TAF4b transcription networks regulating early oocyte differentiation

Megan A. Gura[1], Soňa Relovská[2], Kimberly M. Abt[1], Kimberly A. Seymour[2], Tong Wu[3], Haskan Kaya[4], James M. A. Turner[4], Thomas G. Fazzio[3] and Richard N. Freiman[1,2,5]

[1]MCB Graduate Program and [2]Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, 70 Ship St., Box G-E4, Providence, RI 02903, USA

[3]Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

[4]The Francis Crick Institute, London, UK

[5]Corresponding author

EMAIL: Richard_Freiman@Brown.edu

Phone: (401)-863-9633

Keywords: Transcription, oogenesis, oocytes, meiosis I, TFIID, TAF4b, POI, fetal oocyte attrition

ABSTRACT

Establishment of a healthy ovarian reserve is contingent upon numerous regulatory pathways during embryogenesis. Previously, mice lacking TBP-associated factor 4b (*Taf4b*) were shown to exhibit a diminished ovarian reserve. However, potential oocyte-intrinsic functions of TAF4b have not been examined. Here we use a combination of gene expression profiling and chromatin mapping to characterize TAF4b-dependent gene regulatory networks in mouse oocytes. We find that *Taf4b*-deficient oocytes display inappropriate expression of meiotic, chromatin, and X-linked genes. Furthermore, dysregulated genes in *Taf4b*-deficient oocytes exhibit an unexpected amount of overlap with dysregulated genes in Turner Syndrome oocytes.

Using Cleavage Under Targets and Release Using Nuclease (CUT&RUN), we observed TAF4b enrichment at genes involved in chromatin remodeling and DNA repair, some of which are differentially expressed in *Taf4b*-deficient oocytes. Interestingly, TAF4b target genes were enriched for Sp/Klf family and NFY target motifs rather than TATA-box motifs, suggesting an alternate mode of promoter interaction. Together, our data connects several gene regulatory nodes that contribute to the precise development of the mammalian ovarian reserve.

INTRODUCTION

The ability to produce healthy gametes is critical for the continuation of all sexually reproducing organisms, including humans. The process of mammalian gametogenesis begins during early fetal life with the specification and migration of primordial germ cells (PGCs) to the genital ridge. At the genital ridge, PGCs begin the process of differentiation into eggs and sperm in close concert with sex-specific somatic support cells. Thus, to understand the healthy functioning of adult gametes, we must examine multiple stages of development including those that arise in early fetal life. An added layer of complexity is that female XX and male XY germ cells traverse this differentiation process in a highly sex-specific manner (Feng et al., 2014). While some adult male germ cells become self-renewing spermatogonial stem cells (SSCs) during their development, the adult female mammalian germline is a non-renewable and finite resource termed the ovarian reserve that is steadily depleted after birth. The postnatal ovarian reserve is composed of a stockpile of primordial follicles (PFs) that contain individual primary oocytes arrested in prophase I of meiosis I, surrounded by a single layer of flattened somatic granulosa cells (Gura and Freiman, 2018). Menopause results from the timely depletion of the ovarian reserve and the mean age for menopause is 50±4 years. At least 1% of the female

population worldwide experiences a fertility deficit termed primary ovarian insufficiency (POI) where menopause-like symptoms occur prematurely by 40 years of age (Chandra et al., 2013). Thus, genetic and environmental factors that perturb the establishment of the ovarian reserve *in utero* will have negative consequences on adult reproductive and general health outcomes and need to be understood in greater detail.

We previously identified an essential function of TBP-Associated Factor 4b (TAF4b) in the establishment of the ovarian reserve in the embryonic mouse ovary (Grive et al., 2014; Grive et al., 2016). TAF4b is a germ cell-enriched subunit of the transcription factor TFIID complex, which is required for RNA Polymerase II recruitment to promoters in gonadal tissues (Gura et al., 2020). TFIID is a multi-protein complex that contains TATA-box binding protein (TBP) and 13-14 TBP-associated factors (TAFs) and is traditionally considered part of the cell's basal transcription machinery (Antonova et al., 2019). Female mice that have a targeted mutation, which disrupts the endogenous *Taf4b* gene and prevents TAF4b protein from integrating into the larger TFIID complex (called *Taf4b*-deficiency), are infertile and also exhibit hallmarks of POI including elevated follicle stimulating hormone (FSH) levels and a diminished ovarian reserve (DOR) (Falender et al., 2005; Gura et al., 2020; Lovasco et al., 2010a; Lovasco et al., 2015). We recently demonstrated that *Taf4b* mRNA and protein expression are nearly exclusive to the germ cells of the mouse embryonic ovary from embryonic day 9.5 (E9.5) to E18.5 and that *Taf4b*-deficient ovaries display delayed germ cell cyst breakdown, increased meiotic asynapsis, and excessive perinatal germ cell attrition (Grive et al., 2014; Grive et al., 2016; Gura et al., 2020). Therefore, we hypothesize that TAF4b, as part of TFIID, regulates oogenesis and meiotic gene programs. To what degree the transcriptomic pathways in *Taf4b*-deficiency and POI overlap and contribute to their similarities has yet to be explored.

Both human and mouse genetic studies have begun to reveal the molecular mechanisms underlying POI and its related pathologies. The most striking example is Turner Syndrome (TS) where karyotypically single X-chromosome female individuals undergo early and severe DOR and experience short stature, primary amenorrhea, estrogen insufficiency, and cardiovascular malformations (Gravholt et al., 2019). Recent work in mouse models of TS indicates that loss of correct dosage of the single X chromosomes in XO versus XX oocytes leads to pronounced meiotic progression defects and excessive oocyte attrition during ovarian reserve establishment (Sangrithi et al., 2017). In contrast to the high penetrance of TS, 20% of women with a premutation CGG repeat allele in the *FMR1* gene, also located on the X chromosome, experience a related fragile X-associated POI (FXPOI) (Fink et al., 2018). Similar to *Taf4b*, other targeted mouse mutations have resulted in POI-related phenotypes including those in *Nobox* and *Figla*, two transcription factors that regulate oocyte development, however the relevance of specific mutations in their human orthologs and POI in women remains to be explored (Rossetti et al., 2017). More importantly, a better understanding of how these genes promote healthy establishment of the ovarian reserve and the deregulated molecular events that lead to its premature demise is needed.

To better understand the normal function of TAF4b during establishment of the ovarian reserve, we integrated published bioinformatic data with experimental *Taf4b* genomic assays to uncover unexpected links of *Taf4b* with TS and *Fmr1*. We show that in homozygous mutant *Taf4b* E16.5 oocytes, almost 1000 genes are deregulated as measured by RNA-sequencing (RNA-seq). Surprisingly, the X chromosome was enriched for these deregulated genes and *Taf4b*-deficient oocytes display reduced X:autosome (X:A) gene expression ratios. There is a striking overlap of genes deregulated in *Taf4b*-deficient oocytes and XO mouse oocytes, and XO

oocytes express significantly reduced levels of *Taf4b* at E15.5 and E18.5, further illuminating a potential molecular link between these disparate genetic contributors to POI. Further, we show that *Taf4b*-deficiency and TS both result in deregulation of genes involved in chromatin organization, modification, and DNA repair. Finally, CUT&RUN of TAF4b E16.5 XX germ cells identifies direct TAF4b targets, enriched for Sp/Klf zinc finger family and NFY binding sites, that for the first time confirm its promoter-proximal recognition properties, linking TAF4b binding to the critical transcriptional regulation required for the proper establishment of the ovarian reserve.

RESULTS

Taf4b expression peaks at E16.5 in female embryonic germ cells

To observe the dynamics of *Taf4b* mRNA expression at a single cell resolution, we analyzed a single-cell RNA-seq (scRNA-seq) dataset of Oct4-EGFP^{+} oocytes from E12.5, E14.5, and E16.5 mouse ovaries (Zhao et al., 2020). We selected for *Dazl*-positive, high quality (nFeature_RNA > 1000, nFeature_RNA < 5000, nCount_RNA > 2500, nCount < 30000, percent mitochondrial genes < 5%) oocytes and performed pseudotime analysis using Monocle3 (**Fig 1A- B**). We found that *Figla* expression generally increased from E12.5 to E16.5 and over pseudotime, with some of the highest *Figla*-expressing cells appearing in E16.5 cells at the end of the pseudotime profile. We also found that *Stra8* expression, which is a master regulator of meiotic initiation, declined over time and pseudotime as expected. We then compared the expression profiles of *Taf4a* and *Taf4b*. Most cells across the time course had low *Taf4a* expression throughout. *Taf4b* mRNA expression began to rise at E14.5 and appeared highest in

the E16.5 oocytes that were earliest in pseudotime (**Fig 1C**), consistent with previous observations (Gura et al., 2020).

To identify which other genes were highly expressed in *Taf4b*-expressing oocytes, we performed differential gene expression analysis using cells separated into *Taf4b*-expressing (*Taf4b* $\log_2$ expression > 0) and *Taf4b*-off (*Taf4b* $\log_2$ expression = 0) populations and performed differential gene expression analysis (**Table S1**). We found 155 genes significantly (p-value < 0.05) higher in *Taf4b*-expressing cells. We performed gene ontology (GO) analysis of these genes and found that the top categories included "meiotic cell cycle" and "synaptonemal complex organization" (**Fig 1D**, **Table S1**). Taken together, these data suggest that *Taf4b* expression is highest in the E16.5 mouse oocyte and that *Taf4b* is co-expressed with important meiotic genes. A similar analysis from a second scRNA-seq dataset of whole ovaries from earlier (E11.5 to E14.5) time points supported these findings (Ge et al., 2021) (**Fig S1**).

RNA-seq identifies TAF4b-affected genes in E16.5 XX germ cells

To understand the transcriptome-level changes in *Taf4b*-deficient embryonic oocytes, we performed RNA-seq at E16.5. We sorted Oct4-EGFP$^+$ oocytes from five *Taf4b*-heterozygous (*Taf4b* +/-) and five *Taf4b*-deficient (*Taf4b* -/-) pairs of ovaries and subjected them to ultra-low input RNA-seq (germ cell numbers for each RNA-seq sample can be found in **Table S2**). The resulting principal component analysis (PCA) plot shows each of the *Taf4b*-deficient samples mostly grouping together, with the *Taf4b*-heterozygous samples dispersed throughout (**Fig 2A**, **Table S3**). This patterning of the data is largely due to the litter from which each sample originates, as we were unable to obtain sufficient numbers of our desired genotypes from a single mouse litter, but importantly the different genotypes separate when plotting litter dates individually (**Fig S2A**). We identified 964 differentially expressed genes (DEGs) between *Taf4b*-

heterozygous and *Taf4b*-deficient oocytes, which were defined as protein-coding, average transcripts per million (TPM) expression > 1, and adjusted p-value < 0.05 (**Fig 2B**, **Table S3**). From this list of DEGs, 463 were increased in *Taf4b*-deficient oocytes and will be referred to as "Upregulated DEGs". Some interesting DEGs in this gene set were *Fmr1* (the most common genetic cause of POI), *JunD* (a component of the AP-1 transcription factor complex), and *Sp1* (a DNA-binding transcription factor) (Fink et al., 2018; Mechta-Grigoriou et al., 2001; Vizcaíno et al., 2015) (**Fig S2B**). The remaining 501 DEGs were decreased in *Taf4b*-deficient oocytes and will be referred to as "Downregulated DEGs". As expected, *Taf4b* was a Downregulated DEG, as was another well-known oogenesis gene *Nobox* (**Fig S2C**). Finding *Nobox* as a DEG corroborates previous research which showed that TAF4b directly binds to the promoter region of *Nobox* and promotes its protein expression in E18.5 oocytes (Grive et al., 2016). We also identified *Fam83d* as a Downregulated DEG, which has been implicated to play a role in ovarian cancer (Zhang et al., 2019). For validation of these genes, we performed quantitative real-time PCR (qRT-PCR) on E17.5 Oct4-EGFP^{+} oocytes which corroborated our RNA-seq results (**Fig S2D**). Analysis of known protein-protein interactions (PPIs) using STRING revealed a significant enrichment of PPIs, with major nodes including *Ep300* (a histone acetyltransferase involved in chromatin remodeling) and *Plk1* (a serine/threonine-protein kinase involved in cell cycle regulation) (**Fig S3**). We also performed a similar RNA-seq experiment at E14.5, but found fewer DEGs suggesting that more substantial transcriptomic effects of *Taf4b*-deficiency take place around E16.5 (**Fig S4A-B**, **Table S4**).

We performed GO analysis of all the E16.5 DEGs, as well as separating the Upregulated and Downregulated DEGs (**Fig 2C**; **Fig S2E-F**). We found multiple chromatin organization and modification GO categories associated with Upregulated DEGs and reproduction- and

microtubule-related categories associated with Downregulated DEGs (**Table S3**). Overall, these data suggest that TAF4b impacts the expression of many genes in the developing oocyte transcriptome, particularly those associated with chromatin structure and modification and reproduction. Moreover, the effects of TAF4b on the transcriptome take place after E15.5, correlated with the peak in *Taf4b* expression at E16.5 shown by scRNA-seq and bulk RNA-seq (**Fig 1**) (Gura et al., 2020).

X chromosome gene expression is significantly reduced in *Taf4b*-deficient oocytes

Our E16.5 RNA-seq analysis led us to examine how *Taf4b*-deficiency affects expression of each mouse chromosome. Surprisingly, we observed that there were significantly more Downregulated DEGs on the X chromosome than expected and significantly fewer Upregulated DEGs on the X chromosome (**Fig 3A-B**, **Tables S5-6**). Furthermore, the X chromosome was the only chromosome to exhibit such a phenomenon for both sets of DEGs. We then determined if this skew in DEGs translated into overall reduced X chromosome expression compared to autosomes. When comparing the $\log_2$ fold change between *Taf4b*-heterozygous and *Taf4b*-deficient oocytes, we found that there was significantly lower expression of X chromosome genes versus autosomal genes (**Fig 3C**). Two similar but slightly different dosage compensation calculation methods, the X:A ratio and relative X expression (RXE), further support that the expression of X chromosome genes is reduced in E16.5 *Taf4b*-deficient oocytes (outliers not plotted) (**Fig 3D-E**). However, we did not see a significant difference in X chromosome expression when looking at E14.5 oocytes (**Fig S2C-E**).

Ohno's hypothesis postulates that the expression of the X chromosome is uniquely regulated so that "housekeeping genes" on the X largely remain on par with autosomal housekeeping gene expression (Ohno, 1966). In Sangrithi et al., 2017, the authors annotated the mouse genome for genes expressed (FPKM ≥1) in all tissues they sampled (Sangrithi et al., 2017). We used this set of ubiquitously expressed genes to see if the effects of *Taf4b*-deficiency on X chromosome expression were specific to ubiquitously expressed genes. We found that 39% of our DEGs were members of the ubiquitous genes list (**Fig S5A**), which is higher than the 25% of all genes being ubiquitously expressed. However, when we plotted the $\log_2$ fold change of ubiquitous genes on the X chromosome and autosomes, there was no significant difference between these populations (**Fig S5B**, **Table S7**). Taken together, these data indicate that *Taf4b*-deficiency affects the expression of the X chromosome but it is unclear if *Taf4b* plays a direct role in dosage compensation.

Overlap of mouse genes deregulated by *Taf4b*-deficiency and Turner Syndrome

As this is the first link of *Taf4b* to the regulation of X-linked gene expression, we decided to compare it to a mouse model of Turner Syndrome (TS). TS is a chromosomal disorder where a female individual has one intact X chromosome and a second X chromosome either missing or severely compromised (Gravholt et al., 2019). We re-processed the raw data from Sangrithi et al., 2017, where the researchers used Oct4-EGFP mice covering 4 developmental time points in female XX and XO germ cells, ranging from E9.5 to E18.5 (**Fig 4A**) (Sangrithi et al. 2017). *Taf4b* expression was not significantly different between the karyotypes at E9.5 and E14.5, but it was significantly reduced in E15.5 and E18.5 XO oocytes, whereas *Taf4a* expression was not significantly different at any time point (**Fig 4B-C**, **Table S8**). To examine the potential overlap

of transcriptomic effects between TS and *Taf4b*-deficiency, we compared their DEGs. We first compared E15.5 TS DEGs to our E16.5 *Taf4b* DEGs and found 243 genes shared between the two gene sets, there was a significant overlap ($p < 0.05$, hypergeometric test) (**Fig 4D**). When we used these 243 genes as input for GO analysis, we found DNA-related categories enriched such as "DNA repair" and "covalent chromatin modification" (**Fig 4E**). We then compared E18.5 TS DEGs to our E16.5 *Taf4b* DEGs and found 439 genes shared between the two contexts, which was also a significant overlap ($p < 0.05$, hypergeometric test) (**Fig 4F**). When we used these 439 genes as input for GO analysis, we again found DNA-related categories enriched such as "DNA repair" (**Fig 4G**). When we compared the E15.5 and E18.5 TS X chromosome DEGs to our E16.5 *Taf4b* X chromosome DEGS, we found 14 and 31 shared DEGs, respectively (**Fig S6**). These data indicate that there are shared transcriptomic effects of both TS and *Taf4b*-deficiency in mouse embryonic oocytes, and that these shared effects are related to functions concerning DNA repair and chromatin modification.

