Title: A pre-menopausal single-cell atlas for ovarian drug discovery

Authors: Hannah VanBenschoten[1,+], Zhanel Nugmanova[1,+], Audrey Olmsted[1], Ian Gingerich[1,3], Caroline Kratka[4], Ruixu Huang[1], Francesca E. Duncan[4], Olha Kholod[1,*], Brittany A. Goods[1,2,3*]

[+] equal contribution
* equal contribution

Affiliations:
[1] Thayer School of Engineering, Dartmouth College, Hanover NH 03755
[2] Department of Molecular and Systems Biology, Dartmouth College, Hanover NH 03755
[3] Program in Quantitative Biomedical Sciences, Dartmouth College, Hanover NH 03755
[4] Department of Obstetrics and Gynecology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

Corresponding Author: Brittany A. Goods (britt.anne.goods@dartmouth.edu) or Olha Kholod (olha.kholod@dartmouth.edu).

ABSTRACT

Single-cell atlas efforts have reshaped the way we understand cells across the human body. Despite their power, they have not been effectively used to study ovarian biology in the context of other tissues, nor have they comprehensively incorporated healthy tissues from pre-menopausal donors. A focused pre-menopausal single-cell atlas could both advance our understanding of this life stage and support identification of ovarian gene targets for indications prevalent among younger demographics, such as fertility management. Here, we present an integrated resource of single-cell datasets from pre-menopausal women (PreMeno Atlas), comprising 511,365 cells from 14 tissues, including the ovary. This unified resource enables transcriptomic comparisons across cell types, tissues, and organs within the pre-menopausal context. Our analysis revealed distinct ovarian cell gene signatures, with theca and stromal cells exhibiting the most unique transcriptional profiles among ovarian cell types. We further leveraged the PreMeno Atlas to prioritize ovary-specific genes with potential druggability, particularly G-protein coupled receptors (GPCRs), identifying GPR78, ADRB3, GPR20, and GPR101 as candidate targets. Finally, we assessed species homology of ovarian cell marker genes using a harmonized mouse ovulation and spatial transcriptomics dataset. Collectively, this work establishes the PreMeno Atlas as a resource for ovarian biology research and contraceptive target discovery.

INTRODUCTION

Effective and acceptable female contraception is essential for reproductive autonomy, family planning, and global public health.[1] While hormonal contraceptives, in particular the oral

contraceptive pill (OCP), are among the most widely used forms of birth control,[2] they are not suitable for everyone due to side effects, contraindications, and personal or cultural preferences.[3] As a result, there is a growing demand for non-hormonal alternatives that provide reliable contraception without systemic hormone exposure and dysregulation of hormonally-controlled processes.[4] Developing such methods requires identifying highly specific drug targets in reproductive tissues – specifically the ovaries – to ensure pharmaceutical efficacy while minimizing off-target effects.[5]

In target-based drug discovery, a protein's druggability depends on several key factors: tissue-specific expression to reduce the risk of systemic side effects, functional properties conducive to inhibition, sequence divergence from ubiquitously expressed paralogs, and validation from animal loss- and gain-of-function studies demonstrating its physiologic role.[6,7] Precise understanding of tissue-specific gene expression is therefore a crucial element of drug discovery efforts. Despite the central role the ovary plays in female fertility, including follicle development, oocyte maturation, ovulation, and hormone production, tractable drug targets in ovarian tissue, particularly for nonhormonal contraceptive purposes, are limited. One of the major challenges in ovarian drug discovery occurs when genes identified to be highly expressed in the ovary or uterus are also expressed in non-reproductive tissues, or alternatively when genes are exclusive to the reproductive tract, but their disruption causes lethal or subfertile phenotypes.[8,9] These issues underscore the need to rigorously characterize ovary-specific genes in humans relative to other tissues in the body to de-risk drug discovery efforts.

Current drug discovery and gene specificity efforts rely on tissue-profiling databases such as the Genotype-Tissue Expression Project (GTEx)[10] and Functional Annotation of Mammalian Genome (FANTOM5).[11] While valuable, these datasets predominately represent post-menopausal individuals, making them suboptimal for studying reproductive function and conditions related to fertility. Pre-menopausal tissue in the ovary is biologically distinct, characterized by active follicle development, cyclical hormonal changes, and dynamic gene expression patterns – critical factors for identifying relevant drug targets.[12–14] By contrast, post-menopausal reproductive tissue is characterized by atrophic changes, increased inflammation, cessation of follicular activity, and reduced hormone production and responsivity, which have been shown to impact gene expression patterns.[15–17]. Additionally, many datasets rely on bulk RNA-sequencing (RNA-seq), which averages gene expression changes across entire tissues, masking cell-type-specific signals and potentially obscuring gene expression signatures of key ovarian cell populations such as oocytes, granulosa cells, and theca cells.[18,19] The dynamic and heterogeneous nature of the ovary and reproductive tract necessitates a targeted data integration approach, as recent single-cell RNA seq (scRNA-seq) studies and Atlas efforts have highlighted the complex cellular composition of cycling reproductive tissues.[20–24]. scRNA-seq data can more rigorously and precisely index cell-type-specific gene expression signatures, providing rich datasets that can be used to mine specific gene signatures at various biological levels (e.g. cells, tissues, and organs). Several of these multi-tissue Atlas efforts, while extremely comprehensive with broad utility, do not include ovarian tissue,[25] or have not focused to date

specifically on pre-menopausal healthy tissues for the specific use-case of drug-target identification in the ovary in otherwise healthy donors.[26–28]

To address these limitations in the specific use-case of pre-menopausal healthy donors, we curated high-quality scRNA-seq datasets derived from healthy tissue in pre-menopausal women, providing a tailored and comprehensive resource for reproductive drug discovery. This Pre-Menopausal (PreMeno) Atlas is comprised of well-annotated datasets obtained from several sources, including the Human Cell Atlas (HCA)[27], Human Transcriptomics Cell Atlas (HTCA)[26], CZI CELLxGENE[29], Tabula Sapiens[25], and the Gene Expression Omnibus (GEO)[30]. The final PreMeno Atlas contains 511,365 single cells, representative of 113 cell types and 14 tissues. We show that there are clear marker genes that define cell types in the human ovary when compared to the rest of the body, including cumulus cells, oocytes, and ovarian stromal/theca cells. We then used this Atlas in several ways to highlight its utility for both research and drug discovery. First, we explore differences in stromal cells across the body, highlighting shared and unique pathways that are enriched in these cell types, inclusive of the ovary. We then show that this resource can be used to identify ovary-specific genes that can serve as possible therapeutic targets for contraceptive development or fertility enhancement. Finally, we integrate these genes with a curated mouse ovary Atlas to support cross-species validation. Taken together, our PreMeno Atlas provides a critical resource for identifying ovary-specific genes, greatly enhancing the capacity for drug target discovery using human datasets.

MATERIALS AND METHODS

Dataset curation:
We focused on mining scRNA-seq datasets from the Human Cell Atlas (HCA), CZ CELLxGENE, Tabula Sapiens, and the Gene Expression Omnibus (GEO). The CZ CELLxGENE was particularly valuable due to its extensive coverage of major organs and high-quality data formats, which included essential female reproductive organs. Our filtering strategy for mining datasets from these databases aimed to ensure that only high-quality datasets representing healthy, mature female tissues were included in the analysis. Examples and results of our filtering approach as of July 2024 can be seen in **Supplemental Figure 1 and Supplemental Table 1**. We selected datasets that included organs of interest from developmentally mature human female donors of age 18-50 (where accessible), and that used 10X Genomics workflows to generate whole-cell scRNA-seq. The latter was criteria were chosen to minimize technical batch variables in the dataset.

Metadata harmonization prior to integration:
We harmonized gene nomenclature across all datasets by converting to HGNC symbols using the biomaRt package in R.[31] We then manually performed metadata harmonization, ensuring consistency in metadata fields including experimental conditions, tissue origins, and donor information to ensure that original information from each individual study was maintained. We unified cell type and cell population annotations across datasets based on the Human Cell Atlas

Ontology (CL) tool.[32] This curation was designed to harmonize cell labels based on current HCA nomenclature and field standards presented in this comprehensive database. Detailed mappings of cell type and population harmonization are provided in **Supplemental Table 2, Supplemental Figure 1**. All original cell type mappings are also retained in the final object.

scRNA-seq quality control and processing of 10x sequencing data:
The Seurat pipeline (V5) was used to integrate each study into one final object.[33] For each dataset, cells were retained if they expressed more than 400 genes and had less than 30% mitochondrial and ribosomal gene content; genes were included if detected in more than three cells.. The DoubletFinder R package was applied to each individual sample to exclude doublets and ensure that only high-quality singlets were maintained in the final dataset.[34] Gene expression for each cell was then normalized with SCT in Seurat following standard protocols.[35]

Clustering of the major cell types in the scRNA-seq data was performed by integrating all of the samples together using canonical-correlation-analysis-based integration from the Seurat package (v5). The filtered gene matrices from each sample were normalized using the *NormalizeData* function. To identify highly variable genes, we used *FindVariableFeatures*. We further identified anchors using *FindIntegrationAnchors* to integrate all samples using following parameters: dims = 10, k.filter = 30, anchor.features = 3000 and k.score = 30, which were used for the *IntegrateData* function with dims=10. The integrated dataset was then used for downstream analysis, which included scaling and centering the data using *ScaleData* and finding the most significant principal components (PCs) using *RunPCA*. For visualization, the dimensionality was further reduced using the UMAP methods with Seurat function *RunUMAP* based on 10 PCs. The principal components that were used to calculate the UMAP embedding were the same as those used for clustering.

Identification of organ, cell population, and cell type marker genes:
We used the Wilcoxon rank-sum test via *FindAllMarkers* function to identify genes that were differentially expressed between each organ, cell population and cell type. We considered genes with an adjusted p-value ≤ 0.01 and a log fold change > 1 to be significantly differentially expressed. We selected the top 10 genes with the highest log fold change and lowest adjusted p-value for each cell type as potential marker genes, however, a full set of identified marker genes are presented in **Supplemental Data File 1**.

Pseudobulk analysis for identification of pairwise differentially expressed genes (DEGs):
To identify DEGs between cell types, we employed the DESeq2 R package.[36] We aggregated individual cells into pseudobulk populations for each cell type, enabling the comparison of gene expression profiles between distinct cell populations of interest in a targeted way.[37] Briefly, pseudobulk populations were generated by summing the transcript counts for each gene across all cells within a given cell type, thereby creating a single, representative expression profile for

each cell type. We then performed pairwise comparisons between ovarian cell types and reference cell types, with a false discovery rate (FDR)-adjusted p-value cutoff of ≤ 0.05 to determine significance.

Analysis of stromal cell types from integrated atlas:

To analyze stromal cell heterogeneity, we subsetted 19 stromal cell types from the integrated single-cell dataset. Following subsetting of the initial dataset, we re-applied the Seurat preprocessing steps – namely *NormalizeData*, *ScaleData*, and *RunPCA* – to ensure appropriate normalization, scaling, and dimensionality reduction within the selected subset. Dimensionality reduction was performed using UMAP to assess relationships among stromal populations across different organs. For each stromal cell type, we calculated the average gene expression profile and identified the top five marker genes based on differential expression analysis using *FindAllMarkers* in Seurat.[33] To determine gene markers uniquely associated with ovarian stromal cells, we compared marker gene sets across all stromal populations using set-based intersection analysis. Functional enrichment of ovarian-specific marker genes was performed using Gene Ontology (GO) analysis to infer associated biological processes. To assess the expression of these markers across all stromal cell types, we compiled a gene expression matrix and annotated associated pathways to facilitate interpretation of their functional relevance.

Analysis and integration of mouse ovarian datasets:

Two mouse datasets from prior work were integrated as described above.[22,24] The integrated dataset was then used for downstream analysis, which included scaling and centering the data using *ScaleData* and finding the most significant principal components (n = 10) using *RunPCA* in Seurat. Different resolution parameters for unsupervised clustering were then examined to determine the optimal number of clusters using clustree R package.[38] To annotate the major cell type of each single cell, *FindAllMarkers* was used to find differentially expressed genes in each cluster using the Wilcoxon rank-sum test statistical framework. The top 100 most significantly differentially expressed genes (ranked by average log-transformed fold change and adjusted $P \leq 0.05$) were then carefully reviewed using referenced literature.[24] Based evaluation of marker genes, the unsupervised clusters were subsequently re-annotated with harmonized cell type identities. Additionally, ovarian cell marker gene expression was plotted for the spatiotemporal mouse ovary dataset.[22]

Pathway and GO enrichment analysis:

For gene set enrichment analysis, genes were ranked based on differential expression results filtered for an adjusted $P \leq 0.05$ and ordered by average log-transformed fold change between the cell types of interest. The ranked gene lists were subsequently submitted to Metascape for pathway enrichment analysis.[39] Curated gene sets of KEGG, Biological Processes and Reactome were downloaded from the Molecular Signature Database (MSigDB, http://software.broadinstitute.org/gsea/msigdb/index.jsp) and were used to calculate enrichment

scores. Significantly enriched gene sets were identified with a Benjamini–Hochberg adjusted $P \leq 0.05$. For the identified cell states, we also selected the top 200 genes and performed GO and KEGG enrichment analysis using the clusterProfiler package.

Integration of ovarian marker genes with orthogonal datasets and databases:
We integrated orthogonal databases with ovarian marker genes identified by PreMeno Atlas to query proteomic, phenotypic, and druggability information. Tissue-level protein expression information was accessed from the Human Protein Atlas (HPA, Version 24.0, /www.proteinatlas.org). Data represents protein expression levels based on immunohistochemistry (IHC) staining for 76 cell types in 45 normal tissues and includes coverage of 15,302 genes. Expression level is characterized by HPA as low, medium, high, or not detected; for tissues with multiple cell types, the highest expression level of a given gene in any cell type within that tissue was taken to reflect the tissue-specific expression of that gene. Ovarian cell marker genes were queried in the Contraceptive Infertility Target DataBase[40] (CITDBase, www.citdbase.org) and categorical information from various phenotypic databases were extracted. Human infertility phenotypes were reported by ClinVar, Human Phenotype Ontology (HPO), and Monarch, whereby phenotypic designations are classified as male infertility (M), female infertility (F), infertility in both sexes (CTB), or infertility with sex not specified (I). Mouse infertility phenotypes were reported by Mouse Genome Informatics (MGI) and International Mouse Phenotyping Consortium (IMPC), whereby phenotypic designations are classified as female infertility (F), male infertility (M), infertile with unknown sex (I), female sub-fertile (SubF), male sub-fertile (SubM), sub-fertile with unknown sex (SubI), non-reproductive gene with a knockout but no infertility phenotype (NonRKO), and no knockout (NoKO). A comprehensive list of human G-protein coupled receptors (GPCRs) was retrieved from The International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) Guide to Pharmacology database (GtoPdb, Version 2024.3, www.guidetopharmacology.org). All genes included in PreMeno Atlas were filtered on the GPCR list and a Wilcoxon rank-sum test was run on a pseudo bulked atlas aggregated by organ, comparing ovary with and all other organs. Top GPCRs were sorted with an adjusted p-value cutoff of 0.1 and ranked by log-fold change, and row-normalized expression was plotted for the top 11 ovary-specific GPCRs in each organ.

Data and code availability:
Source data are available in the public datasets and databases listed in **Supplemental Data File 2**. All code for analysis of data is available on GitHub (https://github.com/Goods-Lab).

RESULTS

Curation and metadata harmonization of datasets across several databases to create the PreMeno Atlas

We first established a set of previously published datasets that could be used to create a pre-menopausal Atlas to support drug discovery efforts in the ovary by exploring existing, well-annotated datasets in several Atlas repositories. CELLxGENE, GEO, and HTCA databases were used to identify scRNA-seq datasets of interest following stringent filtering criteria (see **Methods**). The final list of datasets consisted of 14 organs, including ovary, lung, brain, liver, kidney, pancreas, skin, eye, breast, adipose tissue, intestine, heart, stomach, and uterus (**Supplemental Data File 2**). These datasets were manually curated from these databases to include samples derived from healthy premenopausal females (wherever available) and exclude samples derived from males and individuals with a history of reproductive diseases or those receiving hormonal treatments. Example filtering approaches using these databases are shown for the lung and kidney (**Supplemental Figure 1**). Despite efforts to filter in only healthy, female, and less than 50-year-old samples, certain organs, such as the eye and heart, had extremely limited or nonexistent pre-menopausal donor data, and thus are supplemented with postmenopausal donors as necessary to ensure cell representation across tissues (**Supplemental Data File 2**).

Prior to integration, all curated scRNA-seq datasets were harmonized to ensure consistency in data format, gene nomenclature, metadata, and cell type annotations. For metadata, we manually curated each dataset to ensure final integration would be successful. We then used the Human Cell Atlas Ontology tool to harmonize cell type labels. This involved collapsing granular cell types into higher-level labels and standardizing naming across different datasets while retaining the original cell-type annotations (see **Supplemental Table 2**). Thus, we did not change the original published annotations of cell types for each dataset, but instead harmonized cell type labels using the Ontology tool. This approach to metadata harmonization ensured that all datasets were aligned to a common framework, facilitating effective integration and enabling accurate cross-tissue and cross-cell type comparisons. Each final dataset had three levels of annotations, including organ (e.g. ovary), cell population (e.g. epithelial) and cell type (e.g. granulosa).

The harmonized PreMeno Atlas yields ovary-specific markers.

We next integrated each dataset into a final harmonized PreMeno Atlas following well-established scRNA-seq data integration workflows (see **Methods**). The final dataset is comprised of 511,365 cells across 14 tissues, including the ovary (**Figure 1A**). Samples from healthy tissues were derived from 115 donors total, with a median age of 35 (**Supplemental Data File 2**). The integrated dataset encompasses cells from multiple tissues, with some tissues contributing a larger number of cells than others (**Supplemental Figure 2**). The tissues with the top three maximum cell counts were breast (64,555 cells), ovary (60,994 cells), and uterus (53,825 cells), as we aimed to have better representation of organs relevant for female reproductive health. The tissues with smaller overall cell counts were adipose (17,959 cells), skin (18,389 cells), and stomach (17,782 cells), indicating the lack of readily available high-quality scRNA-seq datasets for those organs from healthy female donors. A UMAP of the

integrated dataset at the tissue level reveals distinct clustering by tissue type with overlap driven largely by shared cell populations (**Figure 1B**). For instance, cells from the ovary, skin, heart, and lung formed distinct clusters, indicating that the integration process effectively maintained their unique gene expression profiles. We also investigated the expression of several markers of key cell lineages, such as epithelial cells (EPCAM), hematopoietic (immune) cells (PTPRC), and fibroblasts (VIM) (**Figure 1C**, **Supplemental Figure 3**). The expression of these markers overlaps with major harmonized cell lineages. We plotted top gene markers for each organ (**Figure 1D**), demonstrating the utility of this integrated atlas in identifying tissue-specific individual genes and gene expression signatures. Several tissues showed gene expression correlated with expected markers and function, such as the brain, which showed heightened expression of synaptonemal genes such as ADGRB3, NRXN1, and IL1RAPL1. Other highly specialized organs similarly showed strong expression of canonical organ-specific genes, such as the eye (RHO, ROM1) and heart (VWF). Several ovary-specific genes, such as CTGF and ATP5B, ATP5B have strong known associations with follicle development and luteinization or mitochondrial function in oocytes, respectively.[41,42] Other genes (GNB2L1, TCEB1, SEPP1), have more limited known ovarian associations, but upon further analysis showed enrichment in ovarian stromal cell populations, including theca cells, which represented a large proportion of ovary-derived cells. We included additional cell labels, including cell population which largely represents cell origin (hematopoietic cells, epithelial cells, connective tissue cells, etc.) (**Figure 1E**), and the more granular cell type designation (**Figure 1F**).

These multi-level designations are useful for tasks such as atlas subsetting and pairwise analysis, as the ovary contains multiple cell types that are either abundant throughout the body but specialized for ovarian function (ovarian endothelial cell, ovarian macrophages) or organ-specific (e.g. oocyte, cumulus cell, mural granulosa cell). At the highly granular cell type level, we conducted pairwise differential expression analysis to characterize the relative similarity of ovarian cell types and all other integrated cell types based on gene expression using pairwise differential expression analysis (**Supplemental Figure 4A**). The number of significantly upregulated genes in ovarian cell types compared to reference cell types was the highest among pancreatic B cells, followed closely by numerous pancreatic cell types as well as multiple uterine cell types. Cell types that differed the least from ovarian cell types in terms of upregulated genes were largely derived from the stomach. Ovarian cells generally exhibited few downregulated genes with respect to cells of the intestine, lung, and kidney. Most cell types also showed asymmetry with respect to their upregulated and downregulated gene number compared to ovarian cell types. For instance, while very few upregulated genes exist between all ovarian cell types and stomach B cells, a high number of downregulated genes exist between ovarian cells and stomach B cells. Among ovarian cells, theca and stroma cells exhibited the most upregulated genes relative to the largest number of reference cell types, highlighting their distinct gene signatures and specificity. This may also be a product of the large number of ovarian theca and stromal cells included in the atlas; indeed, oocytes represent a smaller portion of ovarian-derived cells and generally exhibited fewