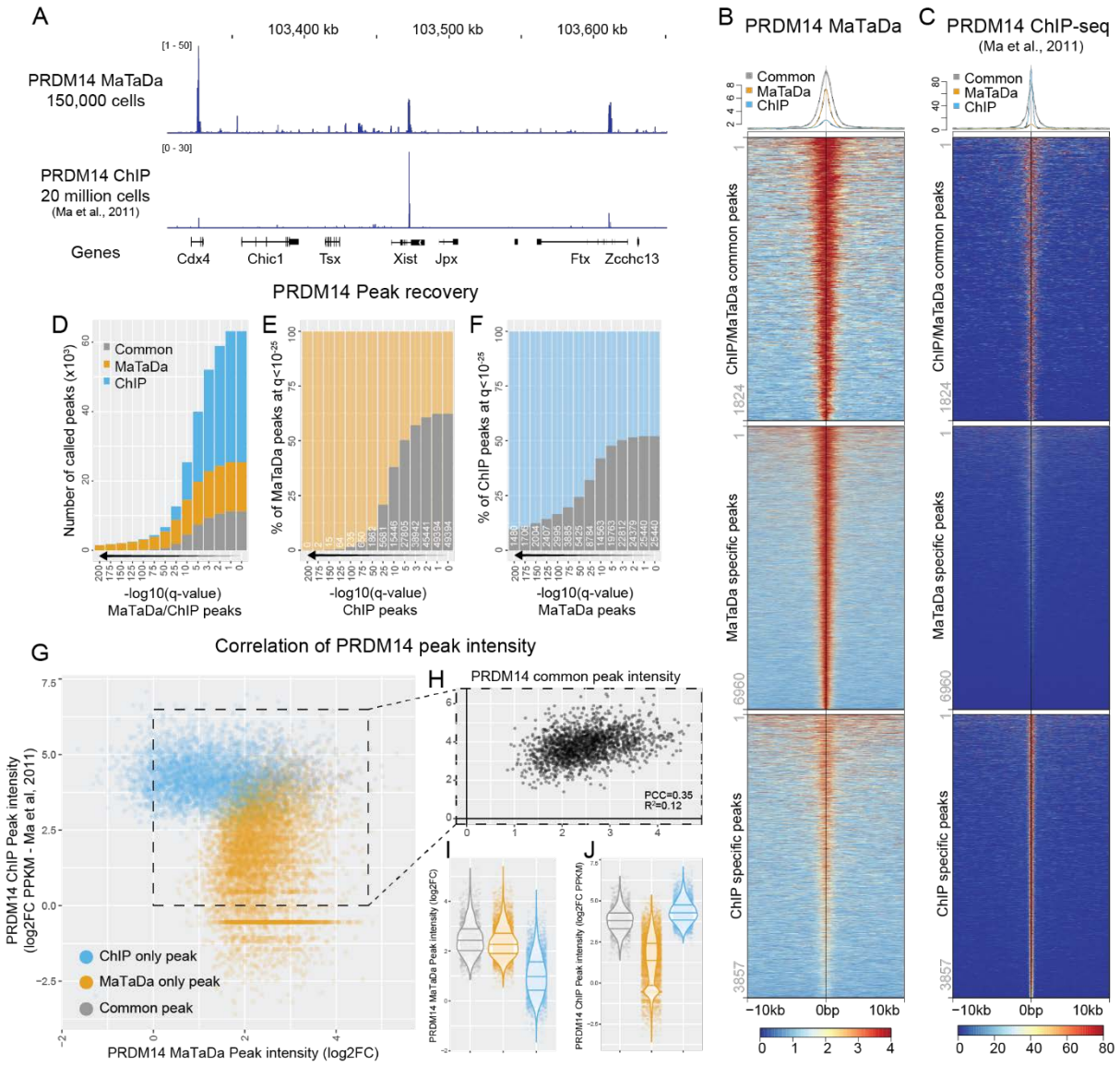
Supplementary Figure 2. Conditional expression of Dam-PRDM14 does not disrupt cellular behaviour