We observed similar results in an independent TS dataset (Hamada et al., 2020). By comparing *Taf4b* expression in XX and XO cells that had been differentiated from mouse embryonic stem cells *in vitro*, we found that *Taf4b* expression was lower in the XO cells that best resembled late embryonic oocytes (**Fig S7A-B**, **Table S9**). Interestingly, when the cells had been further differentiated to be similar to early postnatal oocytes, the trend reversed with *Taf4b* expression being significantly higher in mature oocyte-like cells derived from XO cells. This corroborates the reduction in *Taf4b* expression in mature mouse oocytes and suggests that oocyte expression of *Taf4b* normally decreases postnatally. In contrast, significant differences in *Taf4a* expression occurred in d6PGCLCs and in the latter stages of differentiation that best resembled postnatal oocytes (**Fig S7C**).

CUT&RUN identifies putative direct targets of TAF4b in E16.5 germ cells

To understand which DEGs identified in our E16.5 RNA-seq experiment were likely to be direct targets of TAF4b, we performed Cleavage Under Targets and Release Using Nuclease (CUT&RUN), a technique to map binding sites of specific proteins or histone modifications in the genome. We isolated E16.5 female germ cells using FACS and examined the genomic localization of TAF4b, H3K4me3 (positive control and marker of promoter regions), and IgG (negative control). We performed two replicates of this experiment, with the germ cells in Replicate 1 consisting of 42,416 cells per tube (obtained from 12 embryos) and the Replicate 2 germ cell numbers were 63,079 cells per tube (obtained from 33 embryos). CUT&RUN data analysis using Homer identified 8,129 H3K4me3 peaks and 983 TAF4b peaks in Replicate 1 and 320 H3K4me3 peaks and 1,111 TAF4b peaks in Replicate 2 (**Table S10**). We also found that 90% and 95% of TAF4b peaks were classified as localizing to promoters/transcription start sites ("Promoter-TSS") for Replicates 1 and 2, respectively (**Fig 5A**). Of all the genes that contained TAF4b promoter/TSS peaks, 449 overlapped between the replicates (**Fig 5B**). However, it is clear when looking at some gene tracks (*Polr2a*, for example) that even when a TAF4b peak is identified in only one of the replicates, there is enrichment of TAF4b in the same location in the other replicate, suggesting that some TAF4b binding sites are below the limit of detection by our peak calling criteria (**Fig S8**). When plotting the enrichment profile of TAF4b and H3K4me3 relative to TSSs, we found the highest TAF4b enrichment upstream of the TSS in both replicates (**Fig 5C**). To more closely examine the localization of TAF4b signal near TSSs, we plotted the distance of TAF4b "promoter-TSS" peaks from both replicates to the TSS (**Fig 5D**). There is strong enrichment of TAF4b peaks between -200 bp to +50 bp from the TSS, with the highest number of TAF4b peaks located at -60 to -40 bp away from the TSS.

We performed GO analysis of the shared TAF4b-bound gene promoter-TSSs between the two replicates and found categories related to mRNA processing, DNA repair, and chromatin remodeling (**Fig 5E**). To see if the transcriptomic effects of *Taf4b*-deficiency on the X chromosome arise from greater X chromosome localization, we plotted the expected versus observed number of peaks using the TAF4b peaks that were categorized as "Promoter-TSS" and surprisingly found that there were fewer X chromosome peaks than expected (**Fig 5F**). Given that we found that there are more DEGs between *Taf4b*-heterozyous and -deficient oocytes on the X chromosome than expected, this suggests that there might be an indirect but disproportionately high effect of TAF4b on the X chromosome in E16.5 oocytes. However, we cannot claim this approach thoroughly annotates all TAF4b-bound sites in the developing female germ cell genome.

This CUT&RUN experiment allowed us to begin inquiring as to which DEGs identified in our RNA-seq experiment were putative direct targets of TAF4b. When comparing our DEGs to the "Promoter-TSS" peaks of TAF4b, we found 129 DEGs that had at least one peak near their TSS (**Fig 6A**). GO analysis of these peaks found that the categories enriched in these data pertained to chromatin modifications and organization (**Fig 6B**). A volcano plot of these TAF4b-bound DEGs reveals a skew in the number of Upregulated versus Downregulated DEGs, with 34 TAF4b-bound DEGs being Downregulated and 95 TAF4b-bound DEGs being Upregulated (**Fig 6C**). This means there were three times the number of TAF4b-bound Upregulated DEGs compared to Downregulated DEGs, suggesting that TAF4b may primarily antagonize mRNA levels in developing oocytes. As examples of TAF4b-bound DEGs, we present gene tracks for *JunD*, *Sp1*, *Fmr1*, and *Taf4b* (**Fig 6C**). *JunD*, *Sp1*, and *Fmr1* were all upregulated in *Taf4b*-deficient oocytes. These data suggest that TAF4b negatively regulates the expression levels of

these transcription factors and *Fmr1*. Therefore, we next determined whether the level of Fragile X Mental Retardation Protein (FMRP) encoded by *Fmr1* was also perturbed in *Taf4b*-deficient embryonic oocytes. To do this, we performed immunofluorescent staining of FMRP and the nuclear germ cell marker Tra98 in E16.5 *Taf4b*-wildtype and -deficient ovary tissue sections (**Fig 6E**). First, we focused on germ cell clusters within each tissue section and we then quantified the levels of fluorescence for both the FMRP and Tra98 channels within each individual cluster. We found that there was a modest but statically significant ($p < 0.01$) increase in FMRP signal intensity in *Taf4b*-deficient germ cell clusters compared to wildtype. However, there was no significant difference in Tra98 signal intensity between wildtype and *Taf4b*-deficient germ cell clusters, indicating that the increase in FMRP signal was not due to an increase in oocyte numbers (**Fig 6F**). The increased FMRP levels are consistent with upregulated *Fmr1* mRNA levels in *Taf4b*-deficient E16.5 and E17.5 oocytes as shown by RNA-seq and qRT-PCR, respectively (**Fig S2B&D**).

We then identified conserved motifs in TAF4b "promoter-TSS" peaks and were surprised to find that TATA-box was not among the top 5 motifs from the 129 DEGs that had at least one TAF4b "promoter-TSS" peak (**Fig 7A**). Instead, GC-box motifs which are bound by the Sp/KLF family of transcription factors and the CCAAT-box, bound by NFY, dominated the list. TAF4b peak motif analysis for each female replicate, as well as all TAF4b motifs combined, yielded the same 5 motifs (**Fig S9A-B&F**). Examining the peaks associated with Upregulated and Downregulated DEGs did not reveal any conclusive differences in TAF4b-bound motifs (**Fig S9C-D**). As we had previously noted that the highest number of TAF4b "promoter-TSS" peaks were located -60 to -40 bp away from the TSS (**Fig 5D**), suggesting that TAF4b may be binding a few nucleotides upstream of the canonical -25 to -30 bp TATA-box location in mouse

embryonic germ cells. To further explore this, we created boxplots of the distance to the TSS (no outliers included) of the following: all TAF4b "promoter-TSS" peaks, all TAF4b "promoter-TSS" peaks for genes that were also DEGs, TAF4b "promoter-TSS" peaks for genes that were only Downregulated DEGs, and TAF4b "promoter-TSS" peaks for genes that were only Upregulated DEGs (**Fig 7B**). Their median locations from the TSS were -65 bp, -87.5 bp, -72.5 bp, and -104 bp, respectively, with there being a significant difference in the location between all TAF4b "promoter-TSS" peaks and all TAF4b "promoter-TSS" peaks for genes that were also DEGs. When performing motif enrichment on "promoter-TSS" peaks based on distance to TSS, we found that the same motifs (NFY and Sp1) were in the top three frequently rather than strongly varying based on location (**Fig S9E**). This integration of RNA-seq and CUT&RUN data suggests that TAF4b directly regulates chromatin remodeling and modification genes in oocytes, perhaps through an unconventional protein-protein interaction that prioritizes other motifs just upstream of the TATA-box (discussed below). However, more canonical functions of TAF4b cannot be ruled out, as "TATA-Box (TBP)/Promoter" did appear as a significantly enriched motif in both replicates, it was ranked 140 in Replicate 1 and 137 in Replicate 2.

We then evaluated if the genes that were commonly associated with TAF4b-bound motifs in E16.5 oocytes were dynamic in their expression over germ cell development. We re-examined our re-processed scRNA-seq dataset from E12.5 to E16.5 mouse oocytes for the gene expression profiles of *Nfya*, *Nfyb*, *Nfyc*, *Sp1*, *Sp2*, *Klf3*, and *Sp5* (**Fig 7C**). All genes with the exception of *Nfyc* were relatively unchanging over the time and pseudotime courses in mouse oocytes. *Nfyc* showed its highest expression in the E14.5 cells that were closest to E16.5 in pseudotime. These data indicate that if TAF4b is directly interacting with one or more of these proteins, it might be

TAF4b that provides the dynamic expression in germ cells rather than its potential binding partner.

DISCUSSION

Proper establishment of the ovarian reserve is essential for the reproductive capacity of female mammals, including both humans and mice. This healthy establishment of female gametes is orchestrated through complex oocyte transcription networks that must also properly distinguish germ cell and somatic cell lineages. In addition to the more well-known enhancer-bound transcription activators and repressors, tissue-selective components of the basal transcription machinery can help impart such exquisite regulatory control (Freiman, 2009; Goodrich and Tjian, 2010). We have previously shown that the TAF4b subunit of the TFIID complex is required for proper establishment of the ovarian reserve in the mouse (Grive et al., 2014; Grive et al., 2016; Lovasco et al., 2010b). However, the network of genes regulated by TAF4b to accomplish this critical task has been elusive, until now. Here we show that TAF4b directly and indirectly regulates genes essential for proper meiotic progression during early oocyte differentiation. Integration of RNA-seq and CUT&RUN data in E16.5 mouse oocytes reveals germ cell-intrinsic regulation by TAF4b in the promoter-proximal regions of chromatin modification and organization genes. Furthermore, we discovered an unexpected link between *Taf4b*-deficiency in the mouse to the proper expression of the mouse X chromosome and similarities to the transcriptome of TS, a well- known cause of POI in women (Sangrithi et al., 2017). Surprisingly, TATA-box motifs were not among the top binding motifs in either female oocyte replicate shown by CUT&RUN nor was the peak enrichment of TAF4b at the expected location (**Fig 8A**). Together, these molecular insights suggest TAF4b directly regulates genes

instrumental in establishing the finite ovarian reserve and that TAF4b may have a non-canonical function in mouse oocytes outside of TFIID or in an unconventional version of TFIID (**Fig 8B**).

TFIID was first discovered as a large multi-protein complex required for activator-dependent RNA polymerase II transcription (Dynlacht et al., 1991; Reinberg et al., 1987). Characterization of the composition of TFIID revealed a key DNA binding subunit, TBP, that binds directly to the TATA-box found at the -25 nucleotide position in relation to the TSS of many genes (Hoey et al., 1990). Surprisingly, our embryonic germ cell CUT&RUN data for TAF4b centers its peak of binding to G-C and CCAAT- box sequences at -40 to -60 bp upstream (with TAF4b-bound DEGs containing peaks even further upstream), but still proximal to the TSS. These sequences are well-known binding sites for specificity protein 1 (Sp1) and nuclear factor y (NFY) transcription factors that are known to play extensive roles in promoter proximal transcription and are ubiquitously expressed. Although we do not yet know the significance of these binding sites and the occupancy of TAF4b, there are interesting clues in the published literature. Hibino et al., 2016 showed that there is a direct interaction between human Sp1 and TAF4a through their intrinsically disordered domains. TAF4b lacks most of the large intrinsically disordered regions that TAF4a contains, but TAF4b was not tested in that study. Therefore, we do not know if it might have some capacity to bind to Sp1. Since Sp1 is highly and ubiquitously expressed in cell types throughout the body, a tidy hypothesis would be that Sp1 provides the DNA-binding capacity and TAF4b provides the germ cell expression specificity (**Fig 8B**). We also know that NFY is a protein complex with three components: NF-YA, NF-YB, and NF-YC. Both NF-YB and NF-YC contain histone fold domains and heterodimerize, leading to the NFY protein complex acting in a sequence-specific, histone-like mode of DNA binding (Nardini et al., 2013). TAFs, including TAF4b, also contain histone fold

domains and perhaps this shared feature could enable their cooperation through di-/trimerization in oocytes. Interestingly, *Nfyb* was a Downregulated DEG in our E16.5 RNA-seq experiment and *Nfya* and *Nfyc* were non-significantly decreased (**Table S3**). Furthermore, Replicate 2 of our CUT&RUN experiment contained a TAF4b peak in the "promoter-TSS" region for *Nfya* (**Table S10**). Therefore, we have some limited evidence that the factors whose motifs comprise TAF4b-bound promoter regions a have a connection to TAF4b in our own data. Further molecular investigations are required to sort out the germ cell-specificity of this regulatory logic and the exact protein binding partners of TAF4b.

Similar diversification of selective TFIID subunits has occurred within germline development of highly distant organisms including in insects, vertebrates, and plants. In Drosophila, several testis-specific TAFs (tTAFs) play a critical role in regulating transcription and the timing of spermatogenic differentiation and a germ cell-expressed TBP paralog, TBP-related factor 2 (TRF2) is required for oogenesis (Gazdag et al., 2009; Hiller et al., 2004). The mouse ortholog of TRF2, called TBPL1, is required for spermiogenesis, as is TAF7l that is coordinately expressed with TAF4b in early meiotic oocytes (Zhou et al., 2013a). Interestingly, TBPL1, TAF7l, and TAF9b have also been shown to be critical for muscle and adipocyte differentiation, indicating that this diversification of TFIID subunits has evolved to regulate both somatic and germ cell differentiation, sometimes via the identical subunit (Herrera et al., 2014; Zhou et al., 2013b). The most striking parallel of the early meiotic transcription and chromatin functions of TAF4b shown here lie with a natural variant of TAF4b found in Arabidopsis (AtTAF4b) (Lawrence et al., 2019). This recent study has identified a similar timing of the meiocyte transcriptome regulated by AtTAF4b as we show here for mouse TAF4b. Although arising independently in the plant and animal kingdoms, there appears to be some common

transcription and/or chromatin state that is regulated by TAF4b to ensure the fidelity of meiotic recombination and early oogenesis. We have previously found other TFIID subunits such as *Taf7l* and *Taf9b* to be preferentially and dynamically expressed in embryonic mouse oocytes (Gura et al., 2020). One hypothesis to explain this data is that a germ cell-specific version of TFIID may exhibit different characteristics and targets than canonical TFIID (**Fig 8B**).

One major limitation of this study is the heterogeneity inherent to experiments on E16.5 oocytes, even ones that have been sorted for GFP fluorescence. Because oocytes progress through meiosis I asynchronously, some oocytes will have advanced as far as pachynema while others may still be in leptonema. Furthermore, we know meiotic progression is slowed in *Taf4b*-deficient oocytes (Grive et al., 2016). Therefore, it is not possible to conclude that TAF4b has a specific transcriptomic effect on these substages, such as pachynema, from this aggregated data. Potential experiments to circumvent this issue could be to perform single-cell RNA-sequencing analysis between *Taf4b*-wildtype and -deficient oocytes, which would help improve the resolution of the transcriptomic effects of *Taf4b*-deficiency by allowing us to look at developmental markers and pseudotime trajectories. We could also perform the "3S" method on male germ cells to isolate precise prophase I substages of germ cells, albeit in male mice rather than female (Romer et al., 2018). Despite the challenges of this oocyte heterogeneity, we revealed some consistent trends in our data. One avenue of future research is regarding chromatin modifications and organization because their categories were repeatedly found in our data. This study further corroborates earlier hints of chromatin's relevance to TAF4b where higher yH2AX, less recombination, and more asynapsis were found in E16.5 *Taf4b*-deficient oocytes (Grive et al., 2016). Further exploration of the chromatin state in *Taf4b*-deficient oocytes

will hopefully help bring the worlds of transcription and chromatin biology closer together in these newly born oocytes.

In addition to illuminating the molecular underpinnings of TAF4b function, we discovered unexpected overlaps between *Taf4b*-deficiency and other known causes of POI. Many TS individuals experience POI, which includes not reaching or delayed menarche and primary amenorrhea. Recent research has suggested that excessive prenatal oocyte loss may underlie the ovarian insufficiency in TS. Excessive oocyte attrition at the perinatal DNA damage checkpoint, where oocytes that have not resolved DNA damage or still contain asynapsed chromosomes are eliminated, resulting in a depleted ovarian reserve and its downstream sequelae. Observing deregulation of the X chromosome in *Taf4b*-deficient E16.5 oocytes was surprising and prompted us to compare *Taf4b*-deficiency and a mouse model of TS where females contain a single X chromosome (XO) (Grive et al., 2016). The extensive overlap of the deregulated gene expression in XO and *Taf4b*-deficient early meiotic oocytes was striking. Importantly, *Taf4b* expression itself was compromised in the XO oocytes, indicative of potential mutual regulation between these two genetic changes. A recent report uncovered a unique mechanism of XX dosage compensation in human primordial oocytes and it is possible that TAF4b plays an integral role in this sexually dimorphic mechanism of X-chromosome regulation (Chitiashvili et al., 2020). Another interesting parallel is the association of X chromosome-encoded *Fmr1* gene with TS and *Taf4b*-deficiency. *Taf4b* expression is significantly correlated with *Fmr1* in embryonic human ovaries, mutation of *Fmr1* is one the most common underlying genetic causes of POI, and TS individuals are missing one copy of *Fmr1*. Here we show that TAF4b directly associates with the proximal promoter region of *Fmr1* and the loss of TAF4b increases its mRNA abundance. Interestingly, the peak of TAF4b at *Fmr1* was not localized to

the CGG repeats of the gene, which are the underlying cause of *Fmr1's* contribution to POI incidence (Fortuño and Labarta, 2014). Given the clear links between *Taf4b*-deficiency, TS, FXPOI, and POI presented here, we suspect that a core of the genes identified in this study are required for the proper development of the ovarian reserve in humans and if that quorum is not reached a similar cascade of dysregulated gene expression occurs. Understanding these and other causes of POI will clarify the best ways to manage these related infertility syndromes and improve assisted reproduction therapies.

MATERIALS AND METHODS

Ethics statement

This study was approved by Brown University IACUC protocol #21-02-0005. The primary method of euthanasia is CO_2 inhalation and the secondary method used is cervical dislocation both as per American Veterinary Medical Association (AVMA) guidelines on euthanasia.

Mice

Mice that were homozygous for an Oct4-EGFP transgene (The Jackson Laboratory: B6;129S4-*Pou5f1*^{tm2Jae}/J) were backcrossed to the C57BL/6 line and mated for CUT&RUN collections. Mice that were homozygous for an Oct4-EGFP transgene (The Jackson Laboratory: B6;129S4-*Pou5f1*^{tm2Jae}/J) and C57BL/6 mice heterozygous for the *Taf4b*-deficiency mutation (in exon 12 of the 15 total exons of the *Taf4b* gene that disrupts the endogenous *Taf4b* gene) were mated for mRNA collections. Timed matings were estimated to begin at day 0.5 by evidence of a copulatory plug. The sex of the embryos was identified by confirming the presence or absence of

testicular cords. Genomic DNA from tails was isolated using Qiagen DNeasy Blood & Tissue Kits (Cat #: 69506) for PCR genotyping assays.

All animal protocols were reviewed and approved by Brown University Institutional Animal Care and Use Committee and were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals. Ovaries were dissected out of embryos into cold PBS.

Embryonic ovary dissociation and fluorescence-activated cell sorting

To dissociate ovary tissue into a single-cell suspension, embryonic ovaries were harvested and placed in 0.25% Trypsin/EDTA and incubated at 37°C for 15 and 25 minutes for E14.5 and E16.5 ovaries, respectively, as previously described (Gura et al., 2020). Eppendorf tubes were flicked to dissociate tissue halfway through and again at the end of the incubation. Trypsin was neutralized with FBS. Cells were pelleted at 1,500 RPM for 5 minutes, the supernatant was removed, and cells were resuspended in 100 µL PBS. The cell suspension was strained through a 35 µm mesh cap into a FACS tube (Gibco REF # 352235). Propidium iodide (1:500) was added to the cell suspension as a live/dead distinguishing stain. Fluorescence-activated cell sorting (FACS) was performed using a Becton Dickinson FACSAria III in the Flow Cytometry and Cell Sorting Core Facility at Brown University. A negative control non-transgenic mouse ovary was used for each experiment to establish an appropriate GFP signal baseline. Dead cells were discarded and the remaining cells were sorted into GFP^+ and GFP^- samples in PBS at 4°C for each embryo.

For RNA-seq analysis, GFP^+ cells from each individual embryo were kept in separate tubes and were then spun down at 1,500 RPM for 5 minutes, PBS was removed, and were then resuspended in Trizol (ThermoFisher # 1556026). If samples had roughly less than 50 µL of PBS in the tube, Trizol was added immediately. The number of cells for each sample sequenced can be found in **Table S2**. We used five embryos per genotype and littermates at E16.5 due to the less clear results we received at E14.5. Five embryos per genotype, taken from the same three litters was sufficient to overcome the embryo and oocyte heterogeneity inherent in this tissue type. Samples were stored at -80°C.

For CUT&RUN germ cells, all the collected ovaries were pooled prior to FACS. Sorted cells were then spun down at 1,500 RPM for 5 minutes and were resuspended in 300 µL of PBS, then split into three Eppendorf tubes. These three tubes of germ cells were then used for CUT&RUN. The number of cells for each sample were as follows: Replicate 1 female germ cell samples had 42,416 cells per tube (obtained from 12 embryos) and Replicate 2 female germ cell samples had 63,079 cells per tube (obtained from 33 embryos). We used many embryos per replicate because of the meiotic heterogeneity in oocytes and we used two replicates so that we could better hone in on the genes that were consistently bound by TAF4b.

Single cell RNA-seq data analysis

SRP193506 and SRP188873 were downloaded from NCBI SRA onto Brown University's high-performance computing cluster at the Center for Computation and Visualization. The fastq files were aligned using Cell Ranger (v 5.0.0) count and then aggregated using Cell Ranger aggr. The resulting output from aggr was used as input for Seurat (v 3.9.9) in RStudio (R v 4.0.2) (Stuart et al., 2019). Seurat was used to select for *Dazl*-positive (*Dazl* > 0), high-quality (nFeature_RNA >

1000, nFeature_RNA < 5000, nCount_RNA > 2500, nCount < 30000, percent mitochondrial genes < 5%) oocytes. This data was then passed to Monocle3 (v 0.2.3) for pseudotime analysis and generating UMAP and gene expression (Cao et al., 2019; Qiu et al., 2017; Trapnell et al., 2014). The cloupe file created from Cell Ranger aggr was used as input for Loupe Cell Browser (v 5.0), where the same filtering steps were used (*Dazl* > 0, Feature Threshold > 1000, Feature Threshold < 5000, UMI Threshold > 2500, UMI Threshold < 30000, Mitochondrial UMIs < 5%). These filtered cells were then split into *Taf4b*-expressing (*Taf4b* > 0) and *Taf4b*-off (*Taf4b* = 0) and then "Locally Distinguishing" was run for Significant Feature Comparison. The list of genes significantly associated with *Taf4b*-expressing cells (**Table S1**) was used as input for ClusterProfiler (v 3.16.1) to create a dotplot of significantly enriched gene ontology (GO) categories (Yu et al., 2012).

RNA-sequencing

Embryonic germ cells resuspended in Trizol were shipped to GENEWIZ (GENEWIZ Inc., NJ) on dry ice. Sample RNA extraction, sample QC, library preparation, sequencing, and initial bioinformatics were done at GENEWIZ. RNA was extracted following the Trizol Reagent User Guide (Thermo Fisher Scientific). Glycogen was added (1 µL, 10 mg/mL) to the supernatant to increase RNA recovery. RNA was quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and RNA integrity was checked with TapeStation (Agilent Technologies, Palo Alto, CA, USA) to see if the concentration met the requirements.

SMART-Seq v4 Ultra Low Input Kit for Sequencing was used for full-length cDNA synthesis and amplification (Clontech, Mountain View, CA), and Illumina Nextera XT library was used for sequencing library preparation. The sequencing libraries were multiplexed and clustered on a lane of a flowcell. After clustering, the flowcell was loaded onto an Illumina HiSeq 4000 according to manufacturer's instructions. The samples were sequenced using a 2x150 Paired End (PE) configuration. Image analysis and base calling were conducted by the HiSeq Control Software (HCS) on the HiSeq instrument. Raw sequence data (.bcl files) generated from Illumina HiSeq were converted into fastq files and de-multiplexed using bcl2fastq (v. 2.17). One mismatch was allowed for index sequence identification.

RNA-seq data analysis

Datasets SRP059601 and SRP059599 were from NCBI SRA. All raw fastq files were initially processed on Brown University's high-performance computing cluster. Reads were quality-trimmed and had adapters removed using Trim Galore! (v 0.5.0) with the parameters –nextera -q 10. Samples before and after trimming were analyzed using FastQC (v 0.11.5) for quality and then aligned to the Ensembl GRCm38 using HiSat2 (v 2.1.0) (Andrews, 2010; Pertea et al., 2016). Resulting sam files were converted to bam files using Samtools (v 1.9) (Li et al., 2009). E14.5 heterozygous bam files were downsampled because these samples had been sequenced more deeply than their wild- type and deficient counterparts.

To obtain TPMs for each sample, StringTie (v 1.3.3b) was used with the optional parameters -A and -e. A gtf file for each sample was downloaded and, using RStudio (R v 4.0.2), TPMs of all samples were aggregated into one comma separated (csv) file using a custom R script. To create interactive Microsoft Excel files for exploring the TPMs of each dataset: the csv of aggregated

TPMs was saved as an Excel spreadsheet, colored tabs were added to set up different comparisons, and a flexible Excel function was created to adjust to gene name inputs. To explore the Excel files, please find the appropriate tab (named "Quick_Calc") and type in the gene name of interest into the highlighted yellow boxes. There is an Excel file for each dataset analyzed as a supplementary table.

To obtain count tables, featurecounts (Subread v 1.6.2) was used (Liao et al., 2014). Metadata files for dataset were created manually in Excel and saved as a csv. These count tables were used to create PCA plots by variance-stabilizing transformation (vst) of the data in DESeq2 (v 1.22.2) and plotting by ggplot2 (v 3.1.0) (Love et al., 2014; Wickham, 2016). DESeq2 was also used for differential gene expression analysis, where count tables and metadata files were used as input. We accounted for the litter effect in our mouse oocytes by setting it as a batch parameter in DESeq2. For the volcano plot, the output of DESeq2 was used and plotted using ggplot2. DEG lists were used for ClusterProfiler (v 3.16.1) input to create dotplots of significantly enriched gene ontology (GO) categories for all DEGs, Downregulated DEGs, and Upregulated DEGs. Physical, highest-confidence protein-protein interactions were identified using STRING, with unconnected not shown in the image (Szklarczyk et al., 2019).

For X chromosome analysis, expected numbers of Downregulated and Upregulated DEGs per chromosome were calculated by dividing the average number of observations per chromosome by the average number of total genes per chromosome. Chi-square values and p-values were calculated using the GraphPad QuickCalcs chi-square function, where observed and expected frequencies were used as input (https://www.graphpad.com/quickcalcs/chisquared1/, accessed Jan 2021). Boxplots of $\log_2$ fold change between the autosomes and X chromosomes used the output of DESeq2 as input, based on other publications comparing autosomal and X

chromosome expression (Hirota et al., 2018). The X:A ratio was calculated using pairwiseCI (v. 0.1.27), a bootstrapping R package, after filtering genes for an average TPM > 1 (Duan et al., 2019a; Sangrithi et al., 2017). The RXE was calculated using a custom R script based after filtering genes for an average TPM > 1 and adding pseudocounts for log transformation (log2(x+1)), based on other RXE publications (Duan et al., 2019b; Jue et al., 2013). The "ubiquitous genes" from Sangrithi et al., 2017 were converted from gene names to Ensembl IDs, first by using ShinyGO to convert IDs through the "Genes" tab (Ge et al., 2020). Genes that were not mapped were then used as input for DAVID gene ID conversion, any remaining unconverted gene names were manually entered into the Ensembl database to find matches (Howe et al., 2021; Huang et al., 2007). Venn diagrams were created using BioVenn (Hulsen et al., 2008). All plots produced in RStudio were saved as an EPS file type and then opened in Adobe Illustrator in order to export a high-quality JPEG image.

CUT&RUN

The CUT&RUN performed on E16.5 germ cells followed the protocol in Hainer and Fazzio, 2019. CUT&RUN antibodies were as follows: polyclonal rabbit TAF4b (as previously described (Grive et al., 2016)), monoclonal rabbit H3K4me3 (EMD Millipore # 05-745R), rabbit IgG (ThermoFisher # 02-6102), pA-MNase was a generous gift from Dr. Thomas Fazzio.

For library preparation, the KAPA HyperPrep kit (Roche Cat. No 07962363001) was used with New England Biolabs NEBNext Multiplex Oligos for Illumina (NEB #E7335). After library amplification through PCR, libraries were size selected through gel extraction (~150-650 bp) and cleaned up using the Qiagen QIAquick Gel Extraction Kit (Cat. # 28704). CUT&RUN libraries

in EB buffer were shipped to GENEWIZ (GENEWIZ Inc., NJ) on dry ice. Sample QC, sequencing, and initial bioinformatics were done at GENEWIZ.

The sequencing libraries were validated on the Agilent TapeStation (Agilent Technologies, Palo Alto, CA, USA), and quantified by using Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) as well as by quantitative PCR (KAPA Biosystems, Wilmington, MA, USA). The sequencing libraries were clustered on flowcells. After clustering, the flowcells were loaded on to the Illumina HiSeq instrument (4000 or equivalent) according to manufacturer's instructions. The samples were sequenced using a 2x150bp Paired End (PE) configuration. Raw sequence data (.bcl files) generated from Illumina HiSeq were converted into fastq files and de-multiplexed using bcl2fastq (v. 2.20). One mismatch was allowed for index sequence identification.

CUT&RUN data analysis

Computational scripts regarding CUT&RUN data analysis were based on other CUT&RUN publications (Hainer and Fazzio, 2019). All raw fastq files were initially processed on Brown University's high-performance computing cluster. Reads were quality- trimmed and had adapters removed using Trim Galore! (v 0.5.0) with the parameter -q 10 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Samples before and after trimming were analyzed using FastQC (v 0.11.5) for quality and then aligned to the Ensembl GRCm39 using Bowtie2 (v 2.3.0). Fastq screen (v 0.13.0) was used to determine the percentage of reads uniquely mapped to the mouse genome in comparison to other species. Resulting sam files were converted to bam files, then unmapped, duplicated reads, and low quality mapped were removed using Samtools (v1.9). Resulting bam files were split into size classes using a Unix script. For calling peaks, annotating peaks, and identifying coverage around TSSs, Homer (v 4.10) was used

(Heinz et al., 2010). For gene track visualization, the final bam file before splitting into size classes was used as input to Integrative Genomics Viewer (IGV) (Robinson et al., 2011). A custom genome was created using a genome fasta and gtf file for Ensembl GRCm39.

Pie charts were created using data from Homer output and Venn diagrams were created using BioVenn. For X chromosome analysis, expected numbers of promoter peaks per chromosome were calculated by dividing the average number of observations per chromosome by the average number of total protein-coding genes per chromosome. Chi-square values and p-values were calculated using the GraphPad QuickCalcs chi-square function, where observed and expected frequencies were used as input (https://www.graphpad.com/quickcalcs/chisquared1/, accessed Apr 2021). Dotplots of Promoter-TSS peaks were made using ClusterProfiler. To find which TAF4b peaks were shared between female replicates and RNA-seq DEGs, the Ensembl ID associated with the annotation was used. TSS plots were created using the "tss" function of Homer and plotted using Microsoft Excel. All plots produced in RStudio were saved as an EPS file type and then opened in Adobe Illustrator in order to export a high-quality JPEG image.

Quantitative real-time PCR (qRT-PCR)

Total RNA was extracted from E17.5 female Oct4-EGFP^{+} germ cells using a Direct-zol RNA miniprep kit (Zymo Research, R2027, C1004). Total RNA from all experiments was quantified and checked for purity, and 50 ng was used to prepare 20 µL of cDNA with an iScript cDNA Synthesis Kit (Bio-Rad, 170-8891). Real-time PCR was performed in technical triplicate using 1 µL of DNA template, 12.5 µL of ABI SYBR green PCR master-mix (Applied Biosystems, A25742), and 0.5 µM custom oligos (Invitrogen) for *Fmr1, Sp1, Fam83d, JunD, Taf4b,* or *18S* rRNA in a 20 µL reaction in an ViiA 7 Real Time PCR machine (Life Technologies). Data were

analyzed by the ΔΔCt method, and relative expression levels were normalized to *18S* rRNA. Primer sequences corresponding to genes of interest can be found in **Table S11**.

Ovarian immunofluorescence

Prenatal ovaries were harvested at E16.5, cleaned of excess fat, and fixed in 4% formaldehyde solution for two hours before embedding in Optimal Cutting Temperature (OCT) compound. Ovaries were serially sectioned at 8 μm on a Leica Cryostat onto glass slides and washed in 1X Phosphate-Buffered Saline (PBS) containing 0.1% Tween-20 (Fisher Scientific). Tissue sections were then incubated in blocking buffer [5% Goat Serum (Sigma), and 0.1% Tween-20 (Fisher Scientific) in 1X PBS] for one hour at room temperature. Standard immunofluorescence was performed and slides were stained with rat anti-Tra98 (Abcam, ab82527) and rabbit anti-FMRP (Abcam, ab17722) primary antibodies at a 1:100 dilution. Slides were stained with goat anti-rat Alexa 594 (Invitrogen, A11007) and goat anti-mouse Alexa 488 (Abcam, ab150077) secondary antibodies at a 1:500 dilution and counterstained with 4′,6-diamidino-2-phenylindole (DAPI; Vector Laboratories, Burlingame, CA). A secondary antibody-only control was included to compare background staining. Images were taken at 40X magnification on a Zeiss Axio Imager M1 microscope (Carl Zeiss, Inc., Thornwood, NY).

Quantification of fluorescence intensity

Ovary tissue sections from two E16.5 *Taf4b*-wildtype and *Taf4b*-deficient animals were stained with FMRP/Tra98 and used for fluorescence intensity quantification. Five spatially distributed sections from each ovary were imaged at 40X magnification on a Zeiss Axio Imager M1 microscope (Carl Zeiss, Inc., Thornwood, NY). Fiji-ImageJ software was used to visualize

images and perform quantification of fluorescence intensity (Schindelin et al., 2012). First, germ cell clusters were identified within each tissue section and cropped to generate a 60 μm square image containing the red (Tra98), green (FMRP), and blue (DAPI) channels. The freehand tool was then used to outline the germ cell cluster within the cropped image. Next, either the FMRP or Tra98 channel was selected, and the "Measure" functionality under the "Analyze" menu was used to quantify the mean pixel intensity within the outlined area. The mean FMRP and Tra98 pixel intensity was measured from a total of 53 wildtype and 70 *Taf4b*-deficient germ cell clusters. The mean and standard deviation of the values collected for each genotype were calculated and statistical significance was calculated using an unpaired two-tailed Student's T-test. Jitter plots were generated by using the ggplot2 (v 3.1.0) package in R. All plots produced in RStudio were saved as an EPS file type and then opened in Adobe Illustrator in order to export a high-quality JPEG image.

Data availability

All computational scripts used in this publication are available to the public at the following URL:

https://github.com/mg859337/Gura_et_al._TAF4b_transcription_networks_regulating_early_oocyte_differentiation. The female mouse E14.5 and E16.5 RNA-sequencing data are available from NCBI GEO under accession number GSE174366 and NCBI SRA under accession numbers SRP319538 and SRP319541. The female mouse E16.5 CUT&RUN sequencing data are available from NCBI GEO under accession number GSE186991 and NCBI SRA under accession number SRP344210. The sequencing datasets accessed in this research are from the follow accession numbers: the scRNA-seq mouse data from E12.5 to E16.5 female sorted Oct4-EGFP

ovaries used for Figure 1 was obtained through NCBI GEO: GSE130212; the scRNA-seq mouse data from E11.5 to E14.5 female ovaries used for Figure S1 was obtained through NCBI GEO: GSE128553; the RNA-seq mouse data from E9.5 to E18.5 XX and XO sorted Oct4-EGFP ovaries used for Figure 4 and Figure S6 was obtained through ArrayExpress: E-MTAB-4616; the RNA-seq of *in vitro* differentiated XX and XO mouse ESCs used for Figure S7 was obtained through NCBI GEO: GSE121299.

ACKNOWLEDGMENTS

We thank Drs. Ashley Webb, Erica Larschan, Mark Johnson, and Kathryn Grive for their helpful input throughout these studies. We thank the Center for Computation and Visualization at Brown University for computational resources for scRNA-seq, RNA-seq, and CUT&RUN data analysis. We thank Kevin Carlson and the Brown University Flow Cytometry and Sorting Facility for expertise completing the flow sorting. The Brown University Flow Cytometry and Sorting Facility has received generous support in part by the National Institutes of Health (NCRR Grant No. 1S10RR021051) and the Division of Biology and Medicine, Brown University. As much of our insights were gained by reprocessing publicly available datasets, we greatly appreciate both the researchers that generated and shared the data initially and the respective repositories for making them available. Figure 8 was made using BioRender. We are grateful to the NICHD/NIH for their generous support through awards 1F31HD097933, 1F31HD105340, and 1R01HD091848 to MAG, KMA and RNF, respectively and thank the BSF for their generous support.

Author Contributions: MAG, conception and design of all experiments, collection and assembly of data, data analysis and interpretation, manuscript writing; SR, collection and assembly of data, data analysis and interpretation; KMA, collection and assembly of data, data analysis and interpretation; KAS, collection and assembly of data, mouse colony management, collection of ovarian tissue and cells, data analysis and interpretation; TW, reagents and technical expertise for CUT&RUN data collection, analysis, experimental design, and data interpretation; HK, technical expertise and data interpretation; JMAT, technical expertise and data interpretation; TGF, reagents and technical expertise for CUT&RUN data collection and analysis, experimental design and data interpretation; RNF, conception and design of all experiments, data interpretation, manuscript writing and financial support.

REFERENCES

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/* http://www.bioinformatics.babraham.ac.uk/projects/.

Antonova, S. V., Boeren, J., Timmers, H. T. M. and Snel, B. (2019). Epigenetics and transcription regulation during eukaryotic diversification: the saga of TFIID. *Genes Dev.* **33**, 1–15.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502.

Chandra, A., Copen, C. E. and Stephen, E. H. (2013). Infertility and impaired fecundity in the United States, 1982-2010: data from the National Survey of Family Growth. *Natl. Health Stat. Report.*

Chitiashvili, T., Dror, I., Kim, R., Hsu, F. M., Chaudhari, R., Pandolfi, E., Chen, D., Liebscher, S., Schenke-Layland, K., Plath, K., et al. (2020). Female human primordial germ cells display X-chromosome dosage compensation despite the absence of X-inactivation. *Nat. Cell Biol.* **22**, 1436–1446.

Duan, J., Shi, W., Jue, N. K., Jiang, Z., Kuo, L., O'Neill, R., Wolf, E., Dong, H., Zheng, X., Chen, J., et al. (2019a). Dosage compensation of the X chromosomes in bovine germline, early embryos, and somatic tissues. *Genome Biol. Evol.* **11**, 242–252.

Duan, J., Flock, K., Jue, N., Zhang, M., Jones, A., Al Seesi, S., Mandoiu, I., Pillai, S., Hoffman, M., O'Neill, R., et al. (2019b). Dosage compensation and gene expression of the X Chromosome in sheep. *G3 Genes, Genomes, Genet.* **9**, 305–314.

Dynlacht, B. D., Hoey, T. and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell* **66**, 563–576.

Falender, A. E., Shimada, M., Lo, Y. K. and Richards, J. S. (2005). TAF4b, a TBP associated factor, is required for oocyte development and function. *Dev. Biol.* **288**, 405–419.

Feng, C. W., Bowles, J. and Koopman, P. (2014). Control of mammalian germ cell entry into meiosis. *Mol. Cell. Endocrinol.* **382**, 488–497.

Fink, D. A., Nelson, L. M., Pyeritz, R., Johnson, J., Sherman, S. L., Cohen, Y. and Elizur, S. E. (2018). Fragile X Associated Primary Ovarian Insufficiency (FXPOI): Case Report and Literature Review. *Front. Genet.* **9**, 1–12.

Fortuño, C. and Labarta, E. (2014). Genetics of primary ovarian insufficiency: a review. *J. Assist. Reprod. Genet.* **31**, 1573–1585.

Freiman, R. N. (2009). Specific variants of general transcription factors regulate germ cell development in diverse organisms. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1789**, 161–166.

Gazdag, E., Santenard, A., Ziegler-Birling, C., Altobelli, G., Poch, O., Tora, L. and Torres-Padilla, M. E. (2009). TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. *Genes Dev.* **23**, 2210–2223.

Ge, S. X., Jung, D. and Yao, R. (2020). ShinyGO: A graphical enrichment tool for ani-mals and plants. *Bioinformatics* **36**, 2628–2629.

Ge, W., Wang, J. J., Zhang, R. Q., Tan, S. J., Zhang, F. L., Liu, W. X., Li, L., Sun, X. F., Cheng, S. F., Dyce, P. W., et al. (2021). Dissecting the initiation of female meiosis in the mouse at single-cell resolution. *Cell. Mol. Life Sci.* **78**, 695–713.

Goodrich, J. A. and Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat. Rev. Genet.* **11**, 549–558.

Gravholt, C. H., Brun, S. and Andersen, N. H. (2019). Turner syndrome: mechanisms and management. *Nat. Rev. Endocrinol.*

Grive, K. J., Seymour, K. a., Mehta, R. and Freiman, R. N. (2014). TAF4b promotes mouse primordial follicle assembly and oocyte survival. *Dev. Biol.* **392**, 42–51.

Grive, K. J., Gustafson, E. A., Seymour, K. A., Baddoo, M., Schorl, C., Golnoski, K., Rajkovic, A., Brodsky, A. S. and Freiman, R. N. (2016). TAF4b Regulates Oocyte-Specific Genes Essential for Meiosis. *PLoS Genet.* **12**, 1–18.

Gura, M. A. and Freiman, R. N. (2018). Primordial Follicle. *Encycl. Reprod.* 65–71.

Gura, M. A., Mikedis, M. M., Seymour, K. A., De Rooij, D. G., Page, D. C. and Freiman, R. N. (2020). Dynamic and regulated TAF gene expression during mouse embryonic germ cell development. *PLoS Genet.* **16**, 1–28.

Hainer, S. J. and Fazzio, T. G. (2019). High-Resolution Chromatin Profiling Using CUT&RUN. *Curr. Protoc. Mol. Biol.* **126**, 1–22.

Hamada, N., Hamazaki, N., Shimamoto, S., Hikabe, O., Nagamatsu, G., Takada, Y., Kato, K. and Hayashi, K. (2020). Germ cell-intrinsic effects of sex chromosomes on early oocyte differentiation in mice. *PLoS Genet.* **16**, 1–26.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589.

Herrera, F. J., Yamaguchi, T., Roelink, H. and Tjian, R. (2014). Core promoter factor TAF9B regulates neuronal gene expression. *Elife*.

Hibino, E., Inoue, R., Sugiyama, M., Kuwahara, J., Matsuzaki, K. and Hoshino, M. (2016). Interaction between intrinsically disordered regions in transcription factors Sp1 and TAF4. *Protein Sci.* **25**, 2006–2017.

Hiller, M., Chen, X., Pringle, M. J., Suchorolski, M., Sancak, Y., Viswanathan, S., Bolival, B., Lin, T.-Y., Marino, S. and Fuller, M. T. (2004). Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. *Development* **131**, 5297–5308.

Hirota, T., Blakeley, P., Sangrithi, M. N., Mahadevaiah, S. K., Encheva, V., Snijders, A. P., ElInati, E., Ojarikre, O. A., de Rooij, D. G., Niakan, K. K., et al. (2018). SETDB1 Links the Meiotic DNA Damage Response to Sex Chromosome Silencing in Mice. *Dev. Cell* **47**, 645-659.e6.

Hoey, T., Dynlacht, B. D., Peterson, M. G., Pugh, B. F. and Tjian, R. (1990). Isolation and characterization of the Drosophila gene encoding the TATA box binding protein, TFIID. *Cell* **61**, 1179–1186.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2007). DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, 169–175.

Jue, N. K., Murphy, M. B., Kasowitz, S. D., Qureshi, S. M., Obergfell, C. J., Elsisi, S., Foley, R. J., O'Neill, R. J. and O'Neill, M. J. (2013). Determination of dosage compensation of the mammalian X chromosome by RNA-seq is dependent on analytical approach. *BMC Genomics* **14**,.

Lawrence, E. J., Gao, H., Tock, A. J., Lambing, C., Blackwell, A. R., Feng, X. and Henderson, I. R. (2019). Natural Variation in TBP-ASSOCIATED FACTOR 4b Controls Meiotic Crossover and Germline Transcription in Arabidopsis. *Curr. Biol.*

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

Liao, Y., Smyth, G. K. and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930.

Lovasco, L. a, Seymour, K. a, Zafra, K., O'Brien, C. W., Schorl, C. and Freiman, R. N. (2010a). Accelerated ovarian aging in the absence of the transcription regulator TAF4B in mice. *Biol. Reprod.* **82**, 23–34.

Lovasco, L. A., Seymour, K. A., Zafra, K., O'Brien, C. W., Schorl, C. and Freiman, R. N. (2010b). Accelerated ovarian aging in the absence of the transcription regulator TAF4B in mice. *Biol Reprod* **82**, 23–34.

Lovasco, L. A., Gustafson, E. A., Seymour, K. A., De Rooij, D. G. and Freiman, R. N. (2015). TAF4b is required for mouse spermatogonial stem cell development. *Stem Cells* **33**, 1267–1276.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.

Mechta-Grigoriou, F., Gerald, D. and Yaniv, M. (2001). The mammalian Jun proteins: Redundancy and specificity. *Oncogene* **20**, 2378–2389.

Nardini, M., Gnesutta, N., Donati, G., Gatta, R., Forni, C., Fossati, A., Vonrhein, C., Moras, D., Romier, C., Bolognesi, M., et al. (2013). Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* **152**, 132–143.

Ohno, S. (1966). *Sex Chromosomes and Sex-Linked Genes*. Berlin, New York: Springer-Verlag.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.*

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982.

Reinberg, D., Horikoshi, M. and Roeder, R. G. (1987). Factors involved in specific transcription in mammalian RNA polymerase II. Functional analysis of initiation factors IIA and IID and identification of a new factor operating at sequences downstream of the initiation site. *J. Biol. Chem.* **262**, 3322–3330.

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**, 24–26.

Romer, K. A., de Rooij, D. G., Kojima, M. L. and Page, D. C. (2018). Isolating mitotic and meiotic germ cells from male mice by developmental synchronization, staging, and sorting. *Dev. Biol.*

Rossetti, R., Ferrari, I., Bonomi, M. and Persani, L. (2017). Genetics of primary ovarian insufficiency. *Clin. Genet.* **91**, 183–198.

Sangrithi, M. N., Royo, H., Mahadevaiah, S. K., Ojarikre, O., Bhaw, L., Sesay, A., Peters, A. H. F. M., Stadler, M. and Turner, J. M. A. (2017). Non-Canonical and Sexually Dimorphic X Dosage Compensation States in the Mouse and Human Germline. *Dev. Cell*.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386.

Vizcaíno, C., Mansilla, S. and Portugal, J. (2015). Sp1 transcription factor: A long-standing target in cancer chemotherapy. *Pharmacol. Ther.* **152**, 111–124.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer-Verlag.

Yu, G., Wang, L. G., Han, Y. and He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287.

Zhang, Q., Yu, S., Lok, S. I. S., Wong, A. S. T., Jiao, Y. and Lee, L. T. O. (2019). FAM83D promotes ovarian cancer progression and its potential application in diagnosis of invasive ovarian cancer. *J. Cell. Mol. Med.* **23**, 4569–4581.

Zhao, Z. H., Ma, J. Y., Meng, T. G., Wang, Z. B., Yue, W., Zhou, Q., Li, S., Feng, X., Hou, Y., Schatten, H., et al. (2020). Single-cell RNA sequencing reveals the landscape of early female germ cell development. *FASEB J.* **34**, 12634–12645.

Zhou, H., Grubisic, I., Zheng, K., He, Y., Wang, P. J., Kaplan, T. and Tjian, R. (2013a). Taf7l cooperates with Trf2 to regulate spermiogenesis. *Proc. Natl. Acad. Sci.*

Zhou, H., Kaplan, T., Li, Y., Grubisic, I., Zhang, Z., Wang, P. J., Eisen, M. B. and Tjian, R. (2013b). Dual functions of TAF7L in adipocyte differentiation. *Elife*.

Figures

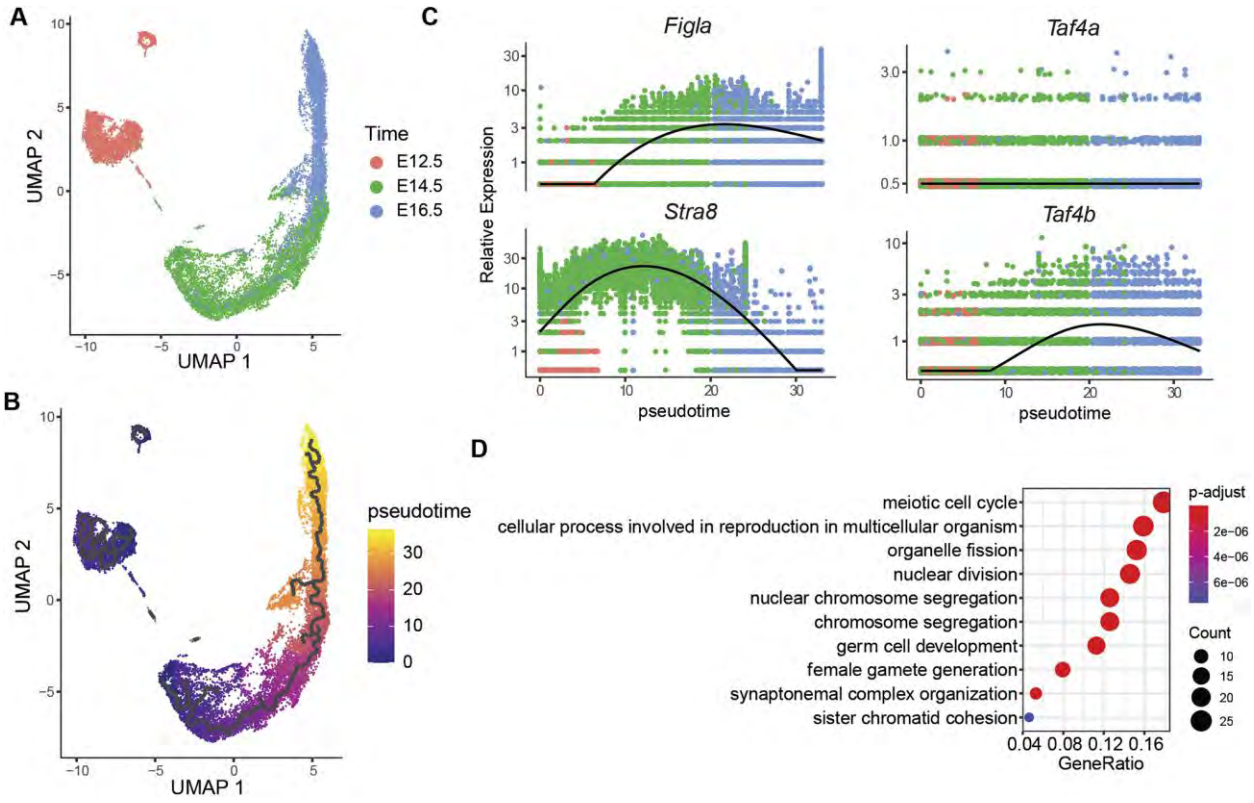


Figure 1. Analysis of scRNA-seq dataset in E12.5 to E16.5 germ cells. (A) Uniform Manifold Approximation and Projection (UMAP) of oocytes colored by embryonic time point. (B) UMAP of oocytes colored by pseudotime analysis. (C) Expression of *Figla*, *Stra8*, *Taf4b*, and *Taf4a* plotted in terms of pseudotime and colored based on embryonic time point. (D) Dotplot of GO results for genes significantly higher in *Taf4b*-expressing oocytes than *Taf4b*-non-expressing oocytes.

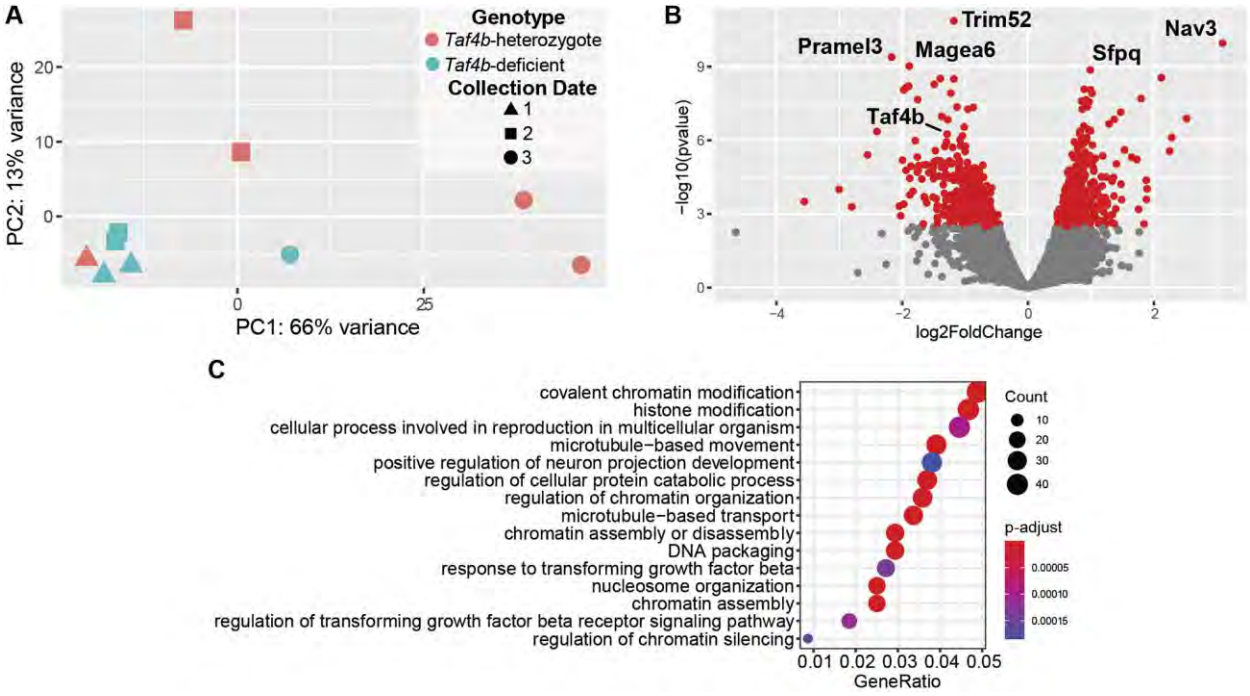


Figure 2. RNA-seq of E16.5 oocytes. (A) PCA plot of the E16.5 samples labeled based on *Taf4b* genotype and collection number. (B) Volcano plot of genes, the significant genes (protein-coding, p-adj < 0.05, avg TPM > 1) are labeled in red and the top 5 DEGs plus *Taf4b* are specified. (C) Dotplot of GO biological process analysis of DEGs.

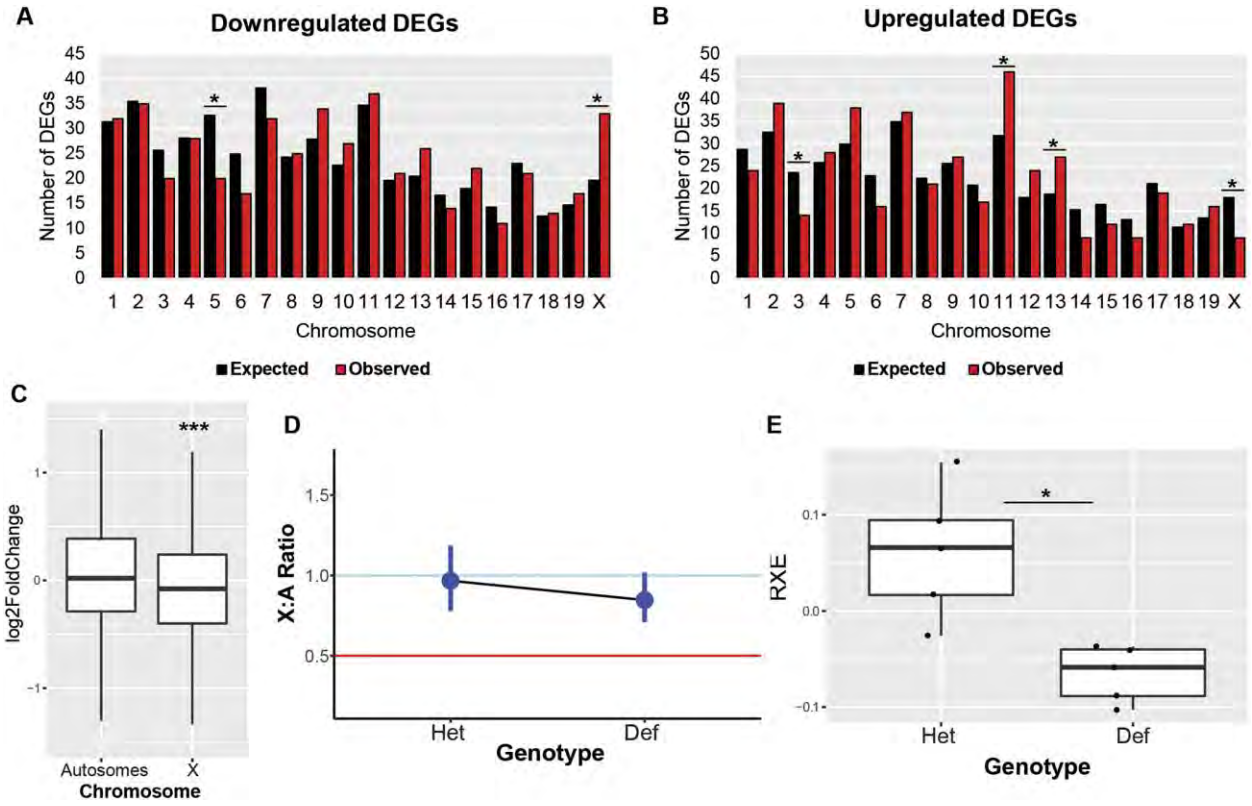


Figure 3. X chromosome gene expression in E16.5 *Taf4b*-deficient oocytes. (A-B) Graphs of expected (black bars) and observed (red bars) numbers of DEGs on each chromosome for the "Downregulated" (A) and "Upregulated" genes (B) DEGs. * = $p < 0.05$, chi-square test. (C) Boxplots of $\log_2$ fold change values from DESeq2 for genes on autosomes versus the X chromosome (outliers removed). *** = $p < 0.0001$, Welch's t-test. (D) X:A ratio plot calculated through pairwiseCI after filtering for average TPM > 1 comparing Het X:A ratio to *Taf4b*-deficient X:A Ratio. (E) Boxplots of relative X expression (RXE) calculations after filtering for average TPM > 1 and adding pseudocounts for log-transformation for *Taf4b*-heterozygous and -deficient samples, * = $p < 0.05$, Welch's t-test.

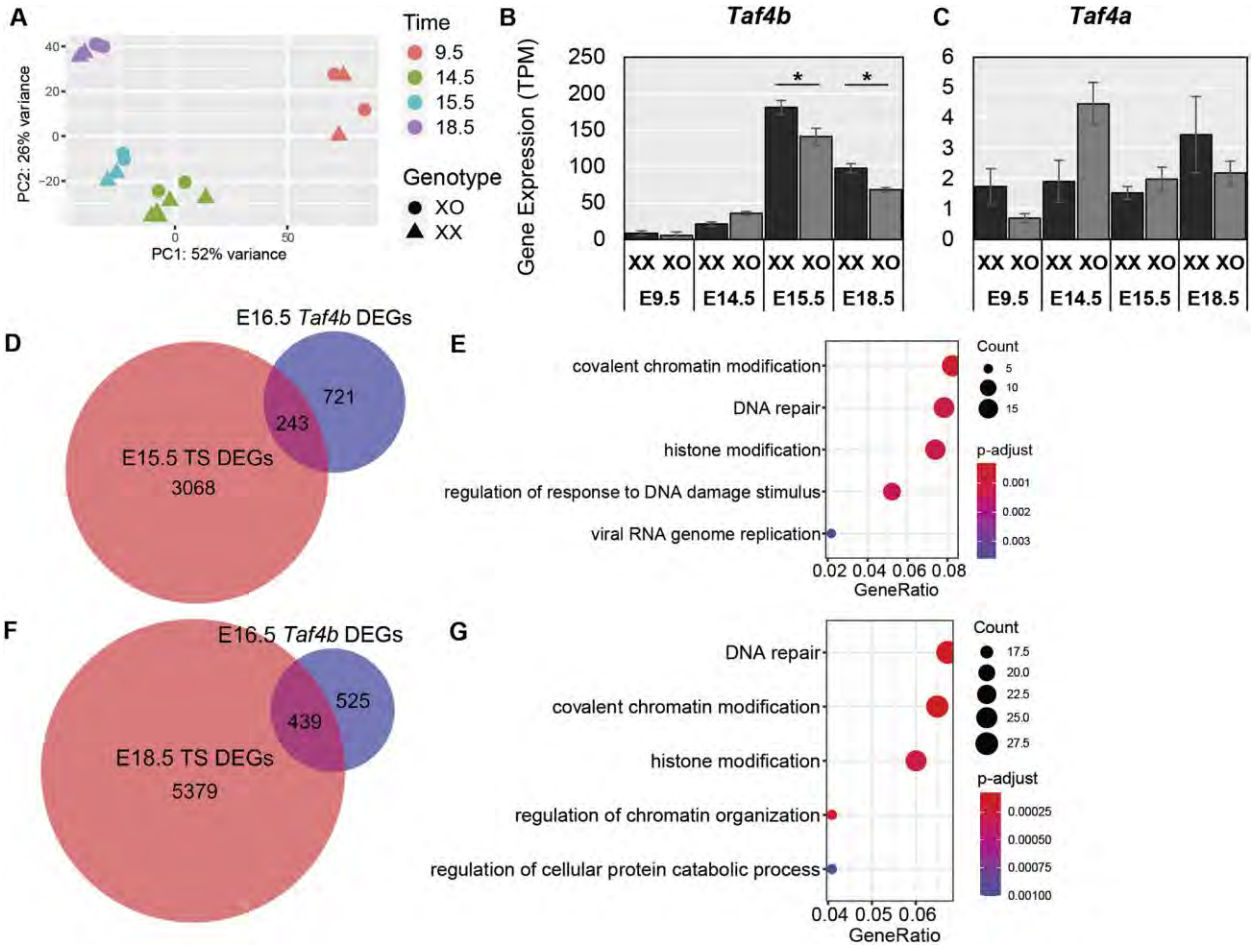


Figure 4. Effects of Turner Syndrome on *Taf4b* and similarities in transcriptomes. (A) PCA plot of the sorted oocytes from Sangrithi et al., 2017, labeled based on embryonic time point and genotype. (B) Expression levels of *Taf4b* in XX versus XO female oocytes (* = $p < 0.05$, avg TPM > 1). Error bars indicate ± standard error of the mean (SEM). (C) Expression levels of *Taf4a* in XX versus XO female oocytes. Error bars indicate ± standard error of the mean (SEM). (D) Venn diagram of E16.5 *Taf4b* DEG list compared with E15.5 TS DEGs (protein-coding, p-adj < 0.05, avg TPM > 1). Significant overlap in Venn diagram ($p < 0.0001$, hypergeometric test). (E) GO biological process dotplot for the 243 DEGs shared between E16.5 *Taf4b* and E15.5 TS RNA-seq experiments. (F) Venn diagram of E16.5 *Taf4b* DEG list compared with

E18.5 TS DEGs (protein-coding, p-adj < 0.05, avg TPM > 1). Significant overlap in Venn diagram ($p < 0.0001$, hypergeometric test). (G) GO biological process dotplot for the 439 DEGs shared between E16.5 *Taf4b* and E18.5 TS RNA-seq experiments.

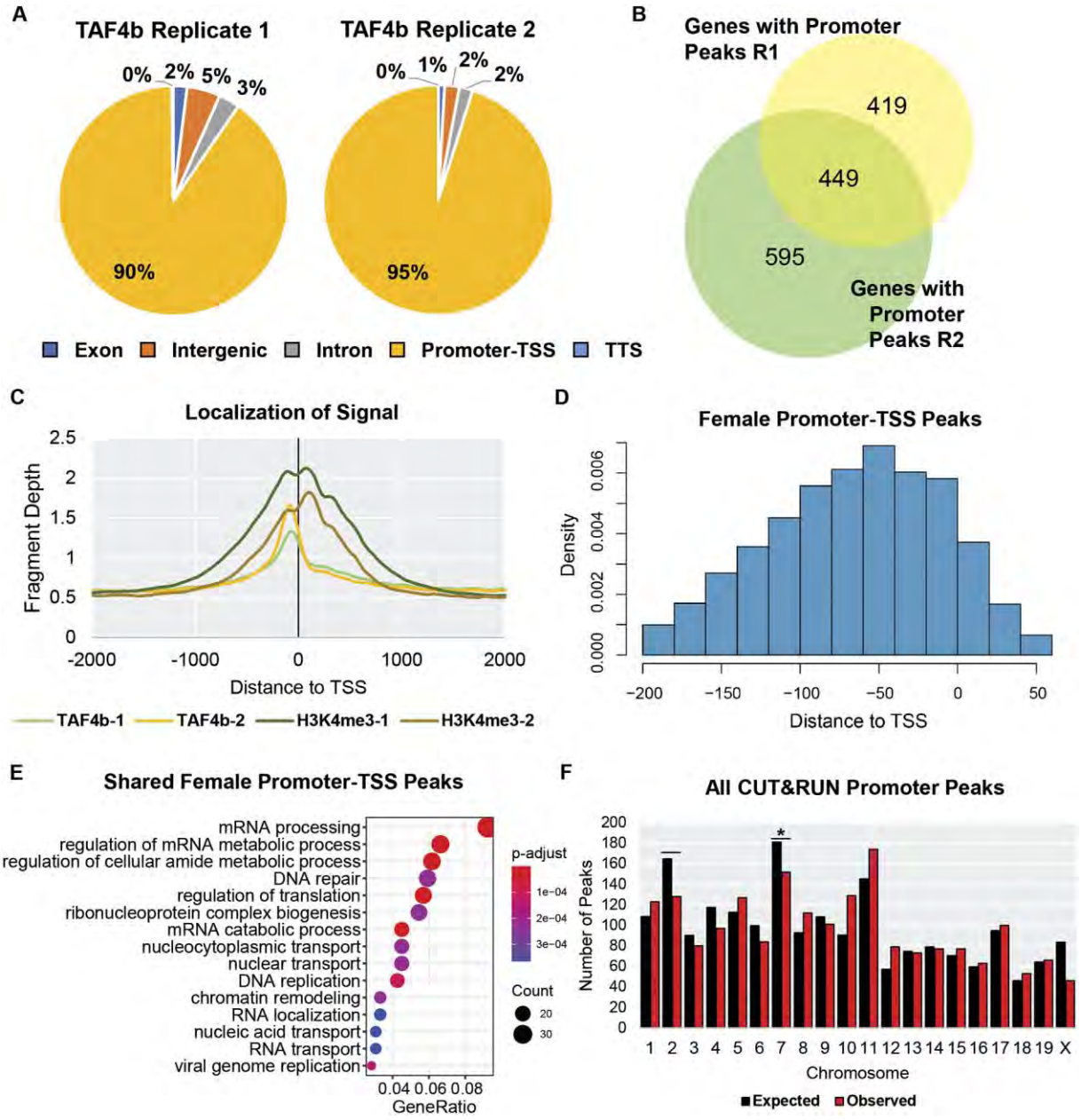


Figure 5. E16.5 germ cell CUT&RUN identifies direct targets of TAF4b. (A) Pie charts of TAF4b peak locations in female germ cell CUT&RUN replicates. (B) Venn diagram of "promoter-TSS" peaks shared between the CUT&RUN replicates. (C) Average enrichment of TAF4b and H3K4me3 signal near TSSs (dotted line) for each replicate. (D) Histogram of TAF4b

"promoter-TSS" peaks in relation to the TSS. (E) GO biological process dotplot for the shared CUT&RUN peaks categorized as "promoter-TSS". (F) Graphs of expected (black bars) and observed (red bars) numbers of promoter peaks on each chromosome in the female replicates (* = $p < 0.05$, chi-square test).

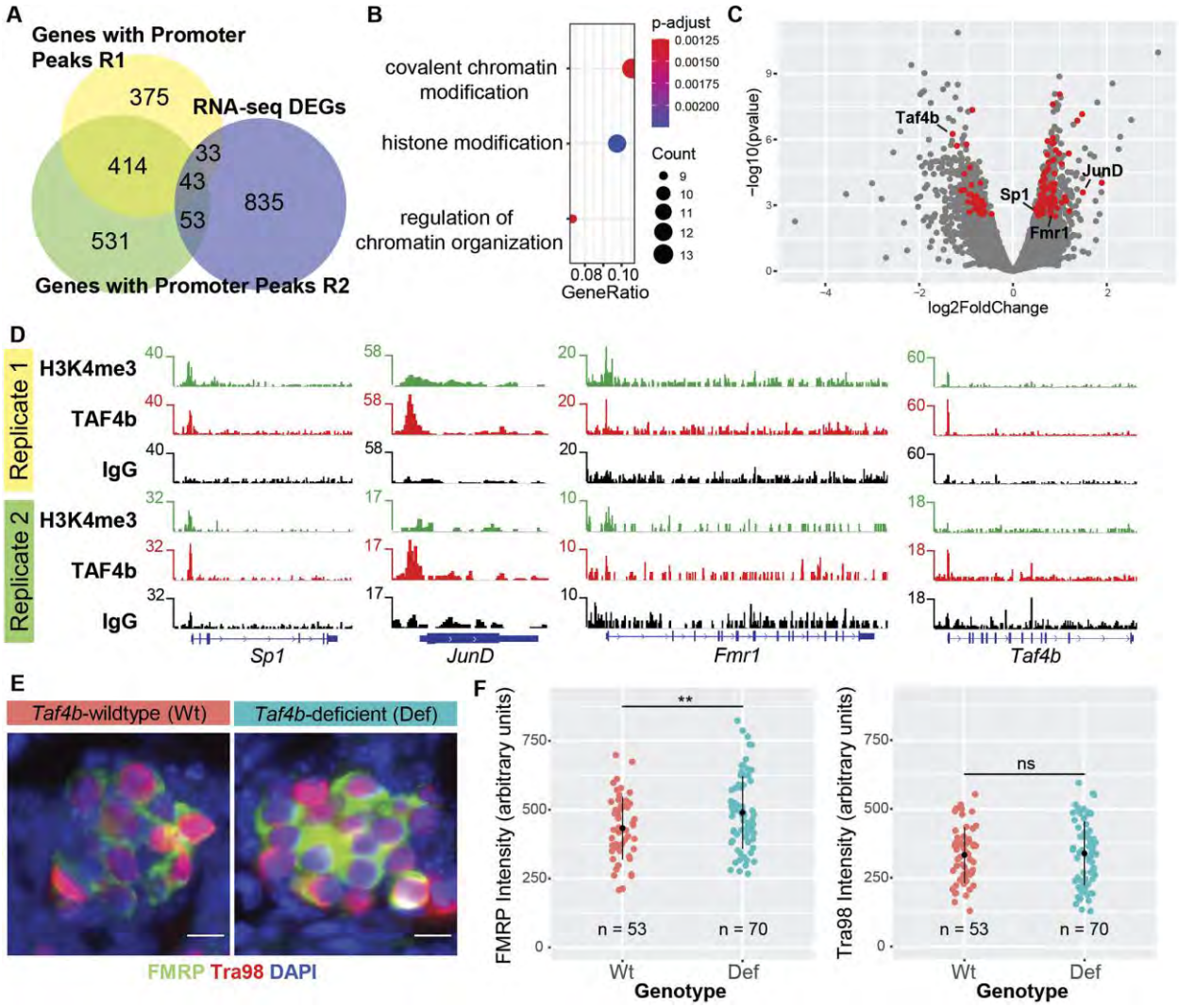


Figure 6. E16.5 CUT&RUN identifies putative direct targets of TAF4b in germ cells. (A) Venn diagram of CUT&RUN promoter peaks and RNA-seq DEGs. (B) Biological process GO dotplot of the 129 genes that are in the list of DEGs and had a promoter-TSS peak in at least one of the two germ cell samples. (C) Volcano plot of the 129 DEGs that had at least one TAF4b promoter peak (red dots). (D) Gene tracks of *Sp1*, *JunD*, *Fmr1*, and *Taf4b*, which were DEGs that had a TAF4b promoter-TSS called in both replicates. (E) Immunofluorescence images of germ cell clusters isolated from E16.5 *Taf4b*-wildtype and *Taf4b*-deficient ovary sections. Tra98

in red is a nuclear germ cell marker, FMRP is in green, and DAPI is in blue. Images were taken at 40X magnification and scalebar is 10 μM. (F) Quantification of FMRP (left) and Tra98 (right) signal intensity. Each dot represents the mean pixel intensity from an image of a single germ cell cluster, and n refers to the number of clusters that were imaged. The black dot represents the sample mean of all images collected and the black bars represent the standard deviation. P-value was calculated using an unpaired two-tailed Student's T test, ** indicates $p \leq 0.01$, ns indicates $p \geq 0.05$.

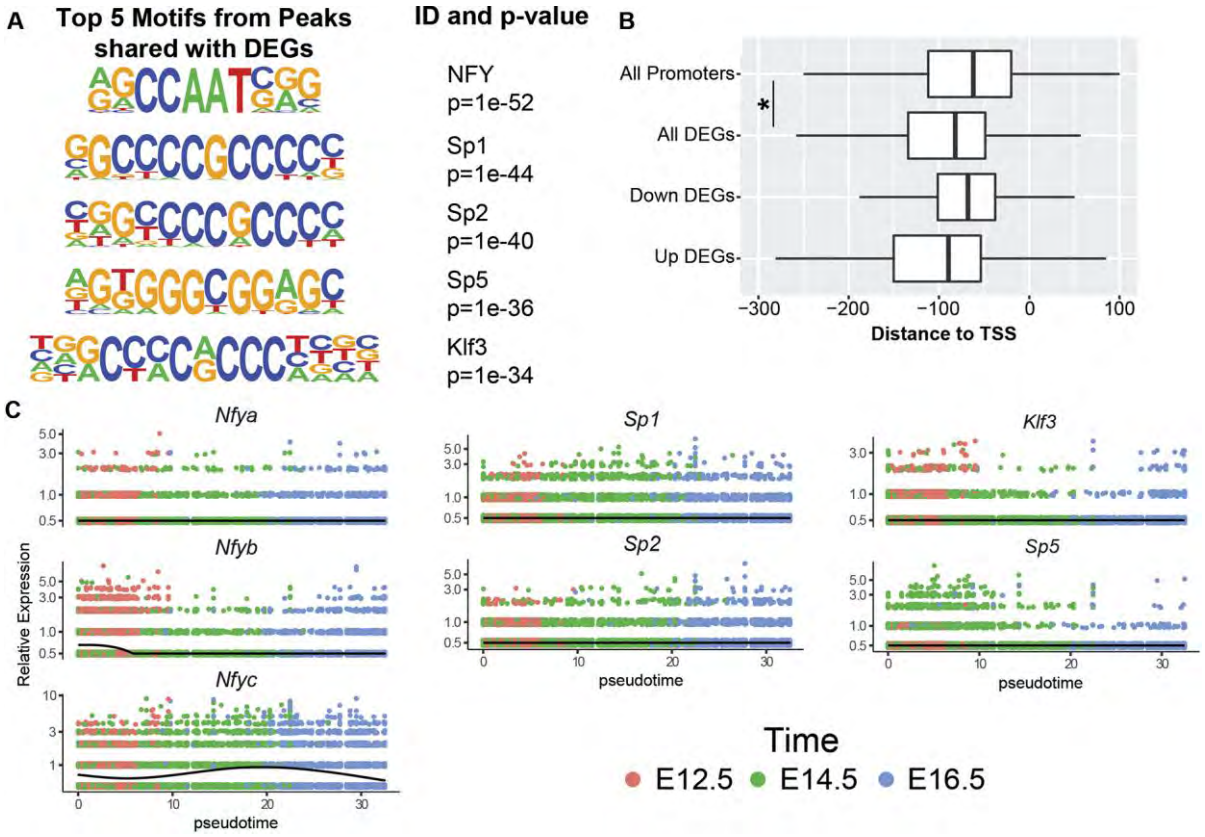


Figure 7. NFY and Sp/Klf binding sites strongly enriched in TAF4b-bound motifs. (A) Top five motifs enriched at TAF4b "promoter-TSS" peaks for genes that were also DEGs, the promoter ID, and the associated p-value. (B) Boxplots (no outliers included) of peaks relative to the TSS for all TAF4b "promoter-TSS" peaks, all TAF4b "promoter-TSS" peaks for genes that were also DEGs, TAF4b "promoter-TSS" peaks for genes that were only Downregulated DEGs, and TAF4b "promoter-TSS" peaks for genes that were only Upregulated DEGs (* = $p < 0.05$, Welch's t-test). (C) Pseudotime expression of *Nfya*, *Nfyb*, *Nfyc*, *Sp1*, *Sp2*, *Klf3*, and *Sp5* colored based on embryonic time points.

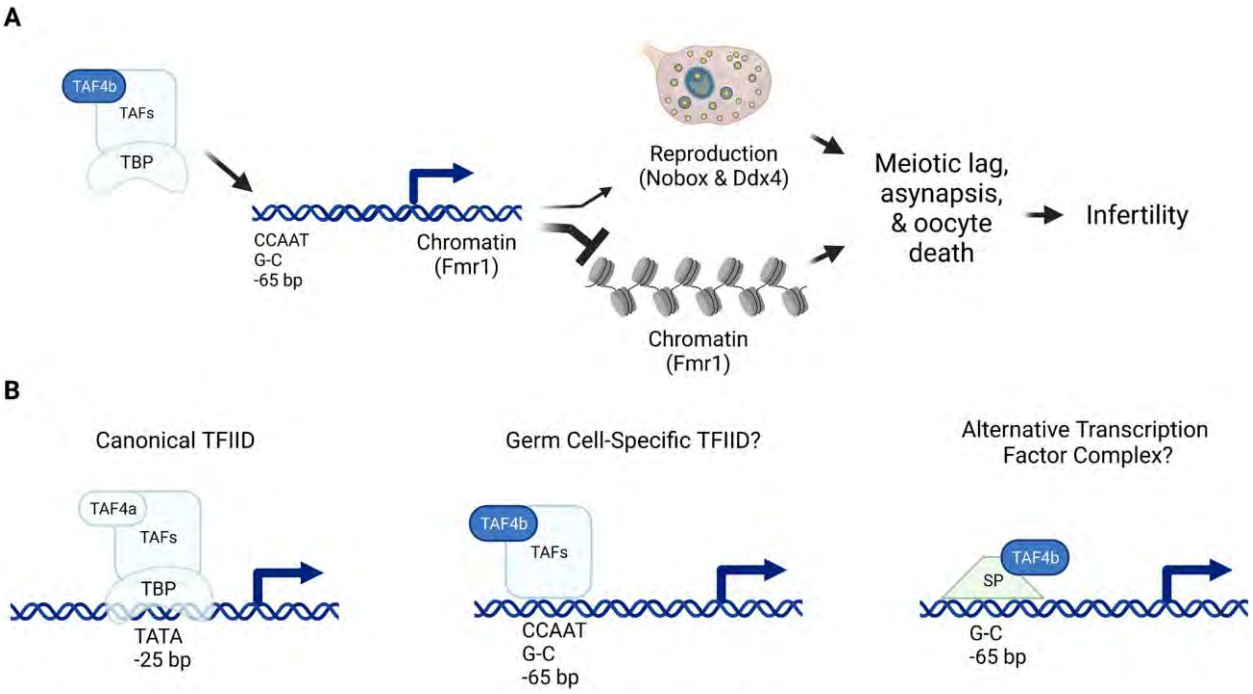


Figure 8. Current hypothesis of TAF4b function in E16.5 oocytes. (A) Model of TAF4b function in E16.5 oocytes based on RNA-seq and CUT&RUN data. (B) Hypotheses regarding molecular function of TAF4b in oocytes including modifying the canonical motifs associated with TAF4b-containing TFIID and location relative to the TSS or having completely different transcription factor binding partner(s).

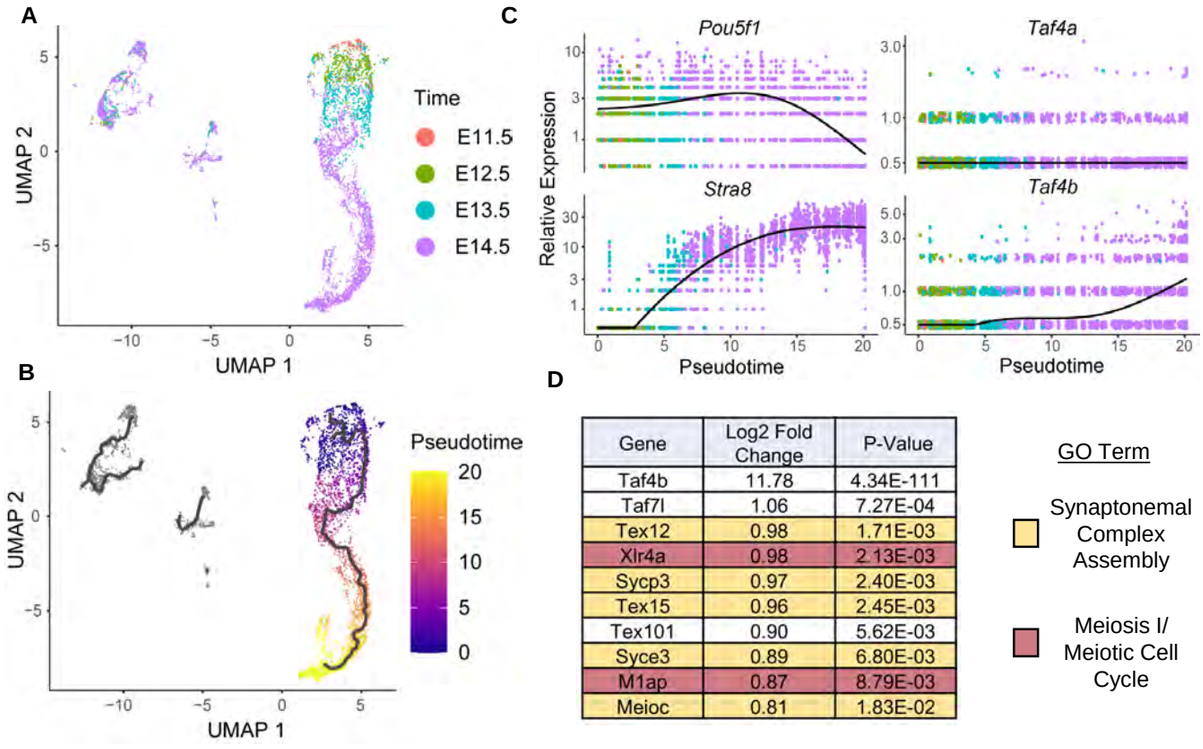
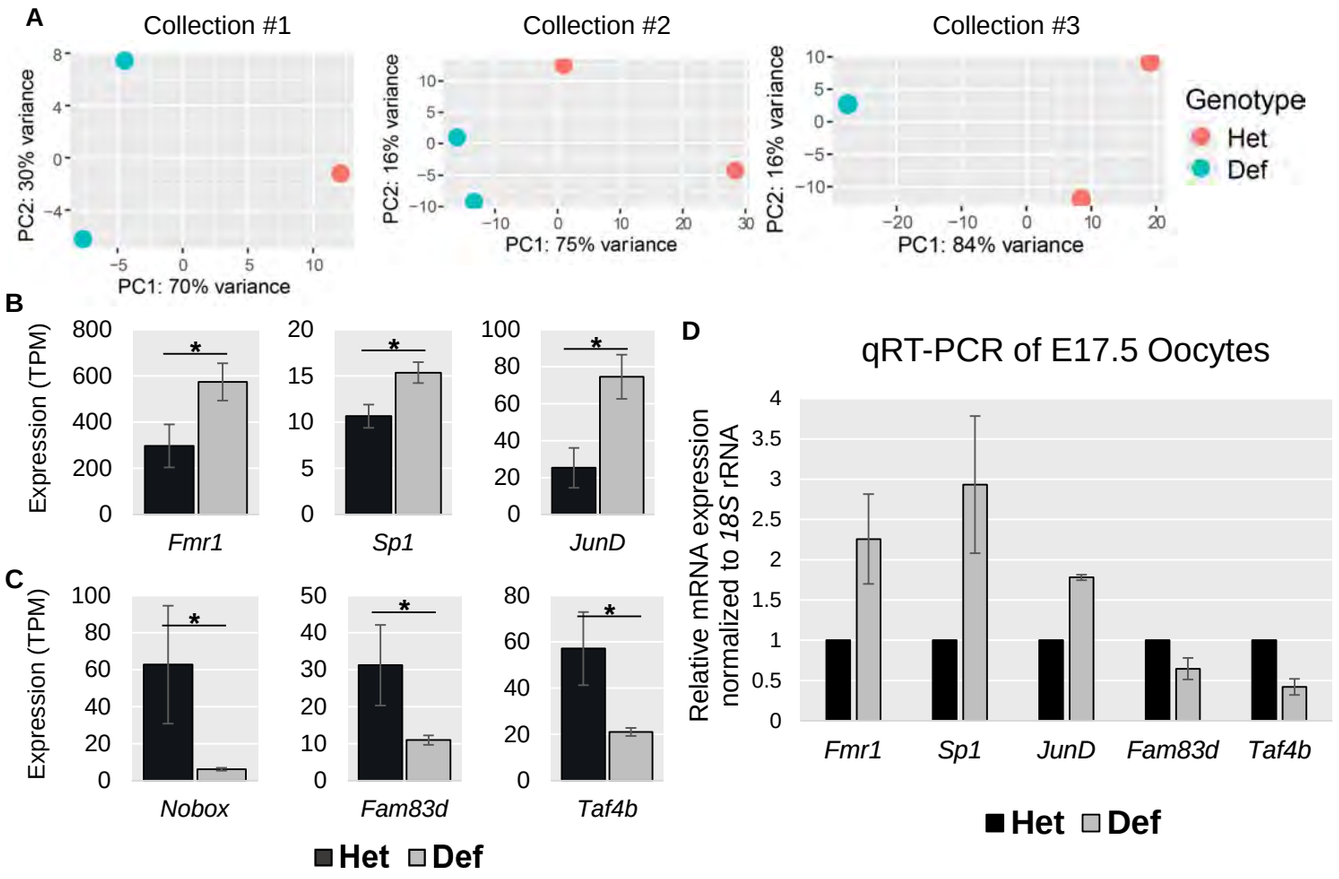
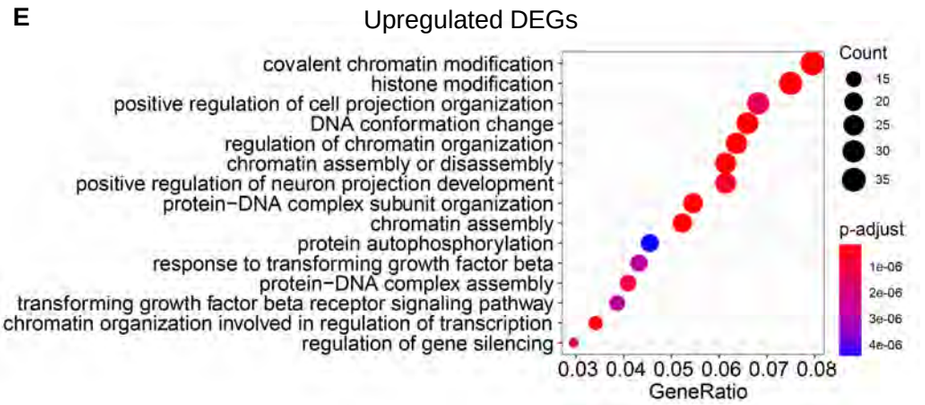


Fig. S1. Analysis of scRNA-seq dataset of E11.5 to E14.5 female germ cells. (A) Uniform Manifold Approximation and Projection (UMAP) of oocytes colored by embryonic time point. (B) UMAP of oocytes colored by pseudotime analysis. (C) Expression of *Pou5f1*, *Stra8*, *Taf4b*, and *Taf4a* plotted in terms of pseudotime and colored based on embryonic time point. (D) Table of top 10 genes (in addition to *Taf4b*) that are expressed significantly higher in *Taf4b*-expressing oocytes. Colors indicate association with gene ontology (GO) term synaptonemal complex assembly (yellow) and meiosis I/meiotic cell cycle (red).





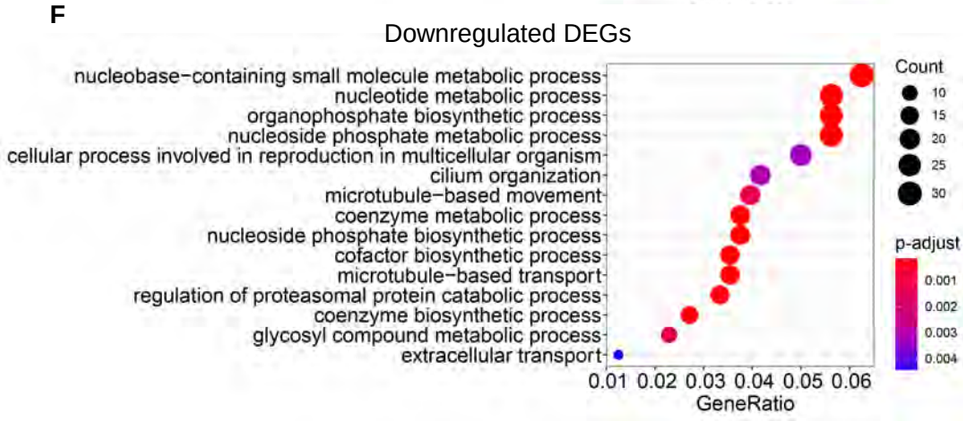


Fig. S2. E16.5 RNA-seq details. (A) PCA plots of different E16.5 RNA-seq collections colored by genotype. (B) Expression levels of upregulated DEGs in TPMs. (C) Expression levels of downregulated DEGs. (D) qRT-PCR results of E17.5 Oct4-EGFP^{+} oocytes for *Fmr1*, *Sp1*, *JunD*, *Fam83d*, and *Taf4b* from *Taf4b*-heterozygous (n=1) and *Taf4b*-deficient (n=2) samples. (E) Biological process GO analysis dotplot of DEGs that were increased in *Taf4b*-deficient oocytes ("Upregulated"). (F) Biological process GO analysis dotplot of DEGs that were decreased in *Taf4b*-deficient oocytes ("Downregulated").

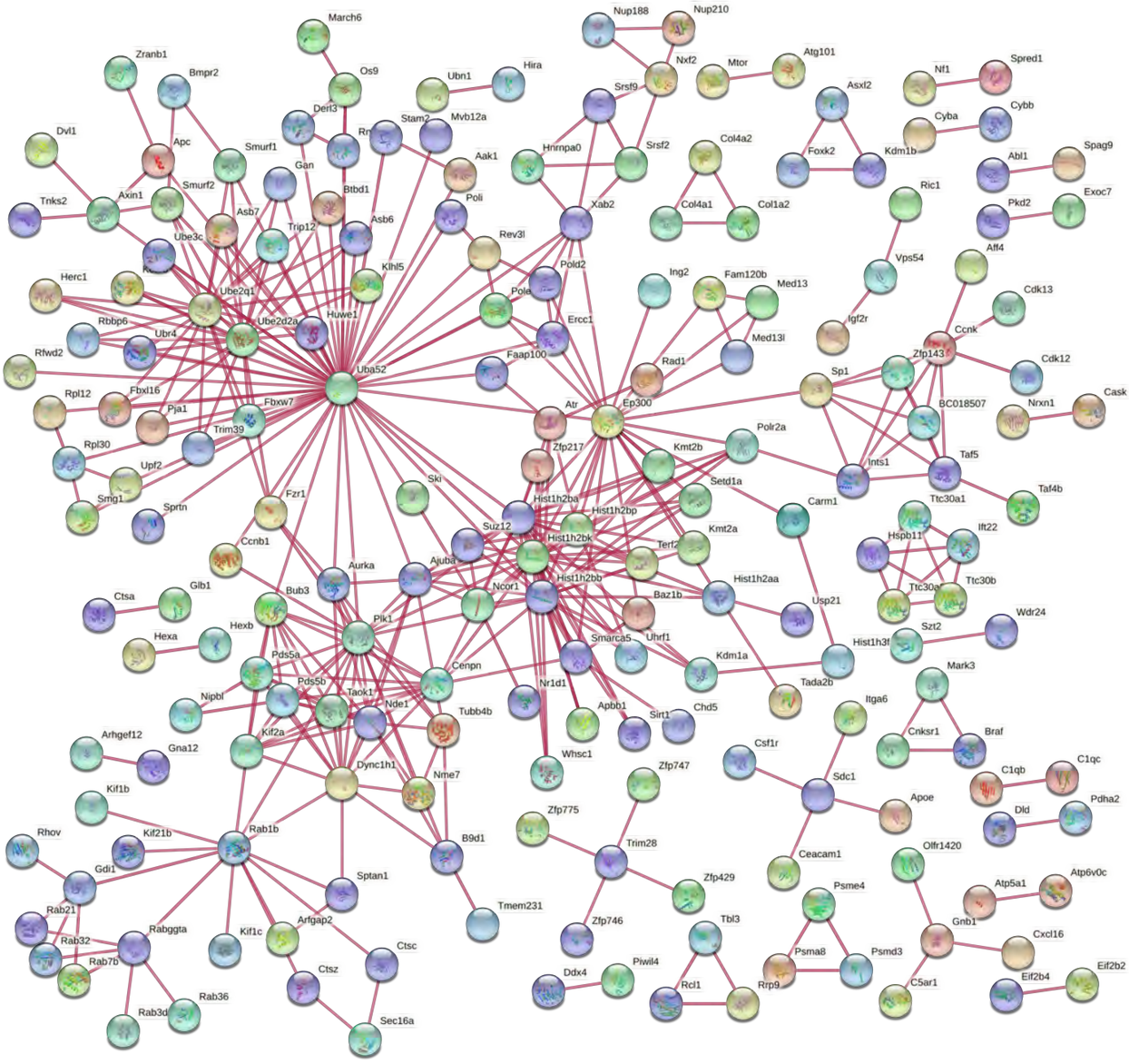


Fig. S3. Protein-protein interactions (PPIs) of E16.5 DEGs. Plot generated by STRING of DEGs from E16.5 RNA-seq that had at least one PPI (physical network, highest confidence) (Szklarczyk et al. 2019). There was a significant enrichment of PPIs ($p < 0.01$).

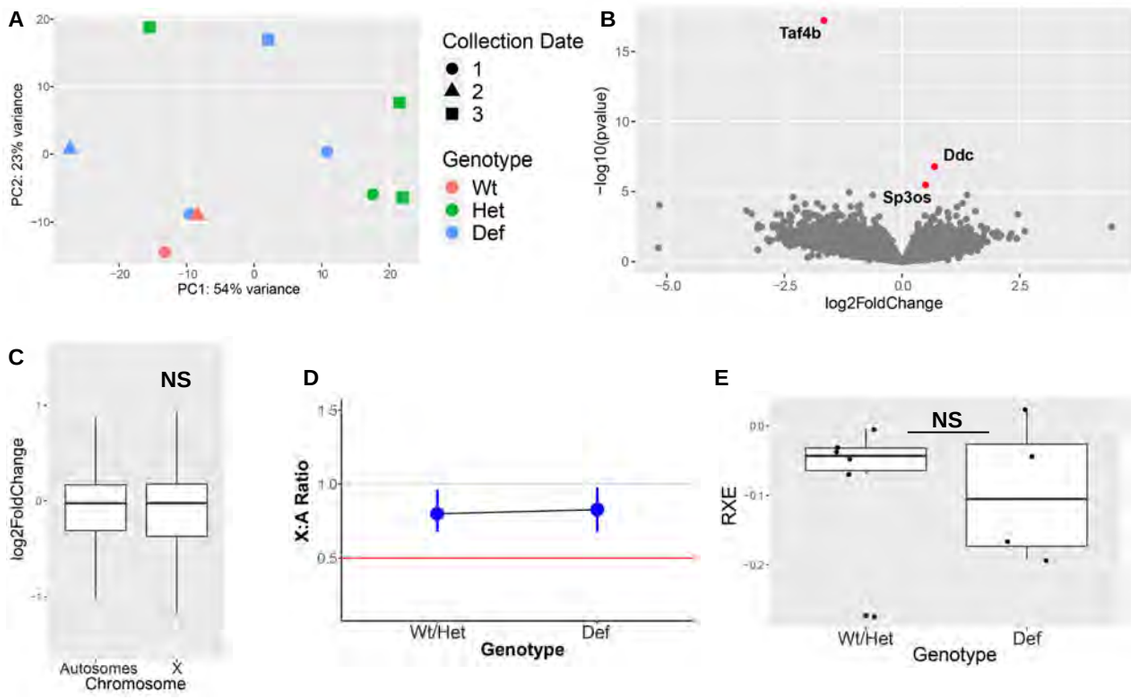


Fig. S4. E14.5 oocyte RNA-seq experiment. (A) PCA of E14.5 RNA-seq with genotype and collection date plotted. (B) Volcano plot of E14.5 RNA-seq with DEGs labeled (red). (C) Boxplots of $\log_2$ fold change values from DESeq2 of all genes comparing autosomal versus X chromosome $\log_2$ fold change (outliers removed), NS = not significant. (D) X:A ratio plot calculated through pairwiseCI after filtering for average TPM > 1 comparing Wt/Het X:A ratio to *Taf4b*-deficient X:A Ratio. (E) Boxplots of relative X expression (RXE) calculations after filtering for average TPM > 1 and adding pseudocounts for log-transformation for each *Taf4b*-Wt, -heterozygous, and -deficient sample. Wt/Het samples compared to Defs, NS = not significant.

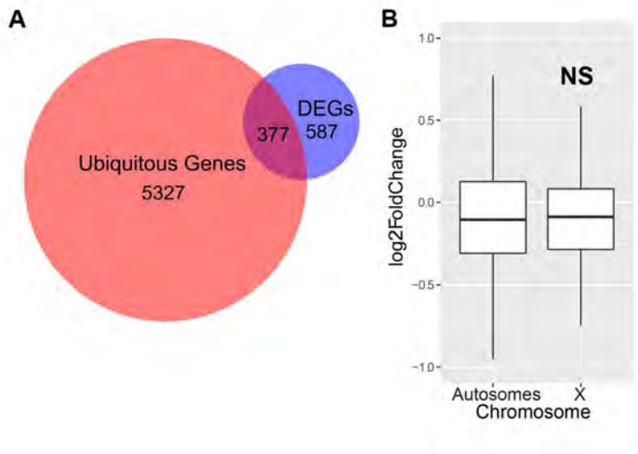
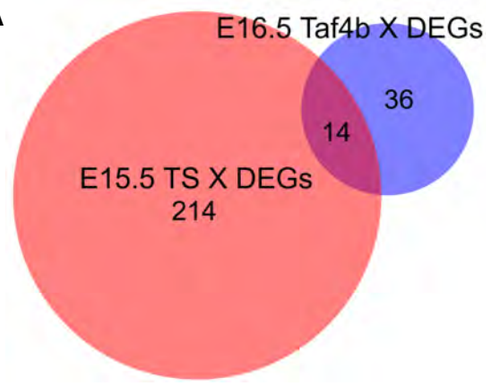


Fig. S5. Ubiquitous gene expression in E16.5 oocytes. (A) Venn diagram of E16.5 DEGs and "Ubiquitous Genes" as identified in Sangrithi et al. 2017. Significant overlap in Venn diagram ($p < 0.0001$, hypergeometric test). (B) Boxplots of $\log_2$ fold change values from DESeq2 of all ubiquitous genes comparing autosomal $\log_2$ fold change versus X chromosome $\log_2$ fold change, NS = not significant (outliers removed).

A



B

C

Gene Name	Ensembl Gene ID
4930550L24Rik (Magea13)	ENSMUSG00000046180
Atg4a	ENSMUSG00000079418
Ccdc160	ENSMUSG00000073207
Clcn5	ENSMUSG00000004317
Fmr1	ENSMUSG00000000838
Gdi1	ENSMUSG00000015291
Huwe1	ENSMUSG00000025261
Magea10	ENSMUSG00000043453
Mageb4	ENSMUSG00000035427
Nsdhl	ENSMUSG00000031349
Nxf2	ENSMUSG00000009941
Rps4x	ENSMUSG00000031320
Shroom4	ENSMUSG00000068270
Slc35a2	ENSMUSG00000031156

D

Gene Name	Ensembl Gene ID
1700080O16Rik (Magea14)	ENSMUSG00000031118
Atg4a	ENSMUSG00000079418
AV320801	ENSMUSG00000054994
Bcorl1	ENSMUSG00000036959
Ccdc160	ENSMUSG00000073207
Clcn5	ENSMUSG00000004317
Cybb	ENSMUSG00000015340
Fgf13	ENSMUSG00000031137
Fmr1	ENSMUSG00000000838
Gab3	ENSMUSG00000032750
Gdi1	ENSMUSG00000015291
Gla	ENSMUSG00000031266
Gm15023	ENSMUSG00000079432
Gm364 (Tm9sf5)	ENSMUSG00000079584
Gm5128	ENSMUSG00000094004
Gm7173 (Cfap47)	ENSMUSG00000073077
Huwe1	ENSMUSG00000025261
Mageb18	ENSMUSG00000035427
Mageb4	ENSMUSG00000035427
Mbnl3	ENSMUSG00000036109
Nkrf	ENSMUSG00000044149
Nsdhl	ENSMUSG00000031349
Nup62cl	ENSMUSG00000072944
Nxf2	ENSMUSG00000009941
Pja1	ENSMUSG00000034403
Rhox2d	ENSMUSG00000095698
Rps4x	ENSMUSG00000031320
Shroom4	ENSMUSG00000068270
Slc35a2	ENSMUSG00000031156
Tktl1	ENSMUSG00000031397
Tsga8	ENSMUSG00000035522

Fig. S6. Shared X chromosome DEGs between TS dataset and *Taf4b*-deficiency. (A) Venn diagram of all E16.5 *Taf4b* X chromosome DEGs compared with E15.5 TS X chromosome DEGs (protein-coding, p-adj < 0.05, avg TPM > 1). No significant overlap in Venn diagram (hypergeometric test). (B) Venn diagram of all E16.5 *Taf4b* X chromosome DEGs compared with E18.5 TS X chromosome DEGs (protein-coding, p-adj < 0.05, avg TPM > 1). Significant overlap in Venn diagram ($p < 0.0001$, hypergeometric test). (C) List of the 14 DEGs shared in (A). (D) List of the 31 DEGs shared in (B).

