Efficient Human Germ Cell Specification from Stem Cells via Combinatorial Expression of Transcription Factors

Christian Kramme,[1,2,*] Merrick Pierson Smela,[1,2,*] Bennett Wolf,[1,2] Patrick R. Fortuna,[1,2] Garyk Brixi,[1,2,3] Kalyan Palepu,[1,2,3] Edward Dong,[1,2] Jessica Adams,[1,2] Suhaas Bhat,[1,2,3] Sabrina Koseki,[4,5] Emma Tysinger,[4,5] Teodora Stan,[4,5] Richie E. Kohman,[1,2] Songlei Liu,[1,2] Mutsumi Kobayashi,[6] Toshi Shioda,[6] George M. Church,[1,2] Pranam Chatterjee[1,2,3,4,5,†]

1. Wyss Institute, Harvard Medical School
2. Department of Genetics, Harvard Medical School
3. Department of Biomedical Engineering, Duke University
4. Center for Bits and Atoms, Massachusetts Institute of Technology
5. Media Lab, Massachusetts Institute of Technology
6. Massachusetts General Hospital Center for Cancer Research, Harvard Medical School

*These authors contributed equally
[†]Corresponding author: pranam.chatterjee@duke.edu

Abstract

Germ cells are the vehicle of human reproduction, arising early in embryonic development and developing throughout adult life until menopause onset in women. Primordial germ cells are the common precursors of germline cells in both sexes, undergoing sexual specification into oogonia or gonocytes which further develop into oocytes or spermatocytes during development. Methods for recapitulation of primordial germ cell and oogonia formation have been developed extensively in recent decades, but fundamental technical limitations in their methodologies, throughput, and yield limit their utilization. Recently, transcription factor (TF)-based methods for human primordial germ cell-like cell (hPGCLC) formation, mouse meiotic entry, and mouse oocyte maturation have demonstrated the feasibility of gene overexpression screening in identifying potent regulators of germ cell development. Here we screened 47 folliculogenesis-regulating TFs for their role in hPGCLC and oogonia formation, identifying *DLX5*, *HHEX*, and *FIGLA* whose individual overexpression enhances hPGCLC formation from hiPSCs. Additionally, we identify a set of three TFs, *ZNF281*, *LHX8*, and *SOHLH1*, whose combinatorial overexpression drives direct oogonia-like formation from hiPSCs in a four-day, feeder-free monolayer culture condition with additional feeder-free culture capabilities post-isolation. We characterize these TF-based germ cells via gene and protein expression analyses, and demonstrate their broad similarity to *in vivo* germ cells. Together, these results identify novel regulators of human germ cell development and establish new TF-based tools for human *in vitro* oogenesis research.

Main

It is estimated that one in ten people in the US struggle with infertility.[1,2] Access to human female germline cell types for study is limited, and little is known about the underlying genetic regulation of human gametogenesis.[3] *In vitro* gametogenesis (IVG) holds promise for the complete modeling of human gametogenesis from stem cells and was recently successfully achieved in mice.[4,5] However, challenges remain in translation of these methods to humans due to species-specific differences, extensive culture periods, low yielding differentiation, and lack of screening tools.[6,7] In this study, we aim to address these current challenges in human IVG to enable derivation of human germ cell types from female human induced pluripotent stem cells (hiPSCs) via transcription factor (TF) overexpression. To do this, we computationally identify 47 TFs and screen for their roles in primordial germ cell (hPGCLCs) and oogonia/oocyte-like cell formation.

We first identify three TFs, *DLX5*, *HHEX*, and *FIGLA,* whose individual overexpression drives potent enhancement of hPGCLC formation in monolayer and aggregate co-culture. We demonstrate that DLX5 overexpression rescues loss of BMP4 during germ cell formation. We additionally identify *LHX8*, *SOHLH1*, and *ZNF281*, whose combinatorial overexpression generates DDX4+ induced oogonia/oocyte-like cells (iOLC) in four days in monolayer culture. We further generate additional TF combinations that drive increased iOLC yield, and subsequently develop methods for feeder-free iOLC expansion. Finally, we elucidate the TF binding activity of all six TFs and show they regulate the core gametogenesis gene regulatory network (GRN). Together, these results establish a foundational platform for targeted derivation and genetic screening of human germ cell types from pluripotent stem cells via TF-directed differentiation. Using these tools, we uncover TFs that play central roles in human gametogenesis and expand the basic knowledge of how diverse TFs regulate and direct human germ cell development.

Results

In silico identification of transcription factors that are central regulators of human folliculogenesis

Based on previous studies, it is known that there are key TFs that signal germline cells to enter meiosis in response to retinoic acid signaling.[8,9] To identify candidate TF regulators driving oocyte maturation events, we curated publicly-available single cell RNA sequencing (scRNA-seq) data of natural folliculogenesis and normalized the gene expression of germ cell transcriptomes at various stages of follicle development: primordial germ cells, intermediate oogonial cells and oocytes within primary, secondary, antral, preovulatory follicles, and MII oocytes.[7,10–12] After constructing this pseudo-time series transcriptomic matrix, we employed unsupervised, regression-based gene regulatory network (GRN) inference via regularized stochastic gradient boosting to generate a representative, directed graph of oogenesis, where nodes represent genes and edge weights correlate to the regulatory effect between TFs and their downstream targets.[13] After conducting standard pruning of the graph for low-weight, spurious edges, we applied the adapted PageRank algorithm from our recently-published STAMPScreen pipeline to prioritize the most critical, “central” TFs within the GRN (Figure 1A).[14] Our ranked list includes both previously-identified regulators of mammalian oocyte transcriptional networks and germ cell development, such as NOBOX, LHX8, and OTX2, as well as potential novel regulators, including ZNF281, SOX13, and DLX5.[9,15] Overall, these results motivated us to combinatorially screen prioritized candidates in hiPSCs to resolve the human oocyte transcriptional network experimentally.

Characterization of the contribution of 47 TFs to germ cell and oogonia formation via overexpression screening

We generated doxycycline-inducible vectors expressing a full length cDNA for each of the 47 TFs identified from our TF prediction algorithm.[16] Each vector harbored a 20 bp barcode on the 3’ UTR of the cDNA and was piggyBac integratable.[17] We constructed a NANOS3-mVenus; DDX4-tdTomato dual reporter hiPSC line (N3VD4T) using CRISPR-Cas9-mediated homology directed repair (HDR), and generated 47 hiPSC lines

harboring integrations of each TF individually through super piggyBac transposase-mediated insertion.[7,18,19] Polyclonal pools for each TF were utilized for screening purposes.

hPGCLCs were induced through epiblast-like intermediates followed by BMP4 induction for 4 days in a monolayer condition adapted from recently described methods (Figure 1B).[20] We further optimized this monolayer induction protocol, finding that elimination of vitamin A from the media and increasing Activin A concentration increased hPGCLC yield while epiblast-like induction length showed marginal improvement in yield (Figure S1). Utilizing this four day monolayer hPGCLC induction protocol, we assessed NANOS3+ hPGCLC yield via flow cytometry in the presence or absence of doxycycline for all 47 TFs in triplicate (Figure 1C). We show that all 47 TFs drive upregulation of NANOS3+ hPGCLC yield, thus highlighting the general utility of our TF prediction algorithm for identifying TF regulators of human germline development. Remarkably, 3 TFs (DLX5, HHEX, and FIGLA) induced a NANOS3+ yield higher than that of three known TF regulators of hPGCLC development: SOX17, TFAP2C, and PRDM1.[6,21,22]

We sought to further elucidate the contribution of DLX5, HHEX, and FIGLA to hPGCLC formation. We find that overexpression of DLX5 alone alone is able to replace exogenous BMP4 in the induction of hPGCLCs, driving potent hPGCLC formation in the absence of BMP4 (Figure S2). We further show that hPGCLC yield increases with TF dosage for all three TFs (Figure S3A). Additionally, overexpression of each of these three TFs individually increased hPGCLC formation in both floating aggregate and monolayer cultures, as quantified by both the NANOS3 reporter and CD38 cell surface marker expression (Figure S3B and S3C). TF overexpression additionally stabilized hPGCLC yield at day 6 of monolayer induction, whereas the no--TF control showed a dramatic decrease in hPGCLC yield at day 6 (Figure S3C). Interestingly, combinatorial overexpression of the three TFs exhibited lower hPGCLC yield compared to individual overexpression (Figure 1D). We additionally observed that homozygous CRISPR KO of all three TFs individually did not reduce NANOS3+ hPGCLC yield, suggesting that these TFs are likely not required for NANOS3+ germ cell formation (Figure S4). Together, these results identify DLX5, HHEX, and FIGLA as enhancers of hPGCLC formation.