(**A**) Cre-ER is constitutively expressed and translocates to the nucleus upon tamoxifen treatment. The PGK promoter drives expression of a floxed puromycin resistance cassette (dark blue). Upon Cre induction the floxed cassette is excised allowing expression of ORF1 (246 amino acids). Rare translational re-initiation results in low-level expression of the Dam-PRDM14 fusion protein (ORF2; grey/green). **(B)** PRDM14 is not overexpressed and Dnmt3b remains high in EpiLCs following PRDM14 MaTaDa induction. **(C)** The expression of key marker genes and PRDM14 targets in mESCs are unaffected by PRDM14 MaTaDa. (**D**) Brightfield microscopy demonstrated that PRDM14 MaTaDa does not impact EpiLC differentiation.

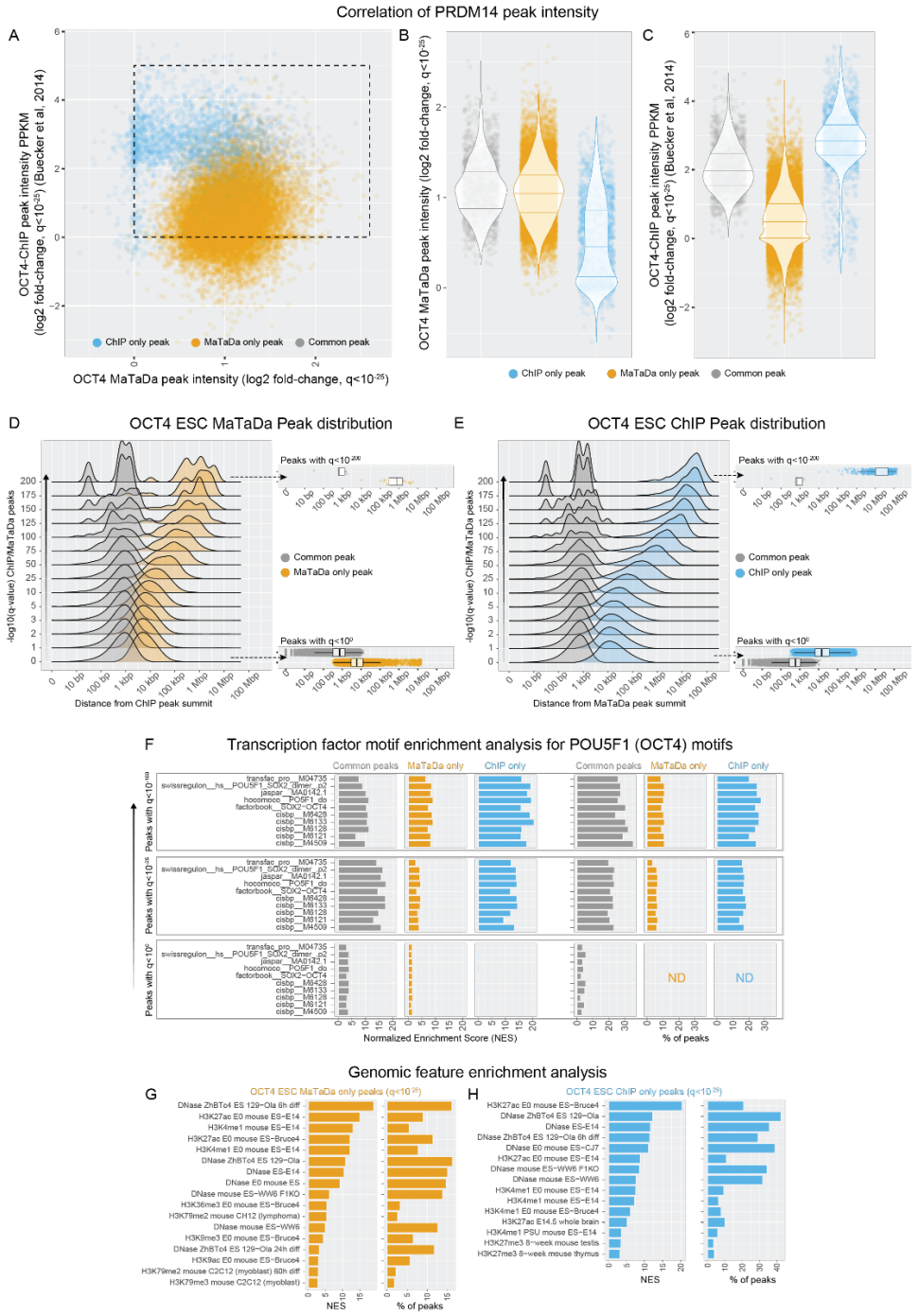


Supplementary Figure 3. MaTaDa Peaks are reproducible

(**A,B**) OCT4 (**A**) and PRDM14 (**B**) ESC MaTaDa peaks are consistently detected between biological replicates. Signal is plotted over a 10 kb window either side of the peak midpoint ($q<10^{-25}$) for each of the three replicates. (**C,D**) Metaplots of the three replicates of OCT4 (**C**) and PRDM14 (**D**) ESC MaTaDa signal over a 10kb window on either side of the peak midpoint. (**E,F**) Correlation matrix showing R^2 values for genome-wide correlation between OCT4 (**E**) and PRDM14 (**F**) ESC MaTaDa replicates.

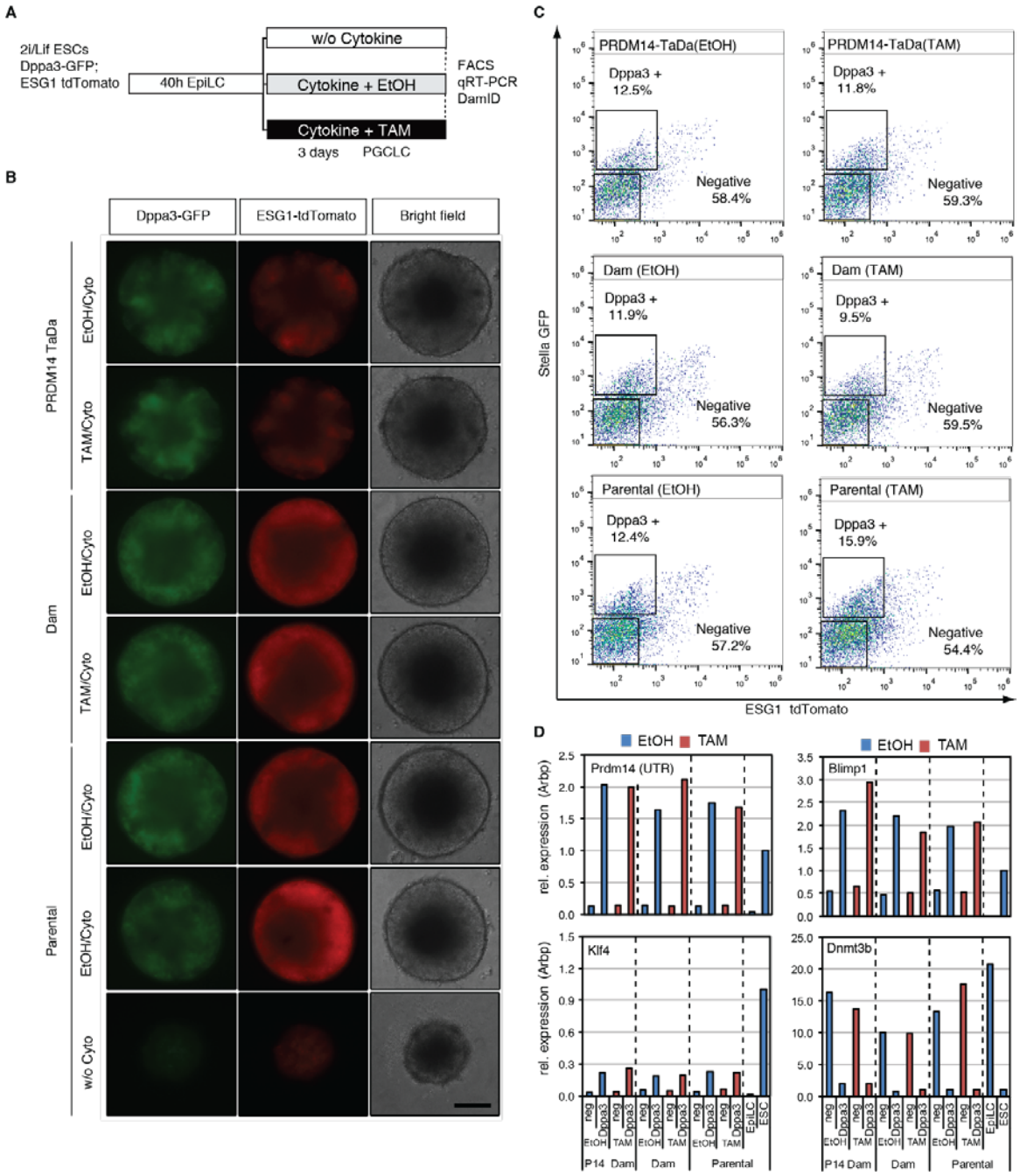


Supplementary figure 4. Correlation between Prdm14 MaTaDa and ChIP-seq in ESC. (**A,B**) PRDM14 MaTaDa (**A**) and ChIP-seq (**B**) ESC signal is plotted over a 10 kb window either side of the peak midpoint, for peaks common to MaTaDa and ChIP-seq (top), specific to MaTaDa only (middle) and specific to ChIP-seq only (bottom) at $q<10^{-25}$. Above are metaplots of the MaTaDa (**A**) and ChIP-seq (**B**) signal over a 10kb window on either side of the peak midpoint. (**C**) Genome browser view of PRDM14 binding at the Xist locus (MaTaDa; average of three replicates) compared to ChIP. MaTaDa data are represented as fold enrichment of Dam-fusion over Dam only; ChIP-seq data are represented as aligned reads. (**D-F**) Number (**D**) or percentage (**E,F**) of peaks called upon changing the q-value for peak-detection, either for MaTaDa and ChIP-seq in parallel (**D**), or compared to a fixed q-value $<10^{-25}$ for MaTaDa (**E**) or ChIP-seq (**F**). Common peaks are grey, MaTaDa-specific peaks are orange, ChIP-seq specific peaks are blue. (**G**) Scatterplot of peak intensity for common, MaTaDa-specific and ChIP-seq-specific peaks at $q<10^{-25}$. (**H**) Scatterplot of peak intensity for peaks common to MaTaDa and ChIP-seq at $q<10^{-25}$. (**I,J**) Distribution of peak intensity in PRDM14 MaTaDa (**I**) and ChIP-seq (**J**) for common, MaTaDa-specific and ChIP-seq-specific peaks at $q<10^{-25}$.