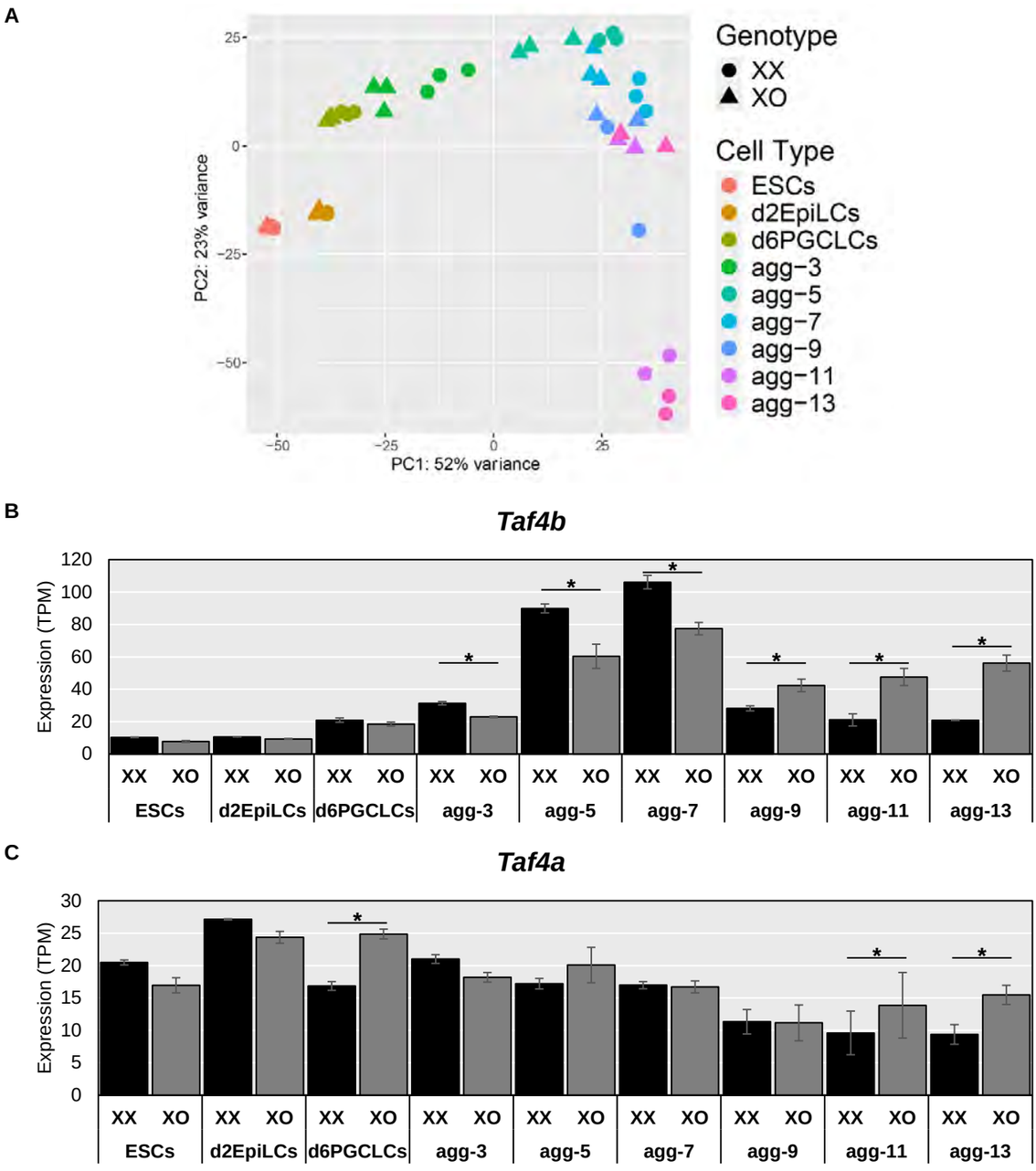


Fig. S7. Effects of XO on *Taf4b* in *in vitro* germ cell differentiation culture system. (A) PCA plot of cultured cells from Hamada et al., 2020, labeled based on cell type and genotype. (B) Expression levels of *Taf4b* in XX versus XO cultured cells (* = $p < 0.05$, protein-coding, avg TPM >1). (C) Expression levels of Taf4a in XX versus XO cultured cells. Error bars indicate ± standard error of the mean (SEM).

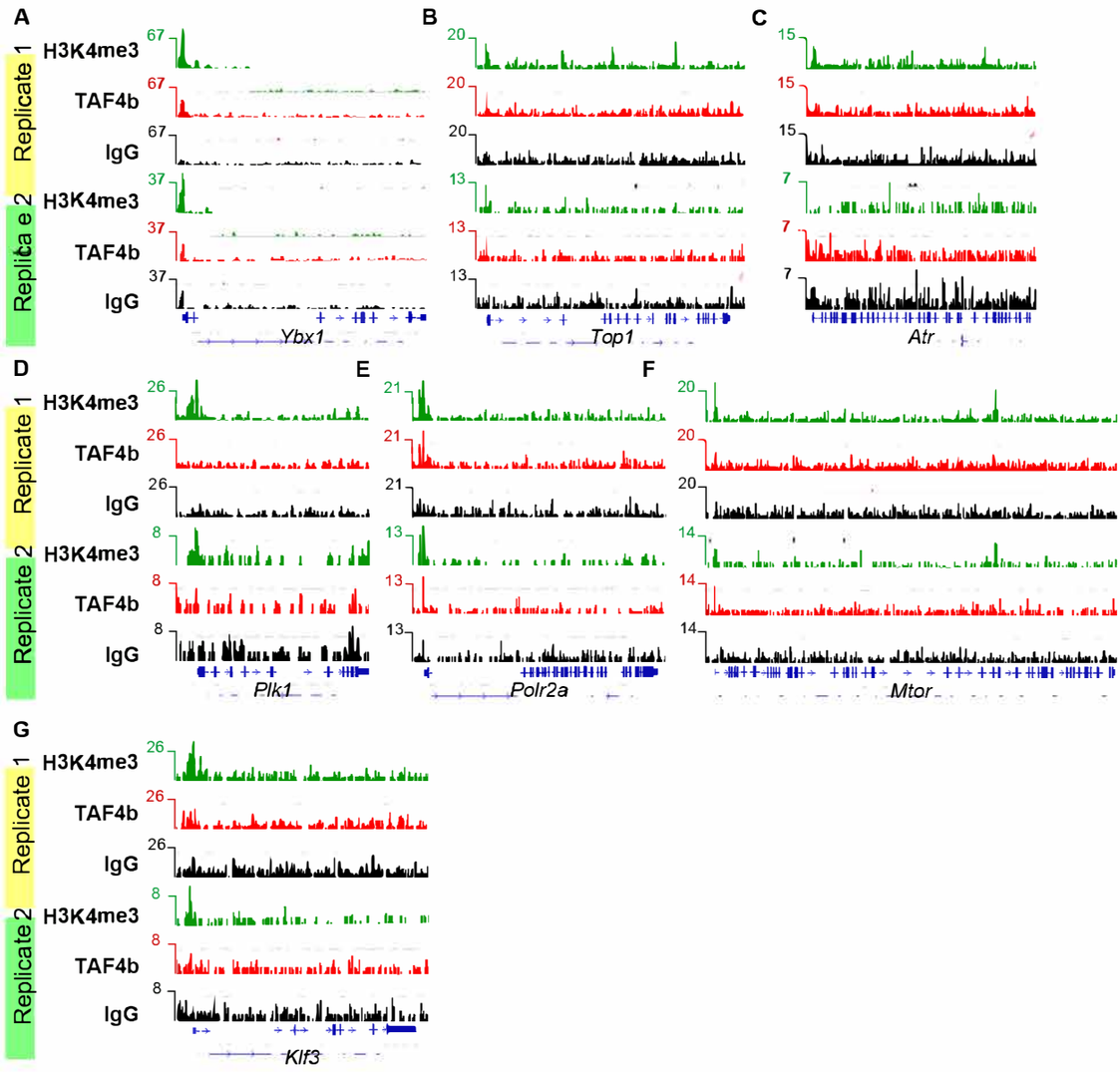


Fig. S8. Selected gene tracks. Genes tracks of *Ybx1* (A) and *Top1* (B), which were TAF4b CUT&RUN "promoter-TSS" peaks shared between the replicates but not DEGs. Gene tracks of *Atr* (C) and *Plk1* (D), which were DEGs but had no TAF4b peaks called. (E) Gene track for *Polr2a*, a DEG that had a TAF4b peak called in only Replicate 1. (F) Gene track for *Mtor*, a DEG that had a TAF4b peak called in only Replicate 2. (G) Gene track for *Klf3*, a non-DEG that had no TAF4b peaks called but did have H3K4me3 peaks called for replicates.

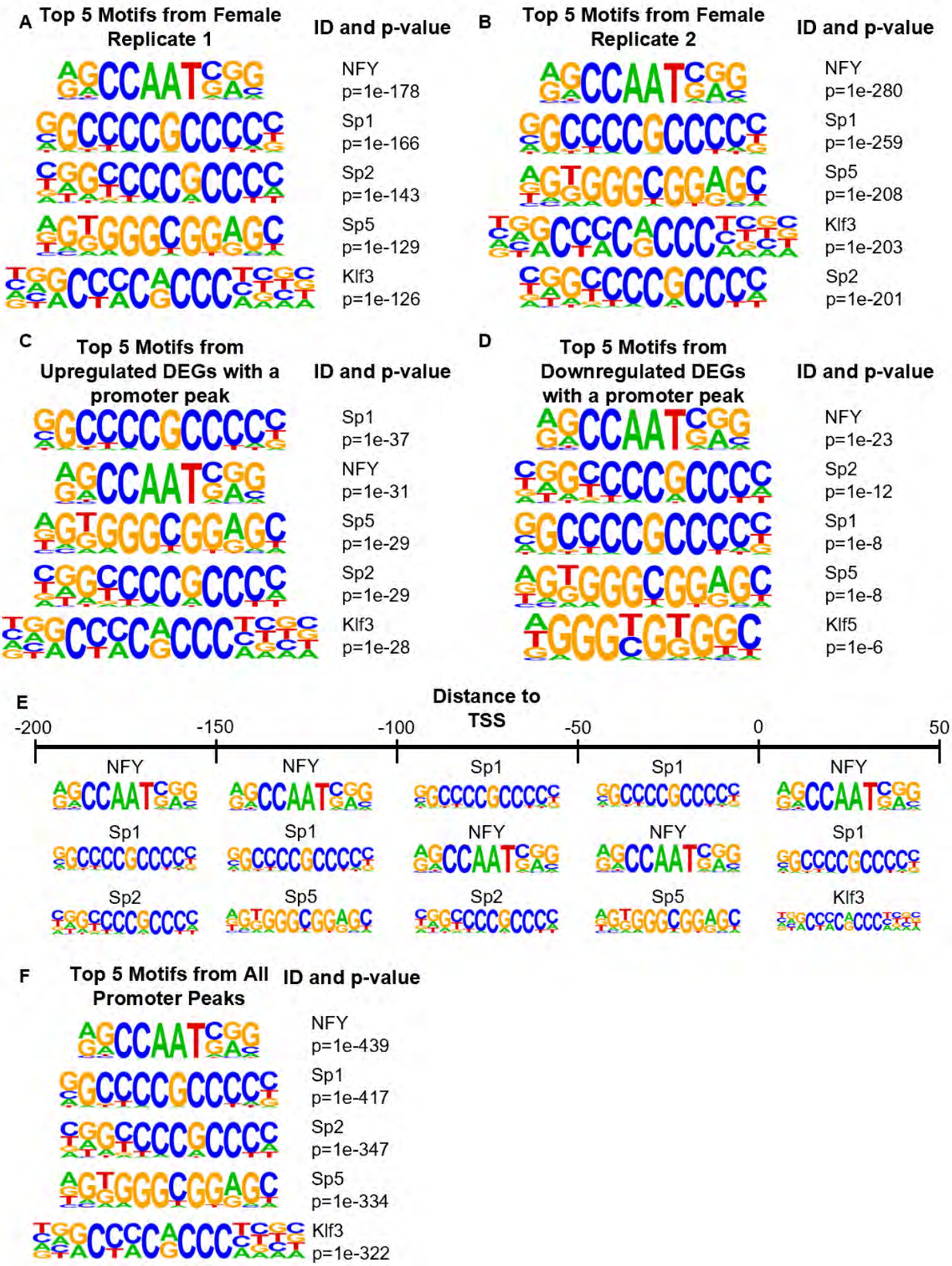


Fig. S9. Strong motif consistency when examining subsets of E16.5 oocyte CUT&RUN data. (A) Top five TAF4b motifs from "promoter-TSS" peaks in female Replicate 1. (B) Top five TAF4b motifs from "promoter-TSS" peaks in female Replicate 2. (C) Top five TAF4b motifs from Upregulated DEGs that had at least one TAF4b "promoter-TSS" peak. (D) Top five TAF4b motifs from Downregulated DEGs that had at least one TAF4b "promoter-TSS" peak. (E) Diagram of the top three TAF4b "promoter-TSS" motifs in 50 bp windows relative to the TSS. (F) Top five motifs enriched at all TAF4b "promoter-TSS" peaks, the promoter ID, and the associated p-value.

Table S1. Output for scRNA-seq analysis.

click here to download Table S1

Table S2. Cell numbers for RNA-seq samples.

Age	Genotype	Sample #	Cell #
E14.5	Wildtype	1	7,942
		2	18,256
	Heterozygous	1	12,553
		2	12,897
		3	2,822
		4	19,308
	Deficient	1	9,112
		2	5,233
		3	9,893
		4	7,035
E16.5	Heterozygous	1	7,199
		2	2,512
		3	3,399
		4	9,369
		5	14,402
	Deficient	1	2,181
		2	3,547
		3	19,089
		4	5,688
		5	9,076

Table S3. Output of E16.5 oocyte RNA-seq. .

Click here to download Table S3

Table S4. Output of E14.5 oocyte RNA-seq.

Click here to download Table S4

Table S5. Chromosome distribution of Downregulated DEGs.

Chromosome	Total Genes	Observed	Expected	Chi^2	p-value
1	1439	32	32.41	0.006	0.9406
2	1630	35	36.71	0.086	0.7694
3	1178	21	26.53	1.217	0.2699
4	1290	28	29.05	0.04	0.8409
5	**1499**	**20**	**33.76**	**6.014**	**0.0142**
6	1145	17	25.79	3.158	0.0755
7	1750	33	39.42	1.135	0.2867
8	1116	25	25.14	0.001	0.9771
9	1281	34	28.85	0.975	0.3233
10	1039	27	23.40	0.581	0.4459
11	1592	38	35.86	0.138	0.7107
12	901	21	20.29	0.026	0.8722
13	940	26	21.17	1.156	0.2823
14	764	16	17.21	0.088	0.7666
15	825	23	18.58	1.092	0.296
16	655	11	14.75	0.982	0.3216
17	1056	21	23.78	0.341	0.5591
18	571	15	12.86	0.365	0.5455
19	673	17	15.16	0.23	0.6313
X	**900**	**41**	**20.27**	**22.094**	**<0.0001**

Table S6. Chromosome distribution of Upregulated DEGs.

Chromosome	Total Genes	Observed	Expected	Chi^2	p-value
1	1439	24	29.95	1.264	0.2609
2	1630	40	33.93	1.172	0.279
3	**1178**	**14**	**24.52**	**4.766**	**0.029**
4	1290	30	26.85	0.392	0.5311
5	1499	38	31.20	1.589	0.2074
6	1145	18	23.83	1.504	0.2201
7	1750	40	36.43	0.381	0.5372
8	1116	21	23.23	0.225	0.6351
9	1281	28	26.66	0.071	0.7898
10	1039	18	21.63	0.638	0.4245
11	**1592**	**48**	**33.14**	**7.181**	**0.0074**
12	901	24	18.75	1.529	0.2162
13	**940**	**31**	**19.57**	**6.977**	**0.0083**
14	764	10	15.90	2.269	0.132
15	825	14	17.17	0.609	0.4353
16	655	9	13.63	1.623	0.2027
17	1056	19	21.98	0.424	0.5148
18	571	12	11.89	0.001	0.9731
19	673	16	14.01	0.292	0.5889
X	**900**	**9**	**18.73**	**5.268**	**0.0217**

Table S7. Ubiquitous gene exploration.

Click here to download Table S7

Table S8. Mouse Turner Syndrome dataset.

Click here to download Table S8

Table S9. I*n vitro* differentiation mouse TS dataset.

Click here to download Table S9

Table S10. TAF4b and H3K4me3 CUT&RUN in E16.5 oocytes.

Click here to download Table S10

Table S11. Quantitative real-time PCR primers

Gene	Primer Sequence
Fmr1	Forward - CAATGGCGCTTTCTACAAGGC
	Reverse - TCTGGTTGCCAGTTGTTTCA
Sp1	Forward – GCCACCATGAGCGACCAAG
	Reverse – GAGTCTGAGAAAAGGCGGCA
Fam83d	Forward - CGTGTCGAGGCTCATTTCC
	Reverse - CCACAGCAATCACCTCTCGG
JunD	Forward - CCCCGGACTCTTTCGAGACT
	Reverse - CCTTAGAGCCCCTACTCGGA
Taf4b	Forward - GATGTTACTAAAGGCAGCCAAGAGT
	Reverse – CTGCTCTGGATCTTCTTTATTGGAG
18S rRNA	Forward – GTAACCCGTTGAACCCCATT
	Reverse – CCATCCAATCGGTAGTAGCG