We next assessed production of production of the DDX4+ oogonia-like cells from hiPSCs cells from hiPSCs via flow cytometry in the presence or absence of doxycycline for all 47 TFs in triplicate (Figure 1E). Compared to control, the overall percentage of DDX4+ cells was not greatly enriched by any single TF. However, a small percentage of cells with elevated DDX4+ expression was identified in the ZNF281, LHX8, and SOHLH1 induction conditions (Figure 1E). We hypothesized that combinatorial expression of these factors may increase DDX4+ yield. To investigate this hypothesis, we overexpressed all combinations of two and all three TFs, and found that overexpression of all three, which we term DDX4 by three TFs or D3 combo, induced a larger percentage of DDX4+ cells (Figure 1F). Remarkably, this high DDX4+ population is obtained in just 4 days in monolayer through direct TF induction during hPGCLC formation. NPM2 is a critical oocyte marker gene, involved in chromatin organization.[9] Employing a DDX4-tdTomato; NPM2-mGreenLantern reporter hiPSC line (D4TP2G), we tested the addition of other TFs and RNA-binding proteins, including DLX5, HHEX, FIGLA, DAZL, DDX4, and BOLL, to the D3 combo to determine if the DDX4+ yield could be increased (Figure 1F). Our results demonstrate that addition of FIGLA, in particular, drives an increase in DDX4+ yield and NPM2+ yield. We also find that addition of all 9 factors, a condition we call DNR3 combo, induces robust DDX4+ yield and a modest NPM2+ yield. To understand the role of these six TFs in fetal gonadogenesis, we assessed their expression using publicly available scRNA-seq data.[11] Our investigation uncovered broad expression of all six TFs through fetal gonad development, with strong upregulation during oogenesis onset, thus supporting their functionality in our system (Figure S5). We therefore conclude that overexpression of ZNF281, SOHLH1, and LHX8 individually or more optimally in combination via a hPGCLC differentiation protocol is a salient platform for DDX4+ oogonia-like cell formation with addition of FIGLA, HHEX, DLX5, DAZL, BOLL, and DDX4 helping to drive higher yields.

TF overexpression drives formation of cells with transcriptomic and protein expression profiles of primordial germ cells and oogonia/oocyte-like cells

hPGCLCs are delineated by upregulation of key genes such as *NANOS3, CD38, SOX17, TFAP2C*, and *PRDM1* and suppression of *SOX2*.[12,19,23] Oogonia-like cells are delineated by expression of *DDX4, DAZL*, as well as other genes such as *ID1/2/3/4, MSX1/2*.[7] We sought to characterize whether our TF-driven NANOS3+ hPGCLCs were bona fide hPGCLCs and if our DDX4+ D3 and DNR3 cells resembled oogonia cells through isolation and bulk RNA sequencing (RNA-seq) (Figure 2A). NANOS3+ hPGCLCs derived from DLX5, HHEX, and FIGLA overexpression exhibit the characteristic upregulation of *SOX17, PRDM1, TFAP2C, NANOS3, CD38* and downregulation of *SOX2*, validating that our TFs drive bona fide germ cell formation. Additionally, our D3 and DNR3 populations express *DDX4, ID1/2/3/4, TBX3, GDF9, HOXB6, MSX1/2,* and *PBX1* while *DAZL* was not expressed (except for exogenous *DAZL)* in DNR3. We furthermore performed immunofluorescence staining for OCT4, SOX17, and ITGA6 and demonstrate that our TF-driven hPGCLCs exhibit nominal protein expression hallmarks of conventional hPGCLCs.[24] Our D3-FIGLA DDX4+ cells exhibit DDX4 protein localization in punctate patterns in the cytoplasm, with no expression of DAZL noted. Taken together, we conclude that overexpression of DLX5, HHEX, or FIGLA during germ cell formation drives characteristic hPGCLC gene and protein expression, while combinatorial overexpression of the D3, D3-FIGLA, and DNR3 combinations drives oogonia-like gene and protein expression.

We additionally performed epigenetic profiling of our hPGCLCs and oogonia-like cells, using enzymatic methylation sequencing (EM-Seq) to profile global levels of 5-methylcytosine (5mC) in our D3 and DNR3 DDX4+ cells compared to hPGCLC and hiPSC controls (Figure S6A-B).[25] We find that our DDX4+ cells have a 5mC pattern more closely resembling hiPSCs and fail to fully demethylate their genome as would be expected for oogonia.[7,9] We additionally utilized time series CUT&RUN to assess TF binding of all six TFs at day 1, 2, and 4 of differentiation, finding broad binding to the core hPGCLC network, BMP4, and oogonia-related genes (Figure S6C-J).[26] Taken together, our TFs demonstrate broad binding to core regulators of germ cell development, yet fail to fully demethylate the genome in the D3 and DNR3 combinations.

We further compared the transcriptomes of our cells to that of previously characterized germ cells in the Gene Expression Omnibus (GEO) database using the recently-described transcriptome overlap metric (TROM) to determine the similarity between our NANOS3+ and DDX4+ cells to other *in vitro* and *in vivo* germline cell types (Figure 2C).[7,10,12,27,28] After batch normalization and TROM analysis, we show that our FIGLA, HHEX, and DLX5-overexpressed NANOS3+ cells indeed resemble primordial germ cells, as expected. Furthermore, our D3 combo DDX4+ cells closely resemble week 26 fetal ovarian oogonia cells, while the DNR3 DDX4+ cells resemble primordial, secondary and (most notably) antral *in vivo* oocytes. Based on these findings we designate the DDX4+ D3 and DNR3 cells as induced oogonia/oocyte-like cells (iOLCs).

As a final characterization, we conducted scRNA-seq on unsorted, day 4 differentiation populations to analyze the heterogeneity of our hPGCLC and iOLC samples. We first merged fetal ovarian and testis germ cell data from previous studies with a pre-curated cell matrix of various somatic cell types to create a reference single cell atlas.[11] Each sample was then mapped onto this reference atlas via the recently-described Symphony algorithm to assign a putative cell identity to each cell in the sample.[29] Our final clustering results demonstrate that our TF-based inductions drive fetal ovarian-like cell formation, as well as an array of off-target cell types including neural cell types, averaging around 7 to 10% ovarian cell types in each population (Figure 2D).

iOLCs expand feeder-free in culture

It is known that *in vivo*, primordial germ cells and oogonia mature and expand, establishing the human ovarian reserve.[7,30] Recent methods have been developed for expansion of hPGCLCs *in vitro* post-isolation.[12,31,32] No

method yet exists for the expansion of human oogonia or oogonia-like cells *in vitro*. We thus sought to determine if our iOLCs could be maintained in culture post-isolation. Serendipitously, we find that by seeding isolated iOLCs onto matrigel-coated plates feeder-free in a media composition recently described for hPGCLC expansion, we are able to maintain and expand the iOLC population (Figure S7A).[12] iOLCs are maintained in culture for over 30 days, retaining heterogeneous expression of DDX4. These DDX4+ iOLCs are larger than DDX4- cells of hiPSCs, averaging 100 μm in width and growing in nest-like structures (Figure S7A and S7B). Expanding iOLCs with enlarged shape are NPM2- and TFAP2C- and show dense nuclear associated granules and cytoplasmic projections. We therefore conclude that our TF-derived iOLCs are capable of feeder-free maintenance and expansion post-isolation.

Discussion

In this study, we identified 47 TFs that may play highly connected roles within the underlying GRN of gametogenesis and further screened these TFs via combinatorial and individual overexpression in oogenesis-reporter hiPSC lines. We find that three TFs, DLX5, HHEX, and FIGLA, individually drive enhancement of hPGCLC yield when overexpressed during hPGCLC specification, while three other TFs, ZNF281, LHX8, and SOHLH1, drive DDX4+ oogonia/oocyte-like cell formation.

We demonstrate that DLX5, HHEX, and FIGLA drive on-target hPGCLC formation, shown in their expression of the core primordial germ cell genes and proteins. These TFs drive hPGCLC formation in a 2D monolayer culture format as well as in floating aggregate methods across different isolation markers, providing broad utility across protocols for enhancement of yield. We furthermore find that DLX5 overexpression rescues and continues to enhance hPGCLC yield in the absence of BMP4. From our study, it is unclear whether this rescue is the result of DLX5 upregulating endogenous BMP4 expression or whether this is through an independent pathway. We find that these three TFs are not required for NANOS3+ germ cell formation but that DLX5 and HHEX knockout reduces CD38+ hPGCLC yield; however, more surface markers and differentiation contexts are required for definitive determination of necessity.

Using CUT&RUN, we find the six characterized TFs bind broadly to the germ cell and greater oogenesis regulatory genes. It is unclear from our system whether this is the canonical function of the TFs normally during hPGCLC formation or a result of forced overexpression. For FIGLA, in particular, it is likely that our overexpression drives a non-canonical function that is not normally seen during hPGC formation, as FIGLA is not normally expressed at PGC developmental timepoints *in vivo*.[11] Nonetheless, this study motivates utilization of our toolkit in a diverse range of hPGCLC screening modalities, which may uncover the role of new TFs in regulating gametogenesis and further elucidate the underlying gene regulatory network of primordial germ cell specification.

We also show that overexpression, both individually and more usefully in combination, of SOHLH1, LHX8, and ZNF281 (D3) drives formation of DDX4+ oogonia-like cells in just four days. We identify higher order combinations, such as the addition of FIGLA, and a 9 gene combo termed DNR3, that generate induced oocyte-like cells (iOLCs) that are DDX4+ and NPM2+. Our iOLC formation is rapid compared to existing methods which generally require 70 to 120 days in xrOvary based methods.[7] Our iOLCs are further specified feeder-free in 2D monolayer culture using our hPGCLC formation protocol, making scaling for production and screening simple and amenable to integration with current hPGCLC screening designs. We also develop a method for iOLC expansion feeder-free post-isolation using media developed for hPGCLC maintenance. [12]

Via transcriptomic analysis, we find that our TFs drive upregulation of key oogonia and oocyte genes, and that iOLCs broadly share similarity with *in vivo* oocytes and fetal germ cells.[10,11] However, our iOLCs are still missing key gene markers such as DAZL and do not seem to have entered or surpassed meiosis. Future

studies will work to establish new methods for initiating meiotic entry in iOLCs and generate transcriptomes with additional similarity to putative oocytes. We further show that our iOLCs slightly demethylate their genomes but fail to fully demethylate as would be expected for oogonia.[7] This likely shows that there is aberrant epigenetic regulation of our iOLCs due to the rapid and direct manner of their specification, in which they may skip hPGCLC formation or simply do not traverse enough cell divisions to undergo demethylation. More research is needed to determine the functionality of these iOLCs and whether they can contribute to further gametogenesis and oocyte formation, such as by entering and performing meiosis and oocyte maturation. In conclusion, we uncover the role of novel TFs in driving primordial germ cell, oogonia and oocyte-like formation, establishing a novel, high-throughput platform for *in vitro* human gametogenesis and reproductive modeling in a rapid and efficient manner.

Methods

Ethics Statement

All experiments on the use of hiPSCs for the generation of hPGCLCs, oogonia-like cells, and oocyte-like cells were approved by the Embryonic Stem Cell Research Oversight (ESCRO) Committee at Harvard University.

Gene Regulatory Network Inference and TF Prioritization

RNA-seq datasets were obtained from the Gene Expression Omnibus (GEO) database, and log2fc values for each aligned gene for each sample were calculated using the DESeq2 package.[33] Gene regulatory networks were inferred utilizing the GRNBoost2 algorithm in the Arboreto computational framework.[13,33] PageRank was calculated for each transcription factor in the resulting network via the NetworkX package, and ranked factors were visualized using Seaborn. The full code and corresponding Jupyter notebooks for the standard STAMPScreen pipeline can be found at: https://github.com/programmablebio/stampscreen.

Cell lines used

For these studies we utilized the ATCC-BXS0116 female hiPSC line, which we term F3, as the parental line for most studies. Additional studies with biological replicates utilized the ATCC-BXS0115 female hiPSC line, which we term F2, as well as an in-house Epi5 episomal footprint-free reprogrammed hiPSC line, termed F66, derived from the NIA Aging Cell Repository (NIA) fibroblast line AG07141. F2, F3, and F66 were subjected to Thermo Fisher Cell ID + Karyostat as well as Pluritest. All three cell lines are karyotypically normal and scored as normal in the Pluritest compared to the assays' control dataset.

hiPSC Culturing

Unless otherwise specified, all hiPSCs were maintained feeder-free on hESC-qualified Matrigel coated plates (Corning), at manufacturer suggested dilution. hiPSCs were maintained in mTeSR media, with mTeSR Plus utilized for standard expansion, passaging and cell line creation and mTeSR1 utilized prior to induction where specified. hiPSCs were passaged in mTeSR1 for at least one passage prior to differentiation in order to remove the stabilized FGF present in mTeSR Plus. Cells were passaged 1:10 to 1:20 every 3-4 days using Accutase and seeded in the presence of 10 μM Y-27632, with media being changed every day for mTeSR1 or every other day when mTeSR plus was utilized. Cells were regularly tested for mycoplasma contamination.

Generation of gametogenesis cell reporter hiPSC lines

Homology arms for target genes (*DDX4*, *NPM2*, *NANOS3, TFAP2C*) were amplified by PCR from genomic DNA. For each gene, a targeting plasmid, containing an in-frame C-terminal T2A-fluorescent reporter reporter of either *tdTomato* (for *DDX4*), *mGreenLantern* (for *NPM2 and TFAP2C*), or *mVenus* (for *NANOS3*), as well as a Rox-PGK-PuroTK-Rox selection cassette, was constructed by Gibson assembly. The plasmid backbone additionally had an MC1-DTA marker to select against random integration. sgRNA oligos targeting the C-terminal region of target genes were cloned into the pX330 Cas9/sgRNA expression plasmid (Addgene 42230). For generation of the reporter lines, 2 µg donor plasmid and 1 µg Cas9/sgRNA plasmid were co-electroporated into F3 hiPSCs, which were subsequently plated in one well of a 6-well plate. Electroporations were performed using a Lonza Nucleofector with 96-well shuttle, with 200,000 hiPSCs in 20 µL of P3 buffer. Pulse setting CA-137 was used for all electroporations. Selection with the appropriate agent was begun 48 hours after electroporation and continued for 5 days.

After selection with puromycin (400 ng/mL), colonies were picked manually with a pipette. The hiPSC lines generated were genotyped by PCR for the presence of wild-type and reporter alleles. Homozygous clones were further verified by PCR amplification of the entire locus and Sanger sequencing. To excise the selection cassette, hiPSCs were electroporated with a plasmid expressing Dre recombinase. Selection was performed with ganciclovir (4 µM) and colonies were picked as described above. The excision of the selection cassette was verified by genotyping. Reporter lines were screened for common karyotypic abnormalities using a qPCR kit (Stemcell Technologies) followed by verification via Thermo Fisher Cell ID + Karyostat and Pluritest services.

cDNA vector creation

Vectors for cDNA overexpression were generated via MegaGate cloning. Full length cDNAs for each TF of interest were either derived from the human ORFeome or synthesized as full-length constructs. All 47 ORFs were cloned into pENTR221 with stop codons and minimal Kozak sequences. MegaGate was utilized to insert ORFs into the final PB-cT2G-cERP2 3' UTR barcode-modified expression vectors (Addgene 175503). Three unique barcodes were selected for each ORF with an average hamming distance of six. The three barcoded vectors for each ORF were then pooled, such that for each individual TF there was a mixture of three barcoded vectors. Sanger sequencing was performed across the entire ORF length to confirm canonical sequence with no amino acid changes.

Generation of inducible TF hiPSC cell lines

Expression plasmids containing TF cDNAs under the control of a doxycycline-inducible promoter were integrated into hiPSCs using piggyBac transposase. To perform the integration, 100 fmol of TF cDNA plasmid, 200 ng piggyBac transposase expression plasmid, and 100,000 to 200,000 hiPSCs were combined in 20 µL of Lonza P3 buffer and electroporated using a Lonza Nucleofector 4D. Pulse setting CM-113 was used for all electroporations. After electroporation, cells were seeded in 24-well plates in mTeSR Plus + 10 µM Y-27632. Selection with 400 ng/ml Puromycin began 48 hours after electroporation and continued for 3-5 days. Cells were then passaged without drug selection for 3 days to allow for non-integrated plasmid loss. Finally, cells were again passaged under drug selection to generate a pure, polyclonal integrant pool. Presence and approximate copy number of integrated TF plasmids was confirmed by qPCR on genomic DNA. For hPGCLC and oogonia generation, polyclonal pools of hiPSCs were utilized, not single cell selected clones. Average copy number was 8-10. The same procedure was performed for generating combinatorial cell lines, in which the 100 fmol of cDNA vector was divided equally between each TF for a pooled nucleofection. For copy number assessment the following was performed: 1) RT-qPCR using SYBR Green master mix was performed

after gDNA extraction using the DNAeasy kit. 2) 10 ng of input gDNA was used per reaction based on the standard curve, with an anneal temperature of 60 degrees. 3) To calculate copy number, the 2ΔCq+1 method was used, with RNAseP as a reference. 4) The resultant value was multiplied by two to account for the two autosomal copies of RPP30.

Generation of hPGCLCs and iOLCs in 2D Monolayer

For generation of hPGCLCs and iOLCs, an identical induction format and media composition was utilized. hiPSCs containing integrated TF expression plasmids were cultured in mTeSR1 medium on Matrigel coated plates. For induction in monolayer, hiPSCs were dissociated to single cells using Accutase and seeded onto Matrigel or vitronectin XF coated plates at a density of 2,500- 3,000 cells per cm2 in mTeSR1 + 10μM Y-27632 and 1μg/ml doxycycline for 6 hours. Media was then removed and washed with DMEM/F12 and replaced with Media 1 (see components list below). After 12-18 hours of induction, Media 1 was removed and washed with DMEM/F12 and replaced by Media 2. After 24 hours, Media 2 was removed and replaced by Media 3. After 24 hours, Media 3 was replaced with Media 4. hPGCLCs and iOLCs could then be harvested for use after 24 hours in Media 4 or additionally after two further days of culture in Media 4 (at day 6 of the protocol). hPGCLCs could be isolated via the NANOS3 reporter expression, CD38 cell surface expression, combinations of both or EpCAM/ITGA6 dual positive cell surface markers. Oogonia-like cells (iOLCs) could be isolated via a DDX4 reporter. hPGCLCs and iOLCs could additionally be generated via embryoid formation through methods established in Yamashiro et al. 2018, Kobayashi et al. 2022, Murase et al. 2021.

Media formulations were as follows: **Basal media (aRB27):** Advanced RPMI, 1X B27 minus Vitamin A, 1X Glutamax, 1X Non-Essential Amino Acids, 10 μM Y-27632, 1X Primocin or Pen-Strep , 1 μg/ml Doxycycline. **Media #1 (Epiblast-induction media):** aRB27 Basal Media, 3 μM CHIR99021, 100 ng/ml Activin A, 0.1 μM PD173074. **Media #2**: aRB27 Basal Media, 1 μM XAV939, 40 ng/ml hBMP4. **Media #3:** aRB27 Basal Media, 1 μM XAV939, 100 ng/ml SCF, 50 ng/ml EGF. **Media #4**: aRB27 Basal Media, 1 μM XAV939, 40 ng/ml hBMP4, 100 ng/ml SCF, 50 ng/ml EGF.

Fluorescent Activated Cell Sorting of hPGCLCs and iOLCs

hPGCLCs and iOLCs were analyzed using flow cytometry on the BD LSRFortessa or Cytoflex LX machine. Negative gates were set using hiPSC controls or unstained cell controls (Figure S8). For cell sorting for post-isolation growth, cells were captured using a Sony SH800 cell sorter. All flow cytometry analysis was conducted using the FlowJo software system (v10.8.1). NANOS3, TFAP2C, DDX4, and NPM2 were captured via fluorescent reporters. Live cells were gated as DAPI negative and as singlets. For cell surface markers, CD38 PerCP-Cy5.5 Mouse IgG (Biolegend 303522), EPCAM-APC-Cy7 Mouse IgG (Biolegend 324245), and Integrin-alpha-PE Rat IgG (Biolegend 313607) were utilized. A 1:60 dilution of each antibody in a dPBS + 3% FBS FACS buffer was utilized. Staining was performed for one hour at 4C at harvest with Accutase and cells were sorted without fixation with additional DAPI staining.

Transcriptomic characterization of TF induced cell lines

Cells were induced according to the above protocols and isolated for bulk RNA-sequencing. Library preparation was performed using the NEB Next ultra-low input RNA-sequencing library preparation kit with PolyA capture module for samples containing less than 10,000 cells. For samples with greater than 10,000 cells the NEBNext Ultra II RNA-sequencing library preparation kit with PolyA mRNA capture module was utilized. Sequencing was performed on Illumina Next-Seq 500 and NovaSeq platforms with a 2 x 100bp configuration. Only samples with RNA quality RIN scores of greater than 8 were utilized for analysis.

For DLX5, HHEX, and FIGLA hPGCLC bulk RNA-sequencing, TFs were induced for four days and cells were sorted for NANOS3+ expression and utilized for bulk RNA-Seq. No TF control cells were also utilized and sorted via NANOS3. For long term culture (LTC) hPGCLCs, data was downloaded from GEO and realigned using our pipeline. For D3 and DNR3 bulk RNA-Seq, TFs were induced for four days and cells were sorted for DDX4+ expression and utilized for bulk RNA-Seq. For reference samples of the ovarian atlas, data was downloaded from GEO and realigned using our pipeline.

Ovarian Atlas Generation and Bulk RNA-Seq Analysis

An automated RNA sequencing (RNA-Seq) analysis pipeline was constructed to determine the extent of similarity between engineered cell types and their natural human counterparts. Briefly, an atlas of 100 deposited RNA-seq FASTQ files was curated from various studies, where ovarian somatic and germ cells were obtained or derived from human samples, further analyzed, and deposited.[1] The final atlas consists of granulosa cell data at various stages of fetal and adult ovarian development,[2-6] as well as oogenesis data from stem cells through primordial germ cell specification, and finally to oogonia and oocyte from various stages of follicular development.[6,7]

In batch, raw data files alongside collected RNA-Seq datasets were aligned to the latest build of the human reference genome (GRCh38) utilizing the Spliced Transcripts Alignment to a Reference (STAR) alignment tool, to construct count matrices aligning sequencing reads to the known set of human genes.[8] The standard DESeq2 analysis package in R was used to estimate variance-mean dependence in count data, and subsequently calculated differential expression of each gene for every sample utilizing a negative binomial distribution.[9]

The Transcriptome Overlap Measure (TROM) method was employed to identify associated genes that capture molecular characteristics of biological samples and subsequently comparing the biological samples by testing the overlap of their associated genes.[10] TROM scores were calculated as the $-\log_{10}$(Bonferroni corrected *p* value of association) on a scale of 0-300. The TROM magnitude is positively correlated with similarity between two independent samples, with a standard threshold of 12 as an generally-accepted indicator of significant similarity.

Single Cell RNA-Sequencing Analysis

FASTQ files were obtained from Nova-Seq outputs, and aligned to the human reference genome (GRCh38) using STAR to generate the standard tripartite output file structure: DGE.mtx, cell_metadata.csv, and all_genes.csv, which together enable scRNA-seq matrix representation in an anndata format. The ScanPy analysis tool was utilized to filter outlier cells and normalize read counts via log transformation. Cells were clustered via the unsupervised UMAP algorithm and the STREAM pipeline was utilized to compute and visualize complex branching trajectories in pseudotime.

Marker gene lists for all human cell types were curated and filtered from PanglaoDB for cell types of interest. Single cells were assigned a score for each cell type by first calculating the average expression of the cell type marker genes and subsequently subtracting the average expression of a reference set of genes, randomly sampled from the overall gene list for each binned expression value.

For atlas analysis, fetal germ cell data was obtained from the recent Li Li, et al (2017) and was merged with a pre-curated cell type matrix to create a reference single cell atlas.[11] After log normalization, 2000 variable genes were selected for both datasets, cells were mapped onto the top 20 principal components created from these variable genes, and the datasets were merged using the Harmony algorithm

(https://github.com/immunogenomics/harmony) to account for batch effects. Using Symphony (https://github.com/immunogenomics/symphony), each sample was mapped onto this reference atlas and each cell classified using the k-nearest neighbors (KNN) soft clustering methodology.[29]

Enzymatic Methylation Sequencing Analysis

Samples of D3 and DNR3 iOLCs were induced through TF overexpression in our standard 4 day protocol. Cells were harvested and sorted for DDX4+ expression using FACS. F2 female and A4 male Long term hPGCLCs were cultured under expansion conditions set about in Kobayashi et al. 2022 and harvested. F3 hiPSCs and PGP1 hiPSCs were grown in mTeSR1 and harvested. Cells for all samples were then harvested and gDNA was extracted using the Qiagen DNAeasy micro kit. The NEB Enzymatic methylation sequencing library preparation kit was then utilized for library preparation and indexing. Indexed libraries were run on a Bioanalyzer to ensure correct size and library quality. Libraries were then sequenced on an Illumina NovaSeq S1 flow cell at 15X genome for each sample. FASTQ reads were then demultiplexed and input for analysis. The bwa-meth (https://github.com/brentp/bwa-meth) pipeline was utilized to align reads from input FASTQ files to the human reference genome (GRCh38). The aligned reads were then de-duplicated with Picard (https://broadinstitute.github.io/picard/) and indexed with samtools (https://github.com/samtools/samtools). DNA methylation calling was achieved using MethylDackel (https://github.com/dpryan79/MethylDackel) for global methylation analysis, and differentially-methylated regions (DMRs) from were detected using metilene (https://www.bioinf.uni-leipzig.de/Software/metilene/). Final analysis and visualization was done via the methrix Bioconductor package (https://www.bioconductor.org/packages/release/bioc/html/methrix.html).

CUT&RUN characterization of TF binding

For CUT&RUN, the HA-tagged TF-containing D4TN3V hiPSCs were seeded and induced in monolayer under 1 µg/ml doxycycline. Unsorted cell populations were harvested at day 1, day 2, and day 4 of induction and subjected to CUT&RUN library preparation. Briefly, the EpiCypher CUTANA CUT&RUN Kit was used in accordance with protocol V2.0 for all CUT&RUN samples. ConA beads were resuspended and 11 µL of solution per CUT&RUN sample was utilized for bead activation. ConA beads were attached to a magnetic separation rack, and the supernatant was removed. Then, 100 µL per reaction of cold Bead Activation Buffer was added and beads were resuspended in solution. This process was repeated once more, and ConA beads were finally resuspended in 11 µL per reaction cold Bead Activation Buffer. Adherent cells were harvested with Accutase, spun down at 600 x g at RT, and supernatant was aspirated.

Cells were subsequently washed with 100 µL per sample with RT Wash Buffer, and this process was repeated once more. Finally, cells were resuspended in 100 µL per reaction of RT Wash Buffer. The samples were then mixed with 10 µL of activated ConA beads per 100 µL washed cells. These samples were incubated at RT for 10 minutes, and then were placed on a magnet to remove supernatant. After removal of supernatant, 50 µL of ice cold antibody buffer was added to each sample, and samples were briefly vortexed to resuspend cells. A defined antibody was then added to each sample, IgG (0.41 µL) or HA (1 µL), depending on the sample. All samples were then vortexed briefly, and placed on a nutator overnight at 4°C. The next day, samples were placed on a magnet until the slurry cleared, and supernatant was removed. Immediately, 200 µL of ice cold Digitonin Buffer was added to the samples, and then was subsequently removed. This process was repeated once more, and then 50 µL of ice cold Digitonin Buffer was added to each sample and samples were briefly vortexed. To each sample, 2.5 µL of pAG-MNase was added and samples were briefly vortexed and incubated for 10 minutes at RT. After 10 minutes, samples were placed on a magnet until the slurry cleared, and supernatant was removed. Immediately, 200 µL of ice cold Digitonin Buffer was added to the samples, and then was subsequently removed. This process was repeated once more, and then 50 µL of ice cold Digitonin Buffer was added to each sample and samples were briefly vortexed. To begin chromatin digestion, 1 µL of

CaCl2 (100 mM) was added to each sample, and samples were vortexed and subsequently incubated on a nutator for 2 hours at 4°C. The chromatin digestion was then stopped by the addition of 33 µL of Stop Buffer to each sample. Samples were briefly vortexed, and then were incubated in a thermocycler for 10 minutes at 37°C to release chromatin and degrade RNA. Samples were then frozen at -20°C until Phenol Chloroform Extraction.

Phenol Chloroform Extraction:

Samples were thawed and briefly spun down on a microcentrifuge. One volume (83 µL) of phenol:chloroform:isoamyl alcohol (25:24:1) was added to each sample and mixed thoroughly. Samples were then transferred to a PhaseLock tube, and subsequently spun down for 5 minutes at RT at 16,000 x g. The upper aqueous phase of each sample was then transferred to a new tube containing 1 µL of glycogen. The samples were then mixed with 800 µL of ethanol, and placed at -80°C overnight to precipitate the DNA. The next day, samples were spun down at 4°C for 10 minutes at 16,000 x g to pellet cDNA. Supernatant was removed, and 1mL of 80% ethanol was added to wash cDNA. Samples were spun down at 4°C for 10 minutes at 16,000 x g to pellet cDNA. Supernatant was removed, and lids were left open to allow all samples to air dry for 10-30 minutes. DNA was then reconstituted in a 0.1X TE buffer by incubation at 37°C for 15 minutes. To remove gDNA, samples were mixed with ProNex beads at a ratio of 1.1X for 5 minutes. Samples were then placed on a magnet, and supernatant was transferred to new tubes containing a quantity of ProNex beads that increased the ratio to 1.9X. After 5 minutes of incubation, samples were placed on a magnet and supernatant was removed. Each sample was washed twice with 200 µL of ProNex wash buffer. Finally, beads were allowed to partially dry for 2-4 minutes, and were then resuspended in 50 µL of 0.1X TE buffer. Samples were placed on a magnet, and eluted DNA was transferred to new tubes and frozen at -20°C until library preparation.

NEBNext Ultra II DNA Library Prep

The NEBNext Ultra II DNA library prep protocol was followed, with halved reaction sizes for all steps. Briefly, 25 µL of each sample was mixed with 1.5 µL NEBNext Ultra II End Prep Enzyme Mix and 3.5 µL NEBNext Ultra II End Prep Reaction Buffer, and was then placed in a thermocycler and heated first at 20°C for 30 minutes, and then at 50°C for 60 minutes. The samples were then mixed with 15 µL NEBNext Ultra II Ligation Master Mix, 0.5 µL NEBNext Ligation Enhancer, and 1.25 µL NEBNext Adaptor for Illumina (0.6 µM). Samples were mixed and then incubated at 20°C in a thermocycler for 15 minutes. After incubation, 1.5 µL of USER Enzyme was added to each sample, and samples were incubated at 37°C for 15 minutes in a thermocycler. Adaptors were then cleaned from samples by bead purification. Samples were mixed with 90 µL of ProNex beads and incubated for 10 minutes. After applying samples to a magnet, supernatant was removed and samples were washed twice with 200 µL ProNex Wash buffer. Samples were allowed to dry, and then resuspended in 17 µL of 0.1X TE buffer. PCR amplification proceeded after transferring 15 µL of each sample to a new tube, and mixing each sample with 5 µL of NEBNext Multiplex Oligos for Illumina and 20 µL of Q5 Master mix (2X). PCR was performed according to the NEBNext Ultra II DNA Library Prep cycling conditions, consisting of 1 cycle of 98°C for 30 seconds, followed by 11 cycles of 98°C denaturation for 10 seconds and 65°C annealing/extension for 75 seconds. After cycling, a final extension at 65°C was conducted for 5 minutes. Samples underwent PCR cleanup with ProNex beads, and were subsequently analyzed by Qubit and in a pooled format on a Bioanalyzer. Samples were sequenced on the Illumina NovaSeq platform using the 2X 100 bp format.

Peak files of each of the replicates were merged to create a conserved peak file for each TF. The bedtools intersect function was used to take the overlapped peak regions from different replicates with a minimum length longer than 100 bp. We further separated the merged peak regions into TSS -proximal and distal-regions based on the distance to TSS. Specifically, TSS-proximal regions (+/- 2kb) were identified by utilizing the diffloop package function getHumanTSS with 2kb of padding on either side. The bedtools intersect function was then used to identify all peaks from each factor's CUT&RUN merged peaks that overlapped TSS

proximal positions. Peaks that overlapped the TSS proximal regions and peaks that did not have any overlap with TSS proximal regions were separately concatenated and merged using bedtools on default parameters for “intersect -wa” to generate factor specific proximal and distal peaksets, respectively. Graphs for TSS relative binding dynamics and genome associations were generated from bed files using the GREAT software suite. For Peak visualization at specific loci, the WashU Epigenome Browser was utilized with bw files. For each TF and genomic region of interest a fixed Y-axis was utilized between 2,000 to 4,000 depending on the peak density. A smoothing function set to 1px was utilized for visualization. All other parameters were set to default with the aggregate method set to ”average”.

Homozygous CRISPR/Cas9 KO of DLX5, HHEX, and FIGLA

CRISPR/Cas9 KO of DLX5, HHEX, and FIGLA was performed using the Synthego CRISPR KO V2 kit. Briefly, the Synthego guide finder was utilized to generate triple sgRNAs for each target with the following sequences: DLX5: UAGGAGAGCAGUAGCCGUGC, AAAGUUGGCGAUUCCUGAGA, GACAGGAGUGUUUGACAGAA HHEX: CCAUGCAGUACCCGCACCCC, CCCGACGCCCUUUUACAUCG, AGUUGGGGGACGGCAGCGUG FIGLA: GCCGCUCCAGCACCAACUGG, CCGCUUGAGCCGGCAGACAG, UUGCGGGGUGCCCAGGAGCG

sgRNAs and Cas9 protein were combined according to manufacturer instructions to form ribonucleoprotein complexes. 500,000 D4TN3V hiPSCs were nucleofected using program CA137 with the sgRNA/Cas9 RNPs and seeded into one well each of a six well in mTeSR Plus media on Matrigel coated wells for each TF. Cells were cultured for 5 days to allow for recovery from nucleofection and to reach 60-70% confluency. Cells were then harvested using Accutase and diluted to 1 cell per 2 µl and seeded into 96 well plates for single cell cloning, in which 2 µl were added to each well. CloneR stem cell single cell cloning media supplement was utilized to improve viability. Single cell clones were then identified and expanded over two weeks and passaged. Genomic DNA was harvested from single cell clones using quick extraction buffer and gDNA was subjected to PCR amplification for each target locus with the following primers. DLX5: DLX5_Fwd: GCTGGTTGGTGGCGCTT, DLX5_Rev: GCCATGTCTGCTTAGACCAGA, DLX5_Seq: CTTGGGACGCGGTTGTAGGC, HHEX: HHEX_Fwd: TGGTAGCAAGCGCGTCC, HHEX_Rev: CGTGTAGTCGTTCACCGTCC, HHEX_Seq: GGGCGGGGAGCTTGGCCGGCGGGGCGGGCC, FIGLA: FIGLA_Fwd: CCACTGCCATTCTCCTGACA, FIGLA_Rev: GAATGGTTGCAGGTGGTCCA, FIGLA_Seq: CCACTGCCATTCTCCTGACAAG.

gDNA amplifications were then sent for Sanger Sequencing to identify indel formations. For each TF, a single cell clone was selected that demonstrated a frameshift mutation with an average 29-43 bp homozygous deletion. Each homozygous knockout clone was then utilized for hPGCLC induction versus wild type and for rescue experiments. For rescue experiments, the single knockout clone was nucleofected with 100 fmol of inducible cDNA vector and 200 ng of piggyBac transposase, with subsequent puromycin drug selection to purify the population. The polyclonal pool was then induced in the presence or absence of 1µg/ml doxycycline to determine rescue.

Dosage and Cytokine requirement hPGCLC assessment

For dosage testing, DLX5, HHEX, and FIGLA-containing D4TN3V hiPSC lines were seeded in triplicate under monolayer induction according to our above protocols. Differential concentrations of doxycycline were added to the culture to determine the effect of TF dosage on hPGCLC yield. For testing we compared 0ng, 50ng, 100ng, 200ng, 400ng, 800ng, 2000ng. For cytokine dependence, DLX5, HHEX, and FIGLA containing D4TN3V hiPSC lines were seeded in triplicate. hPGCLC induction was performed with and without 1µg/ml dox in the presence and absence of hBMP4.

iOLC expansion culture

Feeder-free expansion of iOLCs post-isolation was accomplished on Matrigel coated plates with growth in S-CM media, established by Kobayashi et al. 2022, which is an STO-feeder cell conditioned media supplemented with SCF. The iOLC expansion protocol is identical to the expansion protocol for long term culture of hPGCLCs and shows maintenance of DDX4+ expression over 30 days. The hPGCLC basal medium contained 13% (v/v) KSR, 1x NEAA, 1 mM sodium pyruvate, and 1x penicillin-streptomycin in Glasgow's MEM with 2 mM glutamine (ThermoFisher, 11710035). STO-CM was prepared by maintaining 5.0e6 mitomycin C treated STO cells in 12 mL of hPGCLC basal medium for 24 h, removing cells by centrifugation, and storing frozen at -20°C until use. The complete hPGCLC maintenance medium (S-CM for SCF-supplemented CM) was prepared by adding 0.1 mM b-mercaptoethanol, 50 mg/mL L-ascorbic acid, and 100 ng/mL recombinant human SCF to the CM. hPGCLC expansion culture FACS-enriched NANOS3+ hPGCLCs (5000–20,000 cells) or DDX4+ iOLCs (100-5000 cells) were inoculated onto Matrigel coated plates in a well of six-well plates (Corning, 3506) with SCF-supplemented hPGCLC basal medium containing 10 μM Y27632 + 1 μg/ml doxycycline. Medium was changed every other day without Y27632. During the iOLC and hPGCLC expansion, cells were dissociated using Accutase and passaged onto fresh matrigel every 5–14 days, when cells reached 50% confluency.

Immunofluorescence imaging

Cells cultured on Matrigel-coated ibidi 8-well plates (ibidi, cat 80806) were washed once with 200 μL dPBS and fixed by treatment with 200 μL of 4% paraformaldehyde in dPBS for 10 minutes at room temperature. The dPBS wash was repeated twice, and the cells were permeabilized by treatment with 200 μL 0.25% Triton X-100 in dPBS for 10 minutes at room temperature. The cells were washed with 200 μL PBST (0.1% Triton X-100 in dPBS), and blocked with 100 μL blocking buffer (1% bovine serum albumin and 5% normal donkey serum [Jackson ImmunoResearch, cat 017-000-121, lot 152961] in PBST) for 30 minutes at room temperature. The blocking buffer was removed and replaced with a solution of primary antibodies in blocking buffer, and the cells were incubated overnight at 4°C. The antibody solution was removed and the cells were washed three times with 200 μL PBST. The cells were incubated with 100 μL secondary antibody solution in blocking buffer for 1 hour at room temperature in the dark. The secondary antibody was removed and replaced with 200 μL of DAPI solution (1 ng/mL in dPBS). After 10 minutes the DAPI solution was removed and the cells were washed twice with 200 μL dPBS, and stored in dPBS at 4°C in the dark until imaging (typically a few hours). Imaging was performed on a Leica SP5 confocal microscope. Antibodies used are listed as follows: **Primary Antibodies:** OCT4- Mouse IgG (AB398736) - 1:250 Dilution, DDX4- Mouse IgG (AB27591) - 1:500 Dilution, DAZL- Rabbit IgG (AB215718) - 1:167 Dilution, SOX17- Goat IgG (AB355060) - 1:500 Dilution, ITGA6- Rat IgG (AB493634) - 1:167 Dilution (1:60 for flow cytometry). **Secondary Antibodies:** Mouse IgG-AF647 Donkey (AB162542) - 1:250 Dilution, Rabbit IgG-AF488 Donkey (AB2534104) - 1:500 Dilution, Goat IgG-AF568 Donkey (AB2534104) - 1:500 Dilution.

Quantification and Statistical Analysis

All statistical analysis and specific quantification can be found in the figure legends and methods sections. All graphing was performed using Graphpad unless otherwise noted. For Figure 1C, n=3 replicates were independently seeded for induction from a single hiPSC cell line. For Figure 1D, n=2 replicates were independently seeded for induction from a single hiPSC cell line. Figure 1E, n=3 replicates were independently seeded for induction from a single hiPSC cell line. Figure 1F, n=3 replicates were independently seeded for induction from a single hiPSC cell line. For Figure 2A, n=2 replicates were independently seeded for induction and sequencing from a single hiPSC cell line and averaged for analysis. For Figure 2B, n=3 replicates were independently seeded for induction and imaging from a single hiPSC cell line and images are representative of

overall population. For Figure 2A, n=2 replicates were independently seeded for induction and sequencing from a single hiPSC cell line and averaged for analysis. For Figure 2D, n=1 replicate was seeded for induction and single cell sequencing from a single hiPSC cell line.

Supplementary Figures

S1. Development of an efficient monolayer hPGCLC induction platform

S2. DLX5 overexpression rescues loss of BMP4 in hPGCLC differentiation

S3. DLX5, HHEX, FIGLA overexpression drives hPGCLC formation across differentiation paradigms

S4. DLX5, HHEX, and FIGLA are not necessary for NANOS3+ hPGCLC formation

S5. Expression of TFs during human fetal gonad development

S6. Epigenome profiling of hPGCLC and iOLC formation

S7. iOLC monolayer expansion feeder-free

S8. Representative flow cytometry analysis

Author Contributions

C.K. and P.C. conceived, designed, and directed the study. P.C. developed and implemented GRN inference and centrality analysis algorithms. C.K. performed all TF screening, cell line creation, differentiation protocol development, flow cytometry, cell engineering, iOLC expansion culture methods, and sequencing library preparation. M.P.S. generated reporter cell lines, performed confocal imaging, aided in flow cytometry, and aided in TF screening. B.W. performed CUT&RUN library preparation and analysis. P.R.F. aided in confocal imaging and library preparation. P.C. and G.B. developed and executed single cell analysis pipelines, with assistance from Te.S. P.C., G.B., K.P., S.B., and E.T. conducted bulk RNA and methylation analysis. E.D., J.A., and S.K. assisted in vector cloning, TF screening, and confocal imaging. S.L. aided in single cell atlas integration. M.K. performed fetal gonad TF expression analysis. P.C. supervised the study, with assistance from R.E.K., To.S., and G.M.C.

Data and Materials Availability

All data needed to evaluate the conclusions in the paper are present in the paper and supplementary tables and figures. Data analysis code can be found at: https://github.com/programmablebio/egg. Raw and processed sequencing data will be deposited to GEO upon publication.

Competing Interests

P.C., C.K., M.P.S., and G.C. are listed as inventors for U.S. Provisional Application No. 63/326,656, entitled: “Methods and Compositions for Producing Primordial Germ Cell-Like Cells,” and U.S. Provisional Application No. 63/326,607, entitled: “Methods and Compositions for Producing Oogonia-Like Cells.” P.C. is a co-founder and scientific advisor to Gameto, Inc. C.K. is the Chief Scientific Officer of Gameto, Inc. G.M.C. serves on the scientific advisory board of Gameto, Inc., Colossal Biosciences, and GCTx.

Acknowledgements

This work was funded by the Synthetic Biology Platform at the Wyss Institute for Biologically Inspired Engineering. This work was additionally funded by the Gameto Sponsored Research Agreement at Harvard University and Colossal Sponsored Research Agreement at Harvard University. M.P.S. was supported by the National Science Foundation Graduate Research Fellowship.

References

1. NSFG - listing I - key Statistics from the National Survey of family growth. https://www.cdc.gov/nchs/nsfg/key_statistics/i.htm (2019).
2. GBD 2017 Population and Fertility Collaborators. Population and fertility by age and sex for 195 countries and territories, 1950-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1995–2051 (2018).
3. Zorrilla, M. & Yatsenko, A. N. The Genetics of Infertility: Current Status of the Field. *Curr. Genet. Med. Rep.* **1**, (2013).
4. Hikabe, O. *et al.* Reconstitution in vitro of the entire cycle of the mouse female germ line. *Nature* vol. 539 299–303 (2016).
5. Yoshino, T. *et al.* Generation of ovarian follicles from mouse pluripotent stem cells. *Science* **373**, (2021).
6. Kojima, Y. *et al.* Evolutionarily Distinctive Transcriptional and Signaling Programs Drive Human Germ Cell Lineage Specification from Pluripotent Stem Cells. *Cell Stem Cell* **21**, 517–532.e5 (2017).
7. Yamashiro, C. *et al.* Generation of human oogonia from induced pluripotent stem cells in vitro. *Science* **362**, 356–360 (2018).
8. Nagaoka, S. I. *et al.* ZGLP1 is a determinant for the oogenic fate in mice. *Science* **367**, (2020).
9. Hamazaki, N. *et al.* Reconstitution of the oocyte transcriptional network with transcription factors. *Nature* **589**, 264–269 (2021).
10. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Mol. Cell* **72**, 1021–1034.e4 (2018).
11. Li *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* vol. 20 858–873.e4 (2017).
12. Kobayashi, M. *et al.* Expanding homogeneous culture of human primordial germ cell-like cells maintaining germline features without serum or feeder layers. *Stem Cell Reports* **17**, 507–521 (2022).
13. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
14. Kramme, C. *et al.* An integrated pipeline for mammalian genetic screening. *Cell Rep Methods* **1**, 100082 (2021).
15. Zhang, J. *et al.* OTX2 restricts entry to the mouse germline. *Nature* vol. 562 595–599 (2018).
16. Kramme, C. *et al.* MegaGate: A toxin-less gateway molecular cloning tool. *STAR Protoc* **2**, 100907 (2021).
17. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive *piggyBac* transposase for mammalian applications. *Proceedings of the National Academy of Sciences* vol. 108 1531–1536 (2011).
18. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471 (2013).
19. Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
20. Sebastiano, V. *et al.* Monolayer platform to generate and purify human primordial germ cells in vitro provides new insights into germline specification. *Research Square* (2021) doi:10.21203/rs.3.rs-113078/v1.
21. Kobayashi, T. *et al.* Principles of early human development and germ cell program from conserved model systems. *Nature* **546**, 416–420 (2017).
22. Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178–194 (2015).

23. Kojima, Y. *et al.* GATA transcription factors, SOX17 and TFAP2C, drive the human germ-cell specification program. *Life Science Alliance* vol. 4 e202000974 (2021).
24. Mitsunaga, S. *et al.* Relevance of iPSC-derived human PGC-like cells at the surface of embryoid bodies to prechemotaxis migrating PGCs. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9913–E9922 (2017).
25. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* (2021) doi:10.1101/gr.266551.120.
26. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**, (2019).
27. Li, W. V., Chen, Y. & Li, J. J. TROM: A Testing-Based Method for Finding Transcriptomic Similarity of Biological Samples. *Stat. Biosci.* **9**, 105–136 (2017).
28. Yatsenko, S. A., Wood-Trageser, M., Chu, T., Jiang, H. & Rajkovic, A. A high-resolution X chromosome copy-number variation map in fertile females and women with primary ovarian insufficiency. *Genet. Med.* **21**, 2275–2284 (2019).
29. Kang, J. B. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 1–21 (2021).
30. Wallace, W. H. B. & Kelsey, T. W. Human Ovarian Reserve from Conception to the Menopause. *PLoS ONE* vol. 5 e8772 (2010).
31. Gell, J. J. *et al.* An Extended Culture System that Supports Human Primordial Germ Cell-like Cell Survival and Initiation of DNA Methylation Erasure. *Stem Cell Reports* **14**, 433–446 (2020).
32. Murase, Y. *et al.* Long-term expansion with germline potential of human primordial germ cell-like cells in vitro. *EMBO J.* **39**, e104929 (2020).
33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

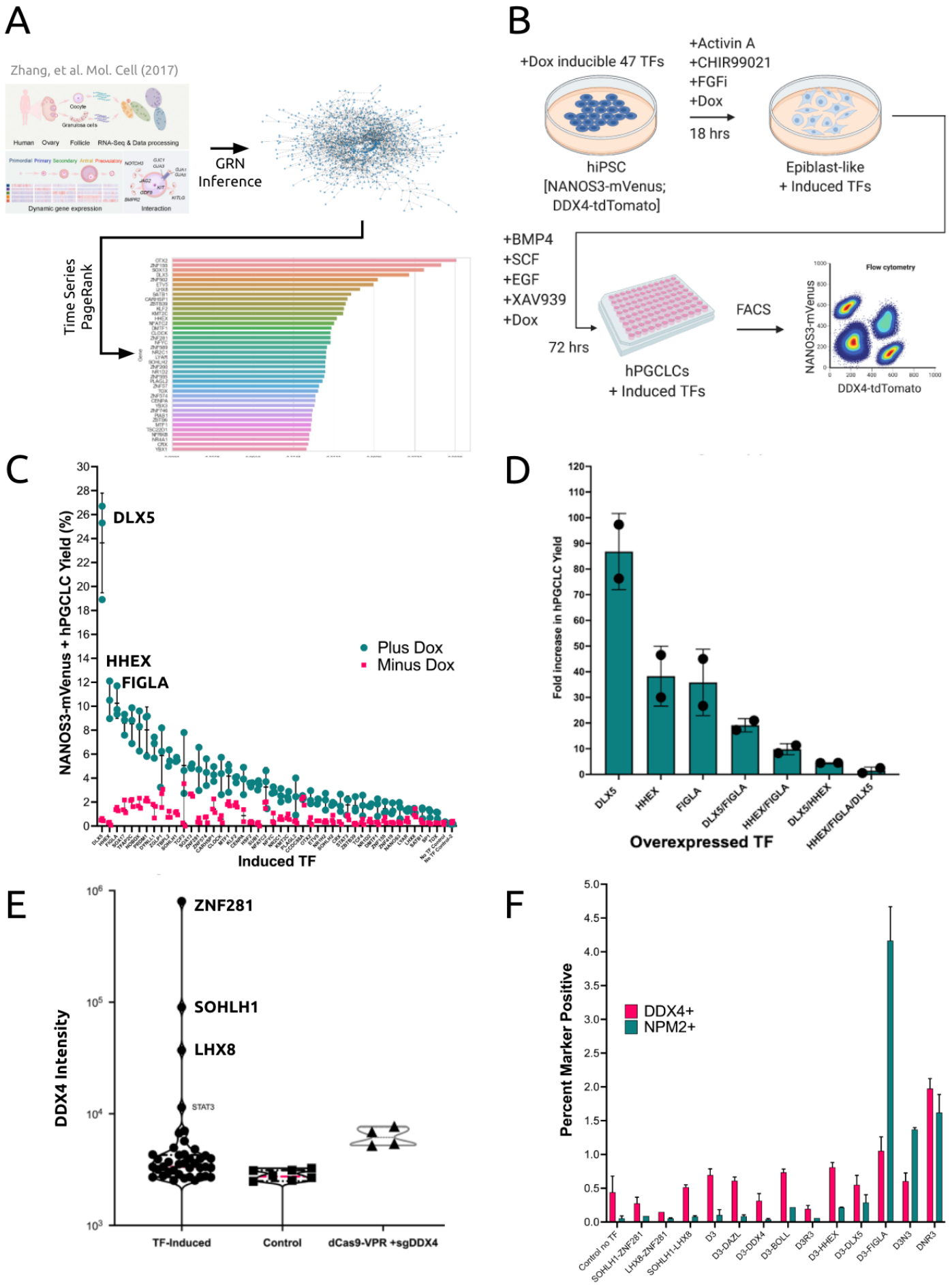

Figure 1. Identification of TFs driving hPGCLC and oogonia-like formation. A) Computational method workflow for inference of oogenesis gene regulatory network and target TF prediction. B) 2D monolayer screening format for TF-assisted hPGCLC and oogonia-like formation C) Individual induction of 47 computationally predicted TFs in hiPSCs during monolayer hPGCLC formation in the presence (teal) or absence (pink) of 1 μg/ml doxycycline. hPGCLC yield was determined via flow cytometry for NANOS3-mVenus expression. Data is plotted as values of three replicates with standard deviation for each TF. D) Combinatorial TF induction in the monolayer protocol in the presence or absence of 1 μg/ml doxycycline. Data is plotted as a fold increase in NANOS3-mVenus positive percent formation in plus versus minus doxycycline with standard deviation for two replicates. E) hiPSCs were induced in triplicate using the monolayer format for hPGCLC formation and DDX4-tdTomato expression was assessed via flow cytometry. Geometric mean fluorescent intensity of DDX4+ cells is graphed for each TF versus control no TF and versus a dCas9-VPR plus sgDDX4 positive control. F) Combinations of TFs were induced in triplicate using the monolayer format and assessed for DDX4-tdTomato expression (pink) and NPM2-mGreenLantern expression (teal). Data is plotted as mean percent positive for each marker with standard deviation.

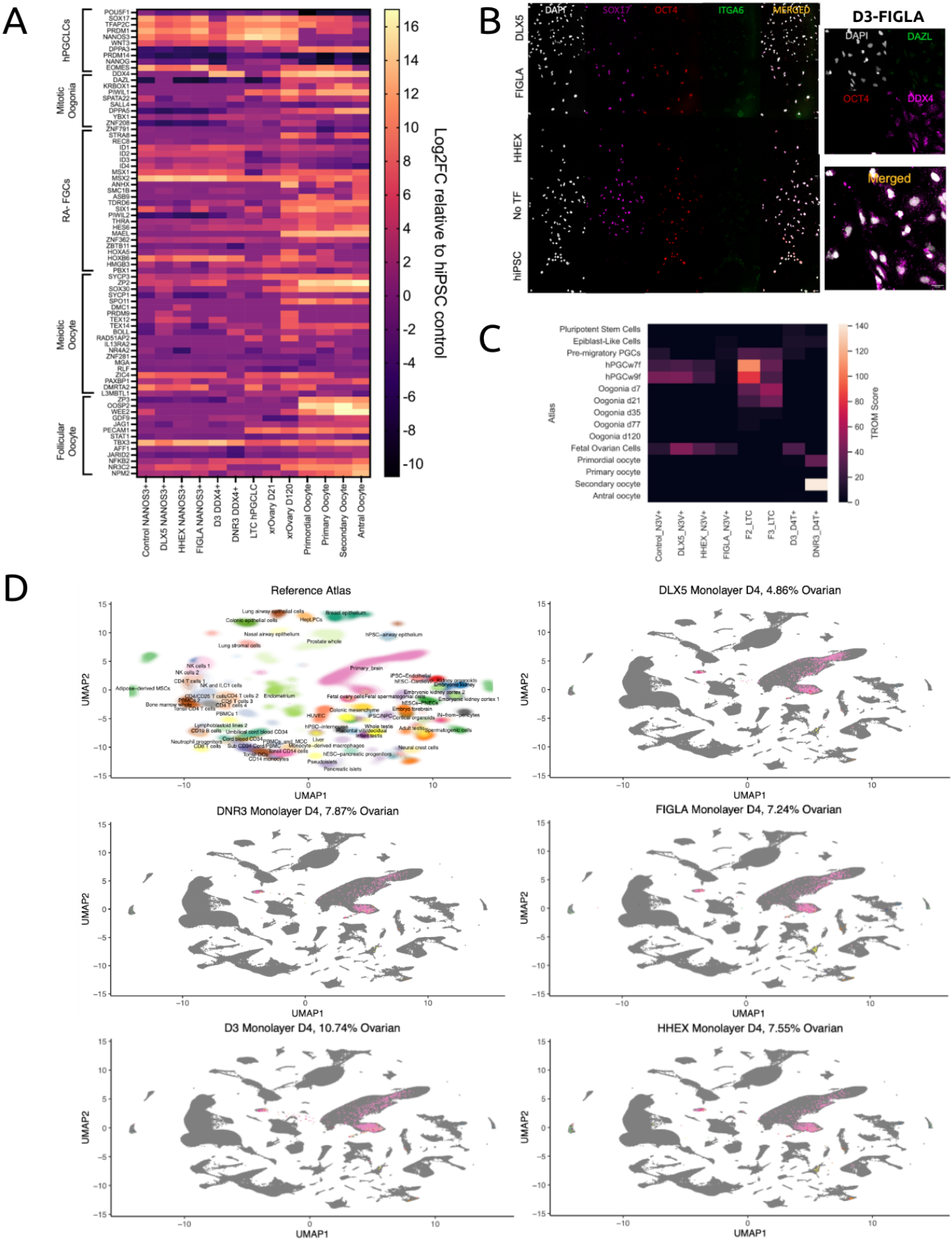

Figure 2. TF overexpression drives on-target germ cell formation A) NANOS3+ cells and DDX4+ cells were isolated at day 4 of differentiation with TF induction in the monolayer format for polyA mRNA bulk RNA-sequencing. Raw reads were aligned via STAR aligner. Raw reads from publicly available reference data were downloaded and re-aligned. Aligned reads were normalized in a single batch using DESeq2 and gene expression of select genes is plotted in a heatmap as a Log2-transformed fold change compared to an hiPSC control. B) (left) hiPSCs and hPGCLCs were subjected to immunofluorescent confocal imaging for SOX17 (purple), OCT4 (red), DAPI (white) and ITGA6 (green). (right) D3-FIGLA iOLCs were isolated for confocal imaging of DAPI (white), DAZL (green), OCT4 (red), and DDX4 (pink). C) Transcriptome overlap measure (TROM) analysis was performed on NANOS3+ hPGCLCs and DDX4+ iOLCs from TF inductions. TROM is measured on a scale of 1-300, with >12 as statistically significant. Reference samples were aggregated from publicly available data and batch normalized with query samples. D) A single cell human atlas was generated via aggregation of human reference samples and human ovarian scRNA-seq samples. Unsorted hPGCLCs and iOLCs were harvested for scRNA-seq. Reads were aligned via STAR aligner and integrated with the cell atlas samples for analysis using Symphony. Query samples are colored by assigned cell type, based upon the color scheme of the reference atlas, and overlaid over reference atlas in gray.