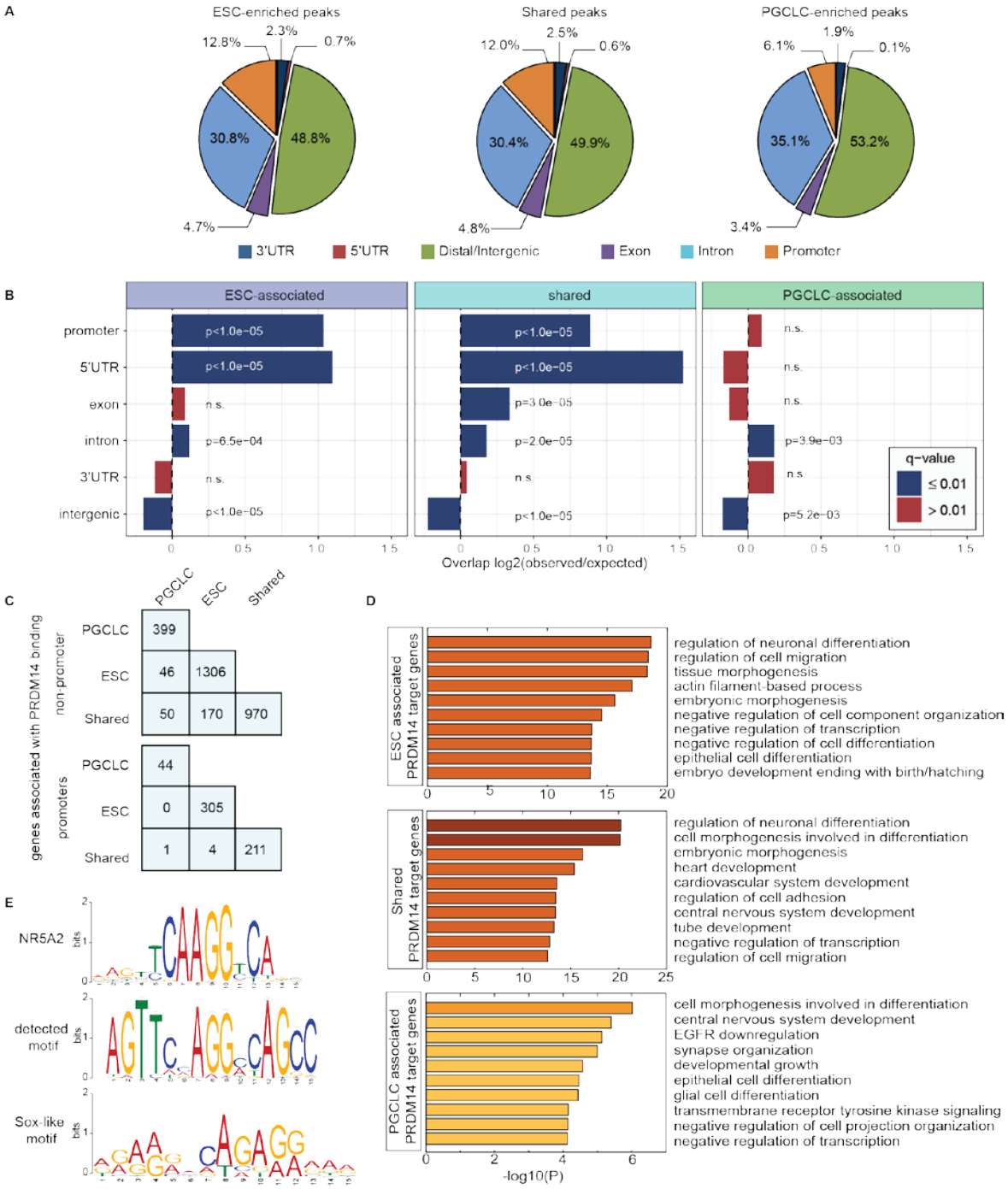


Supplementary Figure 5. Overlap of OCT4 MaTaDa and ChIP peaks in ESC. (**A**) Scatterplot of peak intensity for common, MaTaDa-specific and ChIP-seq-specific peaks at $q<10^{-25}$. ChIP-seq peaks are from (Buecker et al., 2014). Dashed outline for common peaks only is presented in **Fig 2H**. (**B,C**) Distribution of peak intensity in OCT4 MaTaDa (**B**) and ChIP-seq (**C**) for common, MaTaDa-specific and ChIP-seq-specific peaks at $q<10^{-25}$. **(D,E)** Distribution of peak density of the nearest common and MaTaDa-specific (**D**) or ChIP-specific (**E**) peaks relative to OCT4 ESC ChIP (**D**) or MaTaDa (**E**) peak summits at variable q-values. Insets show distribution of the individual peaks respectively at $q<10^{-200}$ and $q<10^{0}$. (**F**) Transcription factor motif enrichment analysis of all available OCT4-motifs in the i-cisTarget database for OCT4 ESC peaks common to MaTaDa and ChIP, or specific to either technique at three different q-values ($q<10^{-100}$, top; q<10-25, middle; $q<10^{0}$, bottom). Normalized enrichment score (NES) and percentage of peaks containing the motif are

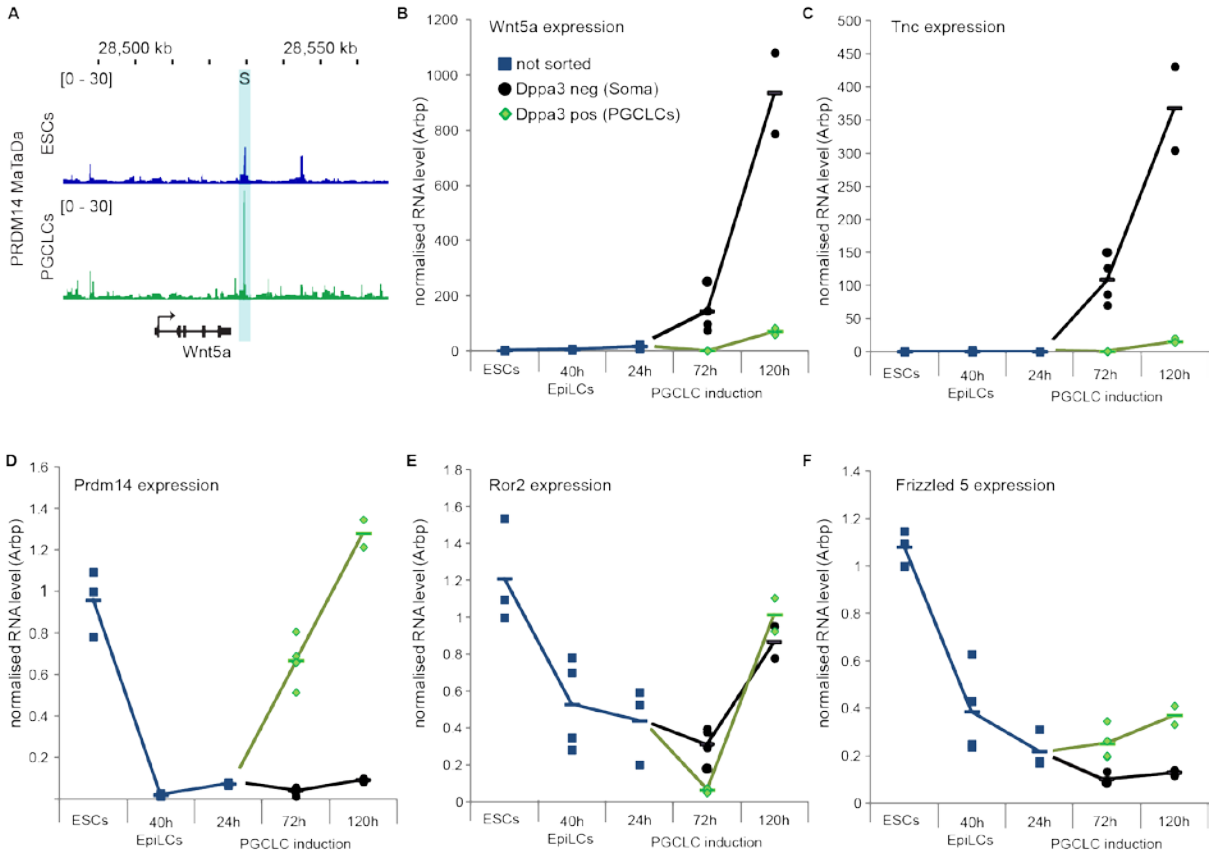
indicated. ND, not determined. (**G,H**) Genomic feature enrichment analysis of OCT4 MaTaDa-specific (**G**) or ChIP-specific (**H**) peaks ($q<10^{-25}$). Normalized enrichment score (NES) and percentage of peaks containing the feature are indicated.



Supplementary Figure 6. PGCLCs specify normally in the presence of PRDM14-Dam (**A**) Experimental outline; 2i/LIF cultured, Dppa3/GFP, Esg1/tdTomato transgenic ESCs and harbouring 1.) no additional transgene (parental), 2.) Dam-PRDM14 or, 3.) Dam inducible transgenes were differentiated into EpiLCs for 40h. During PGCLC specification cell were treated with either ethanol (EtOH) or tamoxifen (TAM) for 72h. (**B**) GFP and tdTomato fluorescence as well as brightfield microscopy images of transgenic PGCLCs cultured in the presence of ethanol or tamoxifen (scale bar 100µm) (**C**) FACS plots of the different transgenic PGCLCs treated with ethanol or tamoxifen during 72h of PGCLC specification. Sorted PGCLCs (Dppa3 positive) and control cells (negative) are indicated in the panels. (**D**) qPCR analysis of PGC (Prdm14, Blimp1), ESC (Klf4) and EpiLC (Dmnt3b) marker genes on parental, Dam-PRDM14 or Dam transgenic cell FACS purified PGCLCs (Dppa3) and control cells (neg), treated with tamoxifen or ethanol.

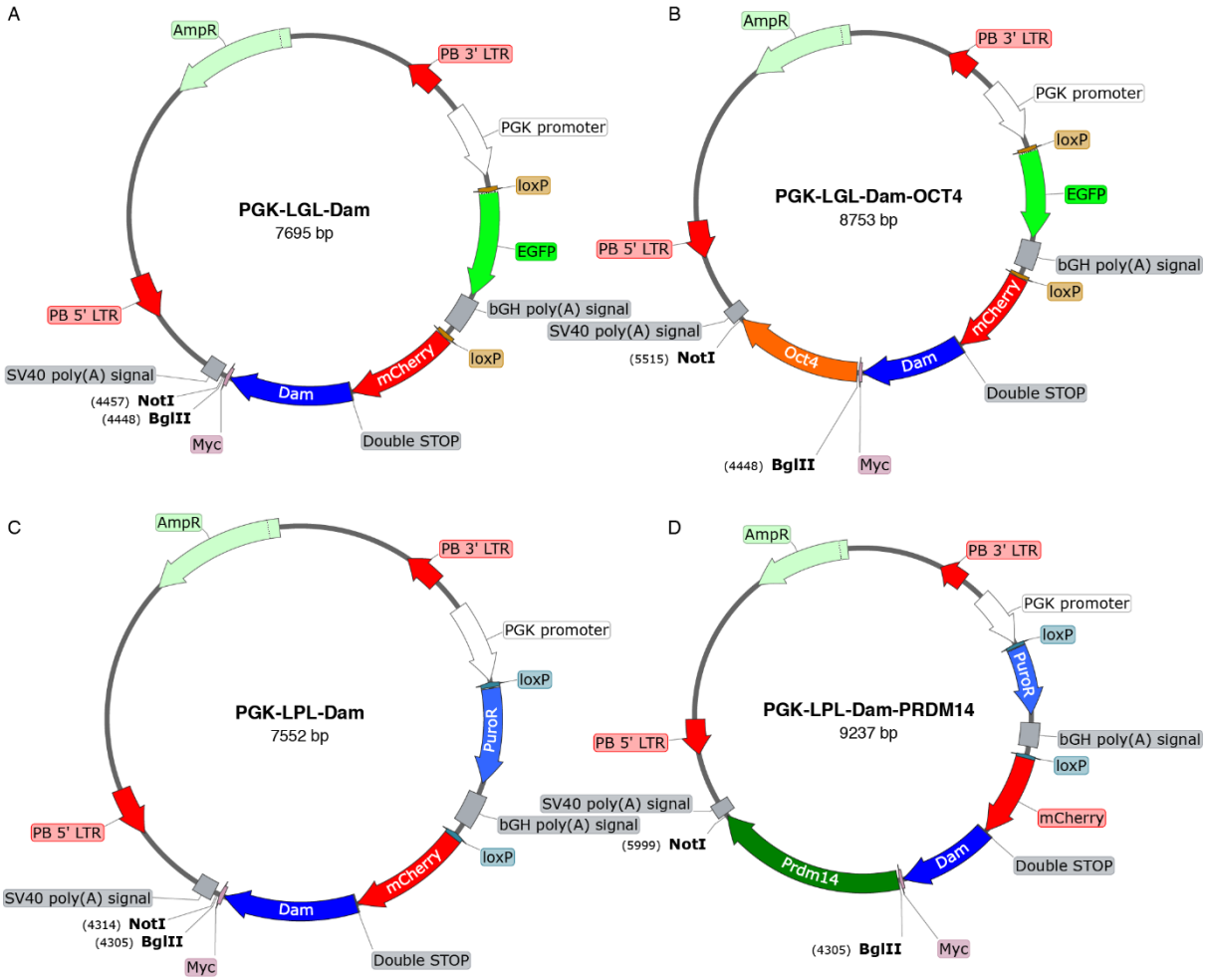


Supplementary Figure 7. Genome wide analysis of PRDM14 binding
(**A**) Genomic distribution of PRDM14 peaks. (**B**) Enrichment of PRDM14 relative to genomic features (Genomic association test). (**C**) PGCLC- and ESC-enriched peaks are associated with distinct sets of genes. (**D**) Gene ontology terms associated with ESC-enriched, shared and PGCLC-enriched PRDM14 peaks. (**E**) Two motifs one of which resembles the NR5A2 motif and the other which is Sox-like were detected de novo by MEME (Bailey et al., 2015).



Supplementary Figure 8: Gene expression kinetics

(**A**) Genome browser view of PRDM14 occupancy in ESCs and PGCLCs at the *Wnt5a* locus (both average of three replicates). Shared (S) binding site between ESC and PGCLC is shown. **(B-F)** Relative gene expression of indicated genes detected by qRT-PCR in 2i/LIF ESCs, EpiLCs (40h after induction) and at 3 time points of PGCLC development (24h, 72h and 120h after induction). Gene expression was normalized to *Arbp* expression. 72h and 120h PGCLCs were FACS purified using the Dppa3-GFP reporter gene. Dppa3-positive PGCLCs are depicted in green and Dppa3-negative somatic cells are depicted in black. Biological replicates (data points) and average gene expression (bar) are shown. (number of biological replicates: ESCs 3; EpilCs 4; 24h PGCLCs 3; 27h PGCLCs 4; 120h PGCLCs 2)



Supplementary Figure 9: MaTaDa Constructs

(A) PGK-LGL-Dam is a conditional vector for establishing stable cell lines that can inducibly express low levels of Dam methylase. The vector contains PB 3'LTRs that allow integration into mammalian genomes with the piggyBac transposase. PGK drives the expression of a floxed GFP cassette. Upon Cre treatment this cassette will be excised allowing transcription of mCherry (ORF1, LT3) and Dam. ORF1 will be highly translated but the presence of two stop codons and a single nucleotide frameshift before Dam results in extremely low levels of Dam expression. (**B**) PGK-LGL-Dam with OCT4 CDS inserted in the multiple cloning site. (**C**) PGK-LPL-Dam, a modification of PGK-LGL-Dam with the GFP cassette replaced with a puromycin resistance cassette for antibiotic selection. (**D**) PGK-LPL-Dam with the PRDM14 CDS inserted into the multiple cloning site.