Human Germ Cell Specification via Combinatorial Expression of Prioritized Transcription Factors

Christian Kramme
Harvard Medical School

Merrick Smela
Wyss Institute

Bennett Wolf
Harvard Medical School

Patrick Fortuna
Wyss Institute

Garyk Brixi
Duke University https://orcid.org/0000-0002-1253-0522

Tian-Lai Zang
Duke University

Sophia Vincoff
Duke University

Kalyan Palepu
Duke University

Sabrina Koseki
MIT Media Lab https://orcid.org/0000-0001-7273-970X

Suhaas Bhat
Duke University

Edward Dong
Wyss Institute

Jessica Adams
Wyss Institute

Richie Kohman
Harvard Medical School

Songlei Liu
Harvard Medical School

Mutsumi Kobayashi
Massachusetts General Hospital Center for Cancer Research

Toshihiro Shioda
Massachusetts General Hospital

George Church

Harvard Medical School

Pranam Chatterjee (✉ pranam.chatterjee@duke.edu)

Duke University https://orcid.org/0000-0003-3957-8478

Article

Keywords:

Posted Date: March 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2652895/v1

Additional Declarations: Yes there is potential Competing Interest. 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, scientific advisor to Gameto, Inc. C.K. is currently the VP of Cell Engineering of Gameto, Inc. G.M.C. serves on the scientific advisory board of Gameto, Inc., Colossal Biosciences, and GCTx.

Human Germ Cell Specification via Combinatorial Expression of Prioritized Transcription Factors

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

1. Wyss Institute, Harvard Medical School
2. Department of Genetics, Harvard Medical School
3. Department of Biomedical Engineering, Duke University
4. Department of Computer Science, Duke University
5. Department of Obstetrics and Gynaecology, School of Medicine, Juntendo University
6. Massachusetts General Hospital Center for Cancer Research, Harvard Medical School

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

Abstract

The generation of germline cell types from human induced pluripotent stem cells (hiPSCs) represents a key milestone toward *in vitro* gametogenesis, which has the potential to transform reproductive medicine. Methods to recapitulate germline cell specification *in vitro* have relied on extensive, long term culture methods, the most notable of which is a four-month culture protocol employing xenogeneic reconstituted ovaries with mouse embryonic ovarian somatic cells. Recently, transcription factor (TF)-based methods have demonstrated the feasibility of exogenous factor expression to directly differentiate hiPSCs into cell types of interest, including various ovarian cell types. The protocols leveraged in these studies, however, utilize more local methods of factor selection, such as basic differential gene expression analysis, and lower-throughput screening strategies via iterative testing of a small set of TFs. In this work, we integrate our recently-described, more globally-representative graph theory and highly-parallelized screening protocols to globally identify and screen 46 oogenesis-regulating TFs for their role in human germline formation. We identify *DLX5*, *HHEX*, and *FIGLA* whose individual overexpression enhances hPGCLC formation from hiPSCs, as well as an additional set of three TFs, *ZNF281*, *LHX8*, and *SOHLH1*, whose combinatorial overexpression drives DDX4+ germ cell formation from hiPSCs. In contrast to previous methods, our protocol employs a simple four-day, feeder-free monolayer culture condition. We characterize these TF-based germ cells via gene expression analyses, and demonstrate their broad similarity to *in vivo* and *in vitro*-derived germ cells. Together, these results identify new regulatory factors that enhance *in vitro* human germ cell specification and further establish unique computational and experimental tools for human *in vitro* oogenesis research.

Main

In the human germline, primordial germ cells (PGCs) undergo sequential differentiation through sex-specific trajectories to form male or female gametes.[1] *In vitro* gametogenesis (IVG) is a powerful technique for the recapitulation of this germline development, often utilizing directed differentiation of pluripotent stem cells to sequentially specify discrete germline cell types.[2] Germ cells are challenging to obtain for research in humans due to inherent ethical and technical limitations in their *in vivo* isolation and utilization, since these critical intermediates span embryonic, fetal, and adult development stages.[3] With an estimated one in eight people in the US struggling with infertility, new cell models of human reproductive development are needed to accelerate research into sex cell formation and genetic regulation, investigate root causes of infertility, and to develop novel assisted reproductive technologies.[4]

Methods for generating primordial germ cell-like cells (hPGCLCs), which highly resemble pre-migratory *in vivo* PGCs, from stem cells show highly varied yield that is protocol and cell line dependent.[1,2,5–13] This variation, combined with methods such as 3D aggregates makes high throughput genetic studies of hPGCLC formation challenging to perform at scale. Recently, human DDX4+ oogonia were specified from pluripotent stem cells through a ~4 month chimeric culture model in xenogeneic reconstituted ovaries (xrOvaries).[14] However, this impressive method remains challenging to utilize, particularly in screening paradigms that require scaled, parallelizable cell type differentiation, due to the extensive culture periods in 3D chimeric aggregates, which require fetal mouse gonad dissection combined with human primordial germ cell specification. To date, no method yet exists for the direct generation of DDX4+ germ cells from stem cells in humans.

In mouse models, a number of studies have demonstrated complete female IVG, yielding developmentally competent metaphase II (MII) oocytes from stem cells.[15,16] Impressively, these various mouse models generated healthy, fertile offspring after fertilization and implantation, showing the power of IVG to successfully model reproductive development from stem cells. In addition, recent mouse studies have demonstrated the utility of transcription factor (TF) based methods for driving and enhancing germ cell development.[17,18] These TF-based methods benefit from their ease of use, rapid developmental timescales, and high efficiency and can serve as complementary or supplementary differentiation platforms to growth factor only methods. Limited TF overexpression methods in human models have been demonstrated in hPGCLCs, with seminal work identifying that the TFs *SOX17*, *TFAP2C*, and *PRDM1* and the GATA members *GATA3/2* comprise an essential TF network for hPGCLC specification.[1,6,10,11,19] Overall however, few TF-based human germ cell specification studies have been completed, likely due to a combination of challenging cell culture techniques, difficulty in utilizing tunable genetic engineering methods, and lack of compelling TF target prediction tools. Outside of this upstream TF network, few studies have identified and validated in human models the role of downstream TFs in driving or modulating germ cell development. To date, no TF induction-based methods have been demonstrated for deriving more advanced human germ cell types such as DDX4+ germ cells, oogonia, spermatogonia or oocytes from stem cells.

We hypothesized that TFs that perform highly connected regulatory roles in the final stages of female gamete development may be capable of driving robust germ cell formation and differentiation when overexpressed in stem cells. In this study, we perform a broad TF overexpression screen to assess the effect of overexpressing putative oogenesis-regulating TFs in human induced pluripotent stem cells (hiPSCs) and assay for formation of reporter positive human germ cell types. Using novel graph theory algorithms to identify central regulators of oogenesis, we infer TFs that may perform highly connected roles in the regulation of human primordial oocyte to ovulated oocyte development stages.[20] We then leverage high throughput, monolayer germ cell differentiation methods amenable to single factor and combinatorial genetic screening for identifying regulators of germ cell development. We first identify three TFs, *DLX5*, *HHEX*, and *FIGLA,* whose individual overexpression drives potent enhancement of hPGCLC formation in monolayer and floating aggregate cultures. We additionally demonstrate that *DLX5* overexpression partially rescues loss of hBMP4 during germ cell formation. We further identify *LHX8*, *SOHLH1*, and *ZNF281*, whose combinatorial overexpression directly generates DDX4+ germ cell formation in four days in monolayer culture. Building on these findings, we produce additional mixed TF and RNA binding protein overexpression combinations that drive increased DDX4+ and NPM2+ germ cell yield. We furthermore demonstrate feeder-free DDX4+ germ cell post-isolation maintenance that builds on and integrates readily with recently described germ cell specification and maintenance methods.[7] Using transcriptomic assays, we demonstrate these TF-based germ cells closely resemble *in vivo* and *in vitro* PGCs, PGCLCs, and oogonia. 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. Furthermore, these methods suggest a novel future platform for human IVG, expanding the breadth of genetic engineering and cell-based tools available for reproductive development studies.

Results

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

Previous studies to identify regulators of mouse oogenesis leveraged differential gene expression of the primordial-to-primary follicle transition, identifying TFs that regulated mouse oocyte maturation and were functionally capable of driving oocyte formation from stem cells.[9] Similarly, to identify candidate TFs that govern the dynamic gene regulatory networks (GRNs) of human oogenesis, which can thus be prioritized for overexpression studies, 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, including primordial, primary, secondary, antral, and preovulatory follicles.[7,14,21,22] After constructing this pseudo-time series transcriptomic matrix, we employed unsupervised, regression-based GRN inference via regularized stochastic gradient boosting to generate a representative, directed graph of oocyte state during folliculogenesis, where nodes represent genes and edge weights correlate to the regulatory effect between TFs and their downstream targets.[23] 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 36 most critical, "central" TFs within the GRN (Figure 1A, Table S1).[20] 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*.[18,24] We additionally include 10 TFs that ranked slightly lower on our predictive list (which only leveraged late stage oogenesis data), but are known to perform important roles in earlier specific stages of germ cell development such as PGC formation or that have been identified in similar mouse studies.[10,11,17,18] The list of the 36 TFs and 10 control TFs is presented in Table S2 with the isoform chosen for screening, if it was predicted or selected as a control and any known involvement in germ cell development. Associated GO term analysis of the 46 factors is shown in Figure 1B, with reproduction and oogenesis terms highlighted as well as select general GO terms. We furthermore assessed the expression of all 46 TFs in single cell RNA-sequencing (scRNA-Seq) datasets from human fetal gonads, which is shown in Figure 1C.[25] As can be seen, each of the 46 TFs is expressed in the fetal gonad germ cells in both male and female development. Overall, these results motivated us to combinatorially screen prioritized candidates in hiPSCs to resolve the human oocyte transcriptional network experimentally.

Identification of TFs regulating germ cell formation via overexpression screening

We constructed a NANOS3-T2A-mVenus; DDX4-T2A-tdTomato dual reporter hiPSC line (N3VD4T) and a DDX4-T2A-tdTomato; NPM2-T2A-mGreenLantern dual reporter hiPSC line (D4TP2G) using CRISPR-Cas9-mediated homology directed repair (HDR) (Figure S1A). Genotyping PCR and Sanger sequencing were used to determine proper insertion of the reporter constructs, with karyotype analysis to confirm the genetic stability of the hiPSC lines (Figure S1B). Using flow cytometry after CRISPR-activation induction of the DDX4 reporter locus, we confirmed that the reporter lines induce expression of the fluorescent marker upon induction of the target gene (Figure S1C). Utilizing common floating aggregate methods described previously, we demonstrated the ability of our N3VD4T reporter line to generate NANOS3+ hPGCLCs that correlate with CD38 and EpCAM populations traditionally used in hPGCLC studies (Figure S1D).

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 2A).[9] A full component list and induction protocol is provided in Supplementary File 2 and the material and methods. Monolayer methods were chosen for screening due to the ease and scalability of use compared to floating aggregate methods. We found that in our N3VD4T reporter line, the protocol as described by Sebastiano et al. resulted in a very low yield of NANOS3+ putative hPGCLCs (0.2%-0.5%) (Figure 2B “Control”).

We generated doxycycline-inducible vectors expressing a full length cDNA for each of the 46 TFs identified for screening using MegaGate cloning into a previously validated and highly optimized backbone vector for tunable expression in hiPSCs from our STAMPScreen pipeline (Figure S2A).[26] Each insert was validated via sanger sequencing to confirm the TF was free of unintended mutations. Each vector harbored a 20 bp barcode on the 3’ UTR of the cDNA, which could be identified in either RNA-seq or PCR of genomic DNA and next generation sequencing (NGS) and was piggyBac integratable (Figure S2B).[27] We generated 46 individual D4TN3V hiPSC lines harboring integrations of each TF individually through super piggyBac transposase-mediated insertion.[6,14,28] Polyclonal pools for each TF were utilized for screening purposes. Through induction of each individual line via doxycycline followed by Barcode capture in RNA-seq, we demonstrated expression of the intended TF and its associated barcode in all 46 lines utilized in the study, confirming each line’s genotype and functional gene expression of each vector (Figure S2C).

Utilizing our four day monolayer hPGCLC induction protocol, we assessed NANOS3+ hPGCLC yield via flow cytometry in the presence or absence of doxycycline for all 46 TFs in triplicate (Figure 2B). We show that 23 TFs (15 computationally predicted and 8 controls) drove a statistically significant increase in NANOS3+ cell yield compared to control, while the other 23 TFs demonstrated a trend towards upregulation, but not significantly. Remarkably, 3 TFs (*DLX5*, *HHEX*, and *FIGLA*) induced a NANOS3+ yield higher (23.6% +/- 2.4% SEM, 10.5% +/- 0.9% SEM, and 10.3% +/- 0.7% SEM) than that of three known TF regulators of hPGCLC development: *SOX17*, *TFAP2C*, and *PRDM1 (8.6% +/- 0.5% SEM, 8.5% +/-0.9% SEM, and 8.0% +/-1% SEM)* respectively.[1,11,19]

We additionally find that overexpression of *DLX5* alone is able to replace exogenous hBMP4 in the induction of hPGCLCs, driving potent hPGCLC formation in the absence of hBMP4, albeit lower than in the presence of hBMP4 (Figure S3A). We show that *DLX5* expression in our system is highly linked to doxycycline concentration, as expected, reaching levels of ~110,000 fold compared to wild type hiPSCs determined by qPCR (Figure S3B). We show that with increasing *DLX5* induction, that endogenous *BMP2* and *BMP4* do not statistically significantly increase, indicating that *DLX5* induction of hPGCLC formation in the absence of BMP4 is likely not mediated through direct upregulation of *BMP4* or *BMP2* (Figure S3B). We further show that hPGCLC yield increases with all three TFs *DLX5*, *HHEX*, and *FIGLA* at day 6 of monolayer hPGCLC formation using both the NANOS3 reporter and EpCAM/Integrin surface markers (Figure S4A). Additionally, overexpression of each of these three TFs individually increased hPGCLC formation in day 6 floating aggregate cultures, as quantified by both the NANOS3 reporter and CD38 cell surface marker expression, with DLX5 showing significant enhancement of hPGCLC yield (Figure S4B). We additionally assessed dual combinations of the three TFs and find that while they increase NANOS3+ hPGCLC yield compared to the no TF control, the combinations perform more poorly than individual TF overexpression (Figure 2C). Together, these results identify individual overexpression of *DLX5*, *HHEX*, and *FIGLA* improves hPGCLC formation.

We next assessed production of DDX4+ cells from hiPSCs via flow cytometry in the presence or absence of doxycycline for all 46 TFs in triplicate (Figure 2A), using the same screening experiment for NANOS3 yield but assessing the DDX4-T2A-tdTomato reporter channel. *DDX4* is broadly expressed throughout germ cell development, arising in gonadal hPGCs after migration to the fetal gonad and increasing in expression during oogonia/spermatogonia formation. Compared to control, the overall percentage of DDX4+ cells was not greatly enriched by any single TF (Figure 2D). However, a small percentage of cells with significantly elevated DDX4-reporter expression intensity was identified in the *ZNF281*, *LHX8*, and *SOHLH1* induction conditions (Figure 2D). We hypothesized that combinatorial expression of these factors may increase DDX4-reporter yield and generate a more reproducible yield of highly DDX4+ cells. To investigate this hypothesis, we generated three independent DDX4-T2A-tdTomato reporter cell lines harboring a polyclonal integration of all three TFs, which we term DDX4 by 3 TFs or D3 combo, and assessed the induced DDX4+ yield compared to a no TF control in biological triplicate (Figure 2E). We find the D3 combo drives a statistically significant increase in DDX4-reporter cell yield (1.22% +/- 0.27% SEM). Remarkably, this high DDX4+ population is obtained in just 4 days in monolayer through direct TF induction during a differentiation protocol designed for hPGCLC formation.

NPM2 is a critical oocyte marker gene, involved in chromatin organization, and has been used in similar mouse studies to assess maturing oocyte-like phenotype formation.[18] We aimed to determine how this *NPM2* reporter in addition to our *DDX4* reporter was expressed in our system, if at all, since a recent study showed that two of our three TFs *SOHLH1* and *LHX8* drive NPM2+ oocyte-like formation in mouse models. We additionally sought to test higher order combinations of our D3 TFs with our NANOS3-driving TFs to determine if any synergistic combinations resulted in higher yields of putative germ cells. Further we wanted to assess the addition of RNA-binding proteins such as *DAZL*, *DDX4*, and *BOLL* which have been shown to regulate germ cell development, but were obviously not selected in our computational tool since they are not transcription factors. Overall, we generated 12 new combination lines that 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 3E). Our results demonstrate that addition of *FIGLA*, in particular, drives a significant increase in DDX4+ yield and significant increase in NPM2+ yield (D3F combination) compared to the D3 combination alone (Figure 2F). We also find that combined induction of all 9 factors, a condition we call DNR3 combo, induces robust, significant DDX4+ yield and NPM2+ yield compared to D3 alone (Figure 2F). We therefore conclude that overexpression of *ZNF281*, *SOHLH1*, and *LHX8* in combination during a hPGCLC differentiation protocol is a salient platform for DDX4+ cell formation with addition of *FIGLA, HHEX, DLX5, DAZL, BOLL*, and *DDX4* helping to drive higher yields.

TF overexpression drives formation of cells with profiles of human primordial germ cells and oogonia

We isolated our reporter positive germ cells from TF-overexpression conditions using fluorescence activated cell sorting (FACS) for subsequent RNA-seq analysis. We performed principal component analysis (PCA) of our D3, D3-FIGLA, and DNR3 DDX4+ and NPM2+ cells as well as our DLX5, HHEX, FIGLA, and no TF control NANOS3+ cells, with culture-derived hPGCLCs and oogonia and *in vivo* oocytes from the Gene Expression Omnibus (GEO) database. We find that our NANOS3+ cells readily cluster with primordial germ cell samples, while our DDX4+ and NPM2+ cells share principal components with oogonia samples and maturing oocytes, stopping short of oocyte maturation at the antral follicle stage (Figure 3A).

hPGCLCs are delineated by upregulation of key genes such as *CD38, NANOS3, SOX17, TFAP2C*, and *PRDM1* and suppression of *SOX2*.[6,7,10] Oogonia-like cells are delineated by expression of *DDX4, DAZL*, as well as other genes such as *ID1/2/3/4, MSX1/2, XIST and MAEL*.[14] To further elucidate the transcriptomic profiles of our derived cell types, we examined the expression of key marker genes in different developmental stages of germ cell maturation. Our results confirm that our monolayer-derived no TF control, *DLX5*, *HHEX*, and *FIGLA* overexpression NANOS3+ cells exhibit gene expression characteristic of standard PGCLC samples, such as upregulation of *SOX17*, *TFAP2C*, and *PRDM1* and downregulation of *SOX2* though DLX5 and FIGLA-driven samples do express key genes at further developmental timepoints (Figure 3B). Additionally, D3, D3-FIGLA, DNR3 DDX4+ samples clearly demonstrate oogonia-specific gene expression and have broad transcriptomic similarity to fetal oogonia via the recently-described transcriptome overlap metric (TROM), expressing *DDX4, ID4, XIST, SYCP1, WEE2, SIX1, MSX2, TFAP2C* and *PRDM1* while *DAZL* was slightly upregulated but not significantly (Figure 3B, Figure S5).[29] Thus, through multiple analysis strategies, our DDX4+ cells consistently resemble oogonia from all compared sample types, leading us to label these samples as induced oogonia-like cells (iOLCs).

We furthermore performed immunofluorescence staining for SOX17, OCT4, and ITGA6 to demonstrate that our monolayer control and TF-driven hPGCLCs exhibit nominal protein expression hallmarks of conventional hPGCLCs. We find our NANOS3-T2A-mVenus+ sorted hPGCLCs exhibit triple positive expression of SOX17, ITGA6, and OCT4, compared to wild type hiPSCs which are dual positive for OCT4 and ITGA6 and negative for SOX17 (Figure 4A).[8] We likewise purified our iOLCs from the D3F combination and performed immunofluorescence staining for endogenous DDX4 and DAZL. Our D3-FIGLA DDX4+ cells exhibit DDX4 protein localization in punctate patterns in the cytoplasm; however, little to no expression of DAZL is noted, confirming our transcriptome findings (Figure 4B). 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.

It is known that *in vivo*, primordial germ cells and oogonia develop and mitotically expand, establishing the human ovarian reserve.[14,30] Recent methods have been developed for expansion of hPGCLCs *in vitro* post-isolation.[5,7,31] No method yet exists for the maintenance 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 the iOLC population (Figure 4C).[7] iOLCs are maintained in culture for over 30 days, retaining heterogeneous expression of DDX4, as can be seen by live cell fluorescent imaging of the tdTomato reporter protein. These iOLCs retain robust expression of the DDX4-T2A-tdTomato reporter, with clear visualization of the cleaved, diffuse fluorescent protein. Since our reporter is not a fusion protein to DDX4 itself, we see the expected pattern of fluorescent protein in both the nucleus and cytoplasm as opposed to just the cytoplasm. We therefore conclude that our TF-derived iOLCs are capable of feeder-free maintenance of DDX4+ identity post-isolation.

Discussion

In this study, we identified 36 TFs that may play regulatory roles within the underlying GRNs of gametogenesis and further screened these TFs alongside known control TFs and RNA binding proteins via combinatorial and individual overexpression in germline-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-like cell formation.

We demonstrate that DLX5, HHEX, and FIGLA drive on-target NANOS3+ hPGCLC formation, shown in the expression of the core primordial germ cell genes *SOX17*, *TFAP2C*, and *PRDM1* and proteins in our isolated TF-based NANOS3+ cells. These TFs enhance 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 partially rescues hPGCLC yield in the absence of BMP4. From our study, it is clear this rescue does not result from DLX5 upregulating endogenous *BMP2* or *BMP4* expression but it is not known what the precise mechanism of action is. Interestingly, we find that while combination overexpression of these TFs improves hPGCLC yield relative to the no TF control, it performs worse than individual overexpression, showing these TFs may not act synergistically in enhancing hPGCLC formation. Based on the consistent performance of the TFs' overexpression across differentiation platforms and isolation techniques, we believe these new tools provide a widely applicable, readily usable method for drastically enhancing hPGCLC yield.

More studies will be needed to determine if these three TFs perform germ cell regulating roles *in vivo* in addition to serving as useful cell engineering tools in our described context. 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*.[22] 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 robustly in combination, of *SOHLH1*, *LHX8*, and *ZNF281* (D3) drives formation of DDX4+ cell formation in just four days. We identify higher order combinations, such as D3-FIGLA, and a 9 gene combo termed DNR3, that generate DDX4+ and NPM2+ cells. Our DDX4+ cell formation is rapid compared to existing methods which generally require 70 to 120 days in xrOvary based methods compared to our four day protocol.[14] In addition, our DDX4+ cells are generated 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 demonstrate maintenance feeder-free post-isolation of our DDX4+ cells using media developed for hPGCLC maintenance, potentially expanding the utility of this cell type for long term studies.[7] Additional studies will be needed to determine the *in vivo* role of these TFs in oogonia and oocyte formation and whether our overexpression-based method for generating these cell types recapitulates the endogenous genetic regulation of oogenesis. As *ZNF281*, *LHX8*, and *SOHLH1* are highly differentially expressed during human folliculogenesis, linked as causal genetic determinants of infertility when mutated, and in the case of, *lhx8* and *sohlh1*, found in mouse models to induce oocyte-like formation from stem cells, it is highly likely they play critical roles during *in vivo* oogenesis. [18,21,32]

Via transcriptomic analysis, we find that our TFs drive formation of DDX4+ and NPM2+ cells that exhibit upregulation of key oogonia and oocyte genes, which we designate as induced oogonia-like cells (iOLCs). We find our iOLCs broadly share similarity with *in vivo* maturing oocytes and fetal oogonia.[21,22] However, our iOLCs are still lacking robust expression of 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. 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. Additional research as well combining our iOLC methods with co-culture methods such as xrOvaries or newer all human methods such as those described recently are needed to determine if these DDX4+ cells are capable of meiotic entry or oocyte maturation.[33] Nonetheless, these four day, directly-generated human germ cell types provide a novel, reproducible, and simple-to-obtain source of important germline intermediates that can be used in reproductive development and genetic screening studies as models of inaccessible *in vivo* counterparts. In conclusion, we identify novel TFs in driving germ cell formation, establishing a rapid, high-throughput platform for *in vitro* human gametogenesis and reproductive modeling in a simple 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.[34] Gene regulatory networks were inferred utilizing the GRNBoost2 algorithm in the Arboreto computational framework.[23,34] 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. Gene Ontology Analysis was performed by input of the prioritized 46 TF screening list to the g:profiler tool and selecting the significantly enriched BP terms. **g:GOSt** performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. These significant GO terms, selected by the adjusted p-value, were then input to the visualizer tool Revigo and plotted as a scatterplot. The axes in the plot have no intrinsic meaning. Revigo uses Multidimensional Scaling (MDS) to reduce the dimensionality of a matrix of the GO terms pairwise semantic similarities. The resulting projection may be highly non linear. The guiding principle is that semantically similar GO terms should remain close together in the plot.

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 Irie et al. 2015, 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 S6). 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, 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, D3F and DNR3 bulk RNA-Seq, TFs were induced for four days and cells were sorted for DDX4+ or NPM2+ expression and utilized for bulk RNA-Seq. For reference samples of the ovarian atlas, data was downloaded from GEO and realigned using our pipeline.

RNA-Seq Analysis

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] Reads per kilobase of transcript per million (RPKM) normalization was conducted for each sample and plotted for state-specific genes with the ComplexHeatmap library in R. Principal component analysis (PCA) was conducted on the normalized matrix in scikit-learn and plotted with the Plotly Express package in Python. The normalized count matrix is provided in Table S3.

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.[28] 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.

Floating aggregate hPGCLC induction and harvesting

iPSCs were differentiated to PGCLCs according to the method of Irie *et al*. Briefly, iPSCs were grown in 4i medium (KnockOut DMEM with 20% KSR, 1X non-essential amino acids, 2 mM L-glutamine, 0.1 mM 2-mercaptoethanol, 20 ng/mL LIF, 8 ng/mL FGF2, 1 ng/mL TGFβ, 3 µL CHIR99021, 1 µM PD0325901, 5 µM SB203580, and 5 µM SP600125) until 70-80% confluency, then harvested using TRYPLE. iPSCs were seeded in 96-well U-bottom low-attachment plates (Corning #7007). Each well contained 4000 iPSCs in 200 µL of PGCLC induction medium (Advanced RPMI with 1X non-essential amino acids, 0.5X B27 supplement, 2 mM L-glutamine, 0.25% polyvinyl alcohol, 10 µM Y-27632, 500 ng/mL BMP2, 100 ng/mL SCF, and 50 ng/mL EGF). The plates were centrifuged (300 *g*, 2 min.) and incubated (37 °C, 5% CO_2) for four days. For PGCLC harvesting and staining, for each sample, nine aggregates were combined in a 1.5 mL tube. The aggregates were washed with PBS, resuspended in trypsin solution (45 µL), and incubated in a shaking heat block (37 °C, 800 rpm). After 8 minutes, 940 µL FACS buffer (3% FBS in PBS) was added, and the aggregates were dissociated by vigorous pipetting. The suspension was passed through a 70 µm strainer and spun down (300 *g*, 3 min). The supernatant was removed, and the pellet was resuspended in antibody solution (40 µL). The suspension was kept on ice in the dark for 30 minutes. Then, 940 µL FACS buffer was added. The cells were spun down again and resuspended in 200 µL FACS buffer with 100 ng/mL DAPI, and analyzed on a BD LSR Fortessa flow cytometer as described above.

iOLC maintenance culture

Feeder-free maintenance 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 2B, n=3 replicates were independently seeded for induction from a single hiPSC cell line. For Figure 2C, n=2 replicates were independently seeded for induction from a single hiPSC cell line. Figure 2D, n=3 replicates were independently seeded for induction from a single hiPSC cell line. Figure 2E, n=6 replicates were independently seeded for induction from three independent hiPSC cell lines, two per line. Results from each cell line were averaged for the duplicates and graphed as biological replicates. For Figure 2F, n=2 replicates were independently seeded for induction from a single hiPSC cell line. For Figure 3A-B, between n= 2 to 3 replicates were independently seeded for induction and sequencing from different hiPSC lines. For Figure 4A-C, n=2 replicates were seeded for imaging from a single hiPSC cell line and images are representative of overall population.

Supplementary Figures

Figure S1: Development of germ cell fluorescent reporter hiPSC lines
Figure S2: Development of TF overexpression screening system for hiPSCs
Figure S3: DLX5 overexpression partially rescues hPGCLC yield in the absence of BMP4 without upregulation of endogenous BMP2 or BMP4 expression
Figure S4: DLX5, HHEX, and FIGLA drive reporter and surface marker positive hPGCLC formation in floating aggregate and monolayer culture formats
Figure S5: TF-induced DDX4+ germ cells significantly resemble primary fetal oogonia
Figure S6: Representative flow cytometry analysis

Supplementary Table 1: PageRank Ranking of all TFs
Supplementary Table 2: TFs utilized for screening
Supplementary Table 3: Normalized gene counts of all sequenced samples

Supplementary File 2: Monolayer germ cell screening reagents and methods

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. assisted with library preparation and analysis. P.R.F. aided in confocal imaging and library preparation. P.C., G.B., T.L.Z., S.V, K.P., and S.B. conducted bulk RNA-seq analysis. E.D., J.A., and S.K. assisted in vector cloning, TF screening, and confocal imaging. S.L. aided in 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, scientific advisor to Gameto, Inc. C.K. is currently the VP of Cell Engineering of Gameto, Inc. G.M.C. serves on the scientific advisory board of Gameto, Inc., Colossal Biosciences, and GCTx.

Acknowledgements

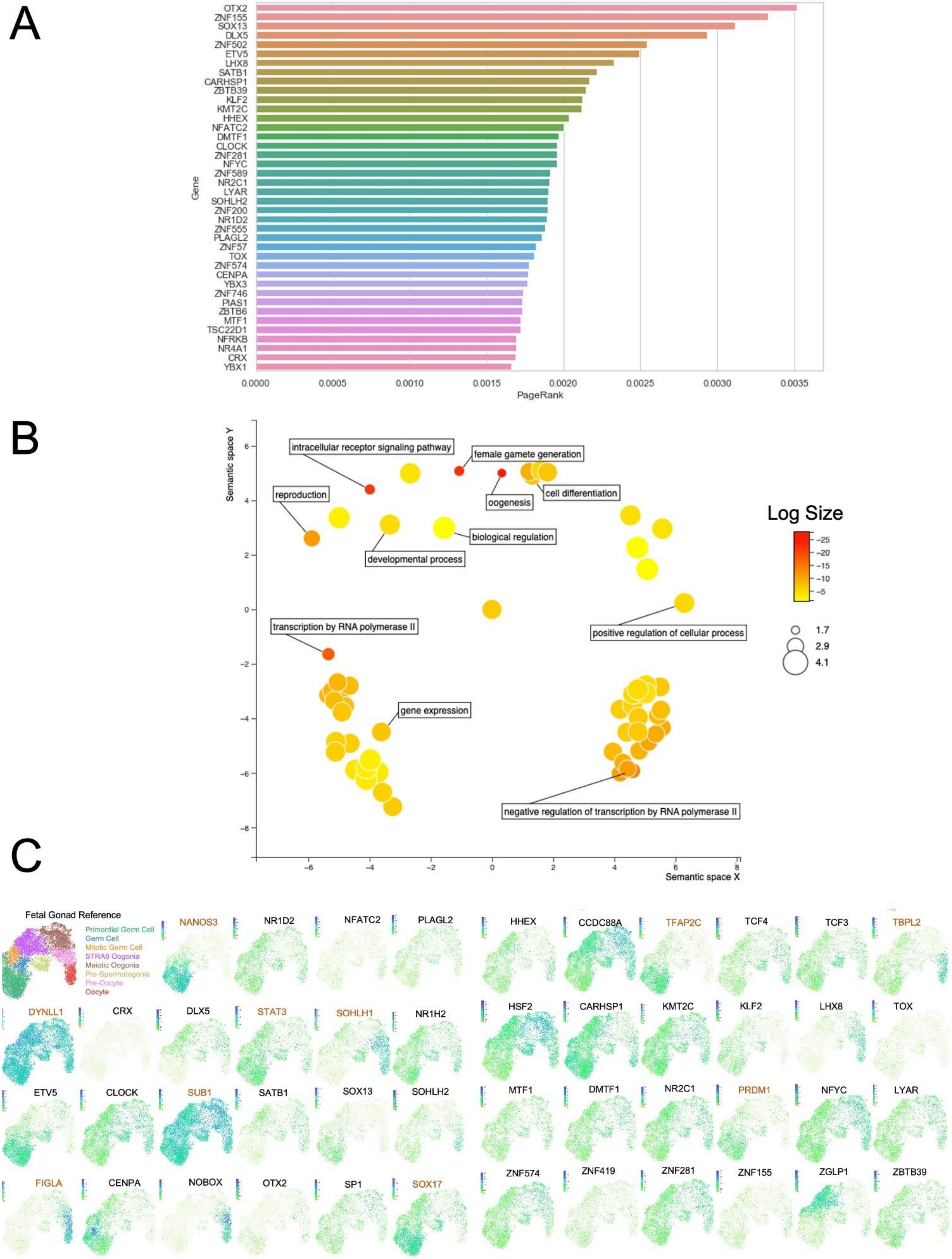
We thank Emma Tysinger and Teodora Stan for technical assistance in the research. This work was funded by the Synthetic Biology Platform at the Wyss Institute for Biologically Inspired Engineering and institutional startup funds at Duke University. This work was additionally funded by the Gameto Sponsored 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. Kobayashi, T. *et al.* Principles of early human development and germ cell program from conserved model systems. *Nature* **546**, 416–420 (2017).
2. Saitou, M. & Hayashi, K. Mammalian in vitro gametogenesis. *Science* **374**, eaaz6830 (2021).
3. Nagaoka, S. I., Saitou, M. & Kurimoto, K. Reconstituting oogenesis in vitro: Recent progress and future prospects. *Current Opinion in Endocrine and Metabolic Research* **18**, 145–151 (2021).
4. NSFG - listing I - key Statistics from the National Survey of family growth. https://www.cdc.gov/nchs/nsfg/key_statistics/i.htm (2019).
5. 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).
6. Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
7. 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).

8. 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).
9. 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.
10. Kojima, Y. *et al.* GATA transcription factors, SOX17 and TFAP2C, drive the human germ-cell specification program. *Life Science Alliance* vol. 4 e202000974 Preprint at https://doi.org/10.26508/lsa.202000974 (2021).
11. 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).
12. Yokobayashi, S. *et al.* Clonal variation of human induced pluripotent stem cells for induction into the germ cell fate†. *Biology of Reproduction* vol. 96 1154–1166 Preprint at https://doi.org/10.1093/biolre/iox038 (2017).
13. Chen, D. *et al.* Germline competency of human embryonic stem cells depends on eomesodermin. *Biol. Reprod.* **97**, 850–861 (2017).
14. Yamashiro, C. *et al.* Generation of human oogonia from induced pluripotent stem cells in vitro. *Science* **362**, 356–360 (2018).
15. Hikabe, O. *et al.* Reconstitution in vitro of the entire cycle of the mouse female germ line. *Nature* vol. 539 299–303 Preprint at https://doi.org/10.1038/nature20104 (2016).
16. Yoshino, T. *et al.* Generation of ovarian follicles from mouse pluripotent stem cells. *Science* **373**, (2021).
17. Nagaoka, S. I. *et al.* ZGLP1 is a determinant for the oogenic fate in mice. *Science* **367**, (2020).
18. Hamazaki, N. *et al.* Reconstitution of the oocyte transcriptional network with transcription factors. *Nature* **589**, 264–269 (2021).
19. Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178–194 (2015).
20. Kramme, C. *et al.* An integrated pipeline for mammalian genetic screening. *Cell Rep Methods* **1**, 100082 (2021).
21. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Mol. Cell* **72**, 1021–1034.e4 (2018).
22. 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 Preprint at https://doi.org/10.1016/j.stem.2017.03.007 (2017).
23. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
24. Zhang, J. *et al.* OTX2 restricts entry to the mouse germline. *Nature* vol. 562 595–599 Preprint at https://doi.org/10.1038/s41586-018-0581-5 (2018).
25. Garcia-Alonso, L. *et al.* Single-cell roadmap of human gonadal development. *Nature* **607**, 540–547 (2022).
26. Kramme, C. *et al.* MegaGate: A toxin-less gateway molecular cloning tool. *STAR Protoc* **2**, 100907 (2021).
27. 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 Preprint at https://doi.org/10.1073/pnas.1008322108 (2011).
28. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471 (2013).
29. 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).
30. Wallace, W. H. B. & Kelsey, T. W. Human Ovarian Reserve from Conception to the Menopause. *PLoS ONE* vol. 5 e8772 Preprint at https://doi.org/10.1371/journal.pone.0008772 (2010).
31. Murase, Y. *et al.* Long-term expansion with germline potential of human primordial germ cell-like cells in vitro. *EMBO J.* **39**, e104929 (2020).

32. Pangas, S. A. *et al.* Oogenesis requires germ cell-specific transcriptional regulators *Sohlh1* and *Lhx8*. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8090–8095 (2006).
33. Pierson Smela, M. D. *et al.* Directed differentiation of human iPSCs to functional ovarian granulosa-like cells via transcription factor overexpression. *Elife* **12**, e83291 (2023).
34. 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).



A
OTX2
ZNF155
SOX13
DLX5
ZNF502
ETV5
LHX8
SATB1
CARHSP1
ZBTB39
KLF2
KMT2C
HHEX
NFATC2
DMTF1
CLOCK
ZNF281
NFYC
ZNF589
NR2C1
LYAR
SOHLH2
ZNF200
NR1D2
ZNF555
PLAGL2
ZNF57
TOX
ZNF574
CENPA
YBX3
ZNF746
PIAS1
ZBTB6
MTF1
TSC22D1
NFRKB
NR4A1
CRX
YBX1
Gene
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
PageRank
B
Semantic space Y
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6
8
Semantic space X
intracellular receptor signaling pathway
female gamete generation
reproduction
oogenesis
cell differentiation
biological regulation
developmental process
positive regulation of cellular process
transcription by RNA polymerase II
gene expression
negative regulation of transcription by RNA polymerase II
Log Size
−25
−20
−15
−10
−5
1.7
2.9
4.1
C
Fetal Gonad Reference
Primordial Germ Cell
Germ Cell
Mitotic Germ Cell
STRA8 Oogonia
Meiotic Oogonia
Pre-Spermatogonia
Pre-Oocyte
Oocyte
NANOS3
NR1D2
NFATC2
PLAGL2
HHEX
CCDC88A
TFAP2C
TCF4
TCF3
TBPL2
DYNLL1
CRX
DLX5
STAT3
SOHLH1
NR1H2
HSF2
CARHSP1
KMT2C
KLF2
LHX8
TOX
ETV5
CLOCK
SUB1
SATB1
SOX13
SOHLH2
MTF1
DMTF1
NR2C1
PRDM1
NFYC
LYAR
FIGLA
CENPA
NOBOX
OTX2
SP1
SOX17
ZNF574
ZNF419
ZNF281
ZNF155
ZGLP1
ZBTB39

Figure 1. *In silico* identification of TFs regulating germ cell formation A) Inferred gene regulatory network generated from folliculogenesis dataset of Zhang et al. 2018. with prioritized TF screening list, ordered by PageRank score imputed from the gene regulatory network and centrality algorithm. B) All 46 TFs utilized in the study were analyzed using the GO term tool g:profiler, and significant GO terms were plotted in Revigo visualizer highlighted terms related to significantly enriched pathways for germline development and transcription factor function. Color and dot size reflect log size of the p-adjusted value. C) The expression of all 46 TFs was visualized using the reproductive cell atlas, utilizing only germline cells obtained from the fetal atlas of Garcia-Alonso et al. 2022. Expression level for each TF is mapped onto the UMAP, with the cell type annotation shown in the reference map. TFs included as controls in the study are highlighted orange.

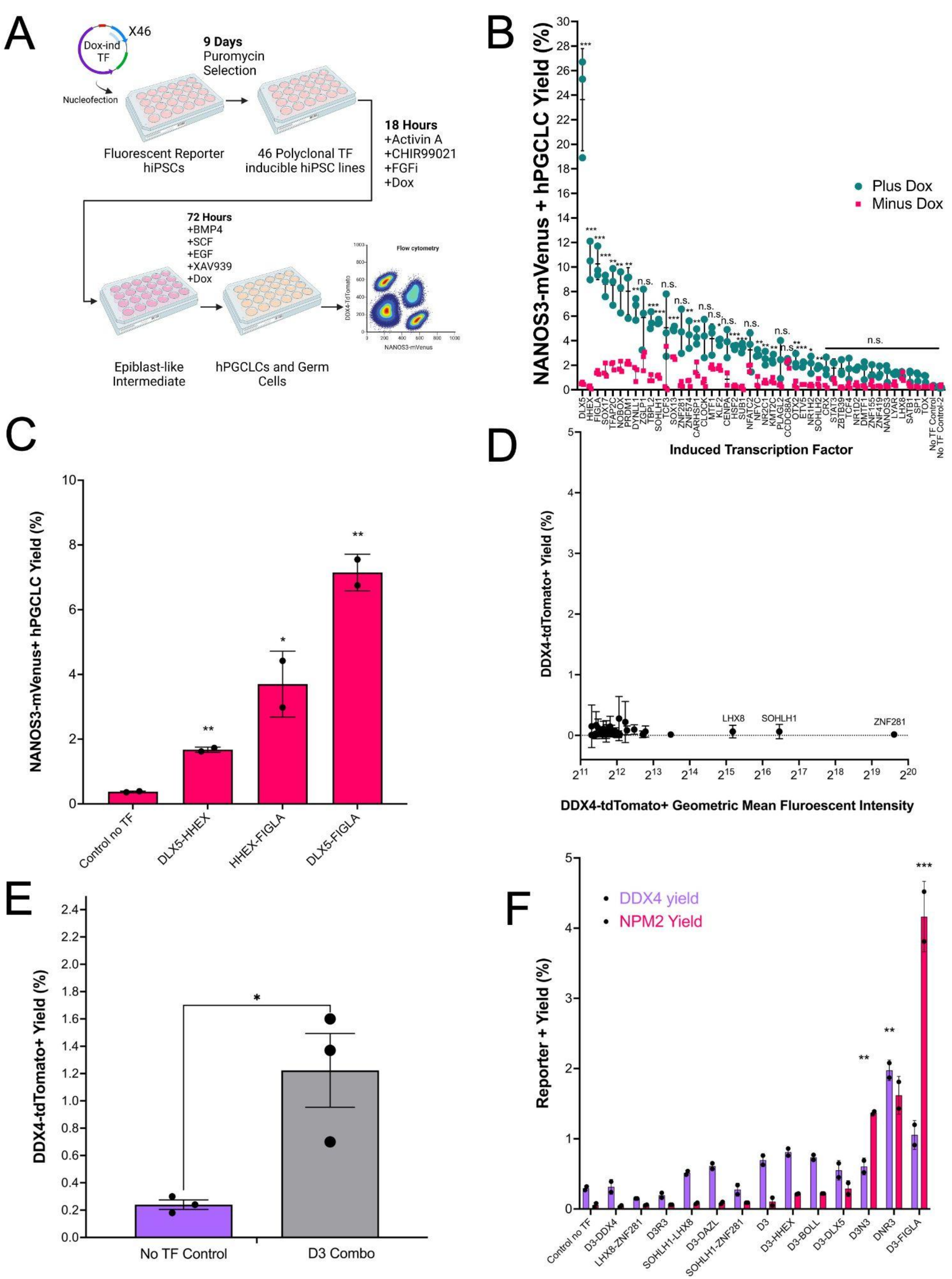

A
Dox-ind TF
X46
Nucleofection
9 Days
Puromycin
Selection
Fluorescent Reporter
hiPSCs
46 Polyclonal TF
inducible hiPSC lines
18 Hours
+Activin A
+CHIR99021
+FGFi
+Dox
72 Hours
+BMP4
+SCF
+EGF
+XAV939
+Dox
Flow cytometry
DDX4-tdTomato
NANOS3-mVenus
Epiblast-like
Intermediate
hPGCLCs and Germ
Cells
B
NANOS3-mVenus + hPGCLC Yield (%)
Plus Dox
Minus Dox
n.s.
Induced Transcription Factor
DLX5
HHEX
FIGLA
SOX17
TFAP2C
NOBOX
PRDM1
DYNLL1
ZGLP1
TBPL2
SOHLH1
TCF3
SOX13
ZNF281
ZNF574
CARHSP1
CLOCK
MTF1
KLF2
CENPA
HSF2
SUB1
NFATC2
NFYC
NR2C1
KMT2C
PLAGL2
CCDC88A
OTX2
ETV5
NR1H2
SOHLH2
CRX
STAT3
ZBTB39
TCF4
NR1D2
DMTF1
ZNF155
ZNF419
NANOS3
LYAR
LHX8
SATB1
SP1
TOX
No TF Control
No TF Control-2
C
NANOS3-mVenus+ hPGCLC Yield (%)
Control no TF
DLX5-HHEX
HHEX-FIGLA
DLX5-FIGLA
D
DDX4-tdTomato+ Yield (%)
LHX8
SOHLH1
ZNF281
DDX4-tdTomato+ Geometric Mean Fluroescent Intensity
E
DDX4-tdTomato+ Yield (%)
No TF Control
D3 Combo
F
DDX4 yield
NPM2 Yield
Reporter + Yield (%)
Control no TF
D3-DDX4
LHX8-ZNF281
D3R3
SOHLH1-LHX8
D3-DAZL
SOHLH1-ZNF281
D3
D3-HHEX
D3-BOLL
D3-DLX5
D3N3
DNR3
D3-FIGLA

Figure 2. Screening for TFs regulating germ cell development. A) Schematic workflow for TF overexpression screening, with time intervals and media supplements. B) Results for NANOS3-T2A-mVenus flow cytometry for triplicate induction conditions in the plus (blue) and minus (pink) doxycycline induction condition for each TF. Data is plotted as a percent of NANOS3-T2A-mVenus+ cells on the y-axis for each TF. Individual dots represent individual wells of the induction condition, seeded separately from a single hiPSC line. Horizontal black line represents the mean of induction replicates. Statistical significance was determined by multiple T-test comparison between the plus and minus dox condition for each TF, with a p-value <0.05 considered as significant. The FDR correction was utilized for multiple hypothesis testing. *** p<0.001, ** P<-.01, * p<0.05. C) NANOS3-T2A-mVenus+ flow cytometry results are visualized for combinatorial TF overexpression of DLX5, HHEX, and FIGLA and the no TF control. Data is graphed as a mean +/- SEM. Individual dots represent individual wells of the induction condition, seeded separately in duplicate from a single hiPSC line. Statistical significance was determined by multiple T-test comparison between each combination and the no TF control, with a p-value <0.05 considered as significant. The FDR correction was utilized for multiple hypothesis testing. *** p<0.001, ** P<-.01, * p<0.05. D) Flow Cytometry results of the screen for the DDX4-T2A-tdTomato reporter are visualized with percent reporter positive on the y-axis, and geometric mean fluorescence intensity of the tdTomato+ cells on the y-axis. Error bars represent the SEM of the induction yield between the triplicates in the plus dox condition. E) Flow cytometry results are visualized for DDX4-T2A-tdTomato reporter yield (percent) for the no TF control and the D3 combo in n=6 replicates (3 independent hiPSC lines, with 2 replicates per line). Individual dots represent the average for the technical replicates for each biological replicate. Data is presented as a mean +/- SEM. Statistical significance was determined by two-sided T-test comparison between each combination and the no TF control, with a p-value <0.05 considered as significant. *** p<0.001, ** P<-.01, * p<0.05. F) Flow cytometry results are visualized for DDX4-T2A-tdTomato (purple) and NPM2-T2A-mGreenLantern (pink) reporter yield (percent) for various combinations of TF inductions. Individual dots represent individual wells of the induction condition, seeded in duplicate from a single hiPSC line. Statistical significance was determined by multiple T-test comparison between each combination and the no TF control, with a p-value <0.05 considered as significant. The FDR correction was utilized for multiple hypothesis testing. *** p<0.001, ** P<-.01, * p<0.05.

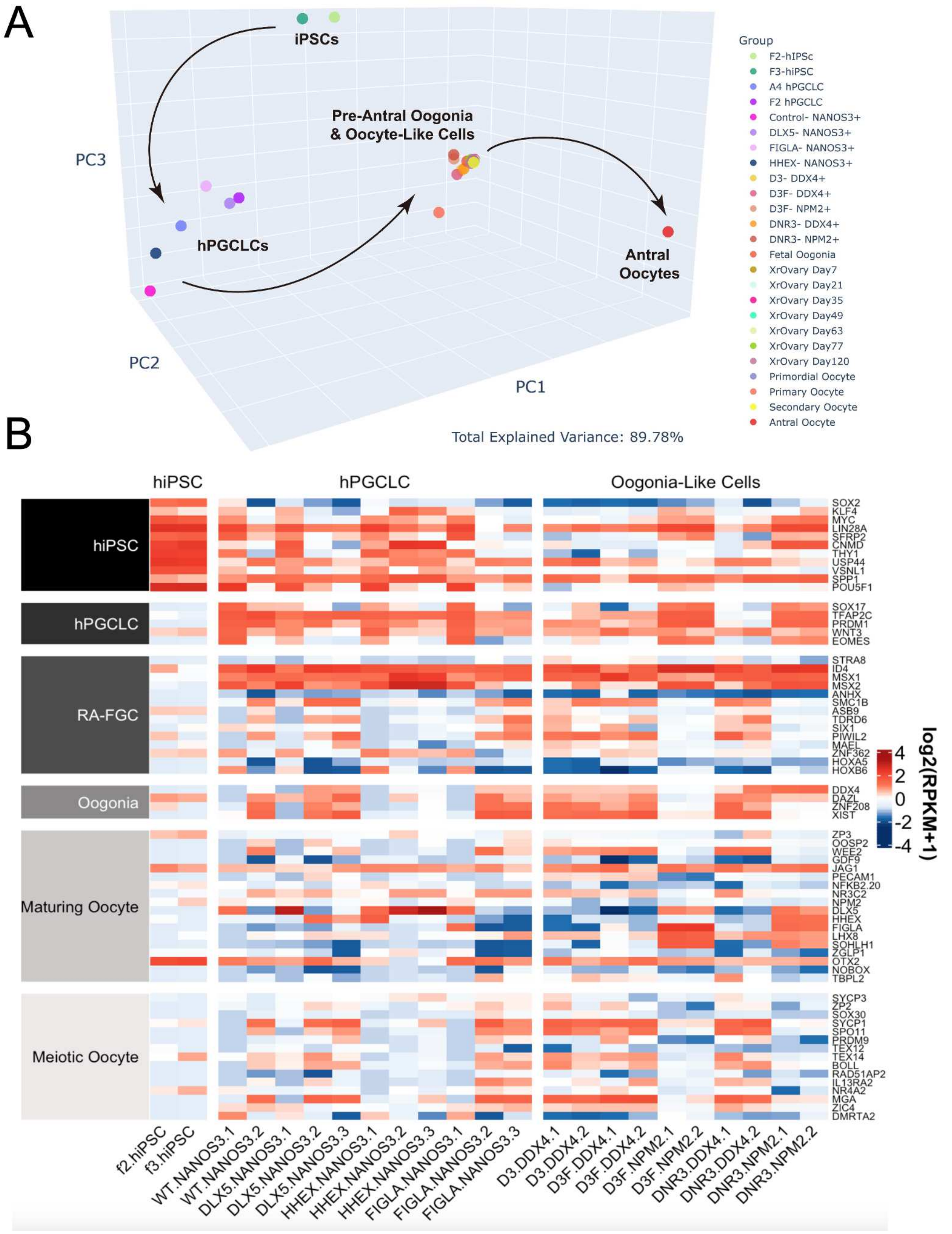

A
iPSCs
Pre-Antral Oogonia & Oocyte-Like Cells
hPGCLCs
Antral Oocytes
PC3
PC2
PC1
Total Explained Variance: 89.78%
Group
F2-hIPSc
F3-hiPSC
A4 hPGCLC
F2 hPGCLC
Control- NANOS3+
DLX5- NANOS3+
FIGLA- NANOS3+
HHEX- NANOS3+
D3- DDX4+
D3F- DDX4+
D3F- NPM2+
DNR3- DDX4+
DNR3- NPM2+
Fetal Oogonia
XrOvary Day7
XrOvary Day21
XrOvary Day35
XrOvary Day49
XrOvary Day63
XrOvary Day77
XrOvary Day120
Primordial Oocyte
Primary Oocyte
Secondary Oocyte
Antral Oocyte
B
hiPSC
hPGCLC
Oogonia-Like Cells
hiPSC
hPGCLC
RA-FGC
Oogonia
Maturing Oocyte
Meiotic Oocyte
SOX2
KLF4
MYC
LIN28A
SFRP2
CNMD
THY1
USP44
VSNL1
SPP1
POU5F1
SOX17
TFAP2C
PRDM1
WNT3
EOMES
STRA8
ID4
MSX1
MSX2
ANHX
SMC1B
ASB9
TDRD6
SIX1
PIWIL2
MAEL
ZNF362
HOXA5
HOXB6
DDX4
DAZL
ZNF208
XIST
ZP3
OOSP2
WEE2
GDF9
JAG1
PECAM1
NFKB2.20
NR3C2
NPM2
DLX5
HHEX
FIGLA
LHX8
SOHLH1
ZGLP1
OTX2
NOBOX
TBPL2
SYCP3
ZP2
SOX30
SYCP1
SPO11
PRDM9
TEX12
TEX14
BOLL
RAD51AP2
IL13RA2
NR4A2
MGA
ZIC4
DMRTA2
log2(RPKM+1)
4
2
0
-2
-4
f2.hiPSC
f3.hiPSC
WT.NANOS3.1
WT.NANOS3.2
DLX5.NANOS3.1
DLX5.NANOS3.2
DLX5.NANOS3.3
HHEX.NANOS3.1
HHEX.NANOS3.2
HHEX.NANOS3.3
FIGLA.NANOS3.1
FIGLA.NANOS3.2
FIGLA.NANOS3.3
D3.DDX4.1
D3.DDX4.2
D3F.DDX4.1
D3F.DDX4.2
D3F.NPM2.1
D3F.NPM2.2
DNR3.DDX4.1
DNR3.DDX4.2
DNR3.NPM2.1
DNR3.NPM2.2

Figure 3. TF overexpression drives germ cell formation with transcriptomic similarity to *in vivo* and *in vitro* controls. A) NANOS3+, DDX4+, and NPM2+ cells were FACS 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 adjusted via $\log_2$(RPKM+1) normalization. Results are represented in 3D with the top three principal components. B) Samples of reporter-positive sorted germ cells used in the PCA of (A) are visualized for specific marker gene expression. Results are plotted as $\log_2$(RPKM+1). Individual marker genes were selected from related studies and used to plot expression across multiple germ cell stages.

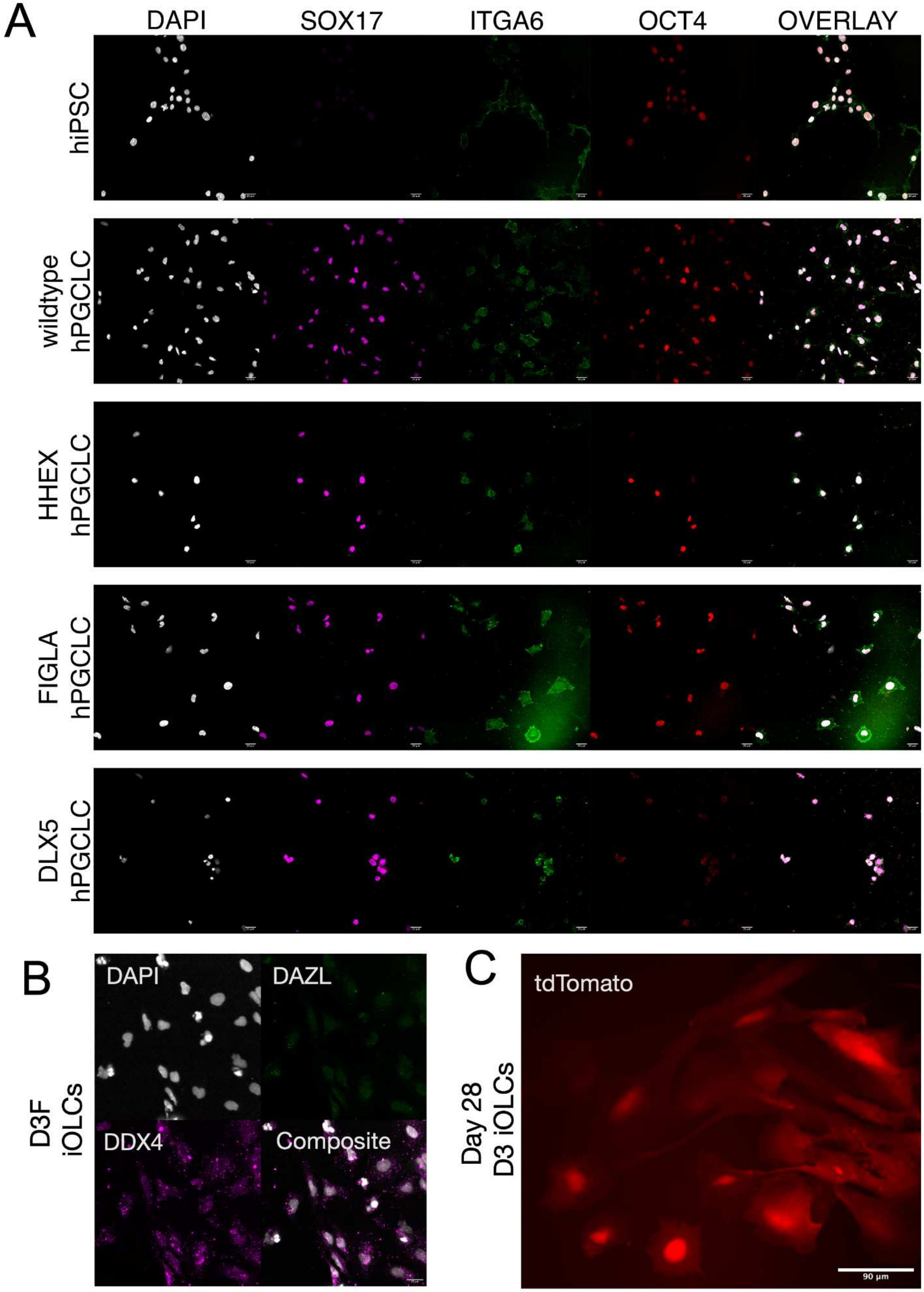

A
DAPI
SOX17
ITGA6
OCT4
OVERLAY
hiPSC
wildtype hPGCLC
HHEX hPGCLC
FIGLA hPGCLC
DLX5 hPGCLC
B
D3F iOLCs
DAPI
DAZL
DDX4
Composite
C
Day 28 D3 iOLCs
tdTomato
90 µm

Figure 4. TF overexpression drives germ cell formation with nominal expression of key germ cell proteins

A) NANOS3+ cells from the no TF, DLX5, HHEX, and FIGLA over expressions were FACS isolated at day 4 of differentiation and stained alongside a wild type hiPSC control for DAPI, SOX17 (germ cell), OCT4 (germ cell/pluripotent), and ITGA6 (germ cell/pluripotent). Representative images were taken using confocal microscopy. Scale bars are 25 µm in all images B) DDX4+ cells from D3-FIGLA combination expression were FACS isolated at day 4 of differentiation and stained for DAPI, DAZL (mature germ cell), and DDX4 (mature germ cell). Representative images were taken using confocal microscopy. Scale bar is 25µm. C) Live cell fluorescent imaging is performed on FACS isolated D3 combination-derived iOLCs grown feeder-free on matrigel coated plates at day 28 post-FACS isolation. tdTomato fluorescent protein derived from the cleaved DDX4-T2A-tdTomato reporter is visualized using the Echo Revolve fluorescent microscope on the Texas Red setting. Scale bar is 90µm.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- krammeetaleggsupplement.pdf
- SupplementaryTable1PageRankofallTFs.csv
- SupplementaryTable2TFsutilizedforscreening.xlsx
- SupplementaryTable3NormalizedCountMatrix.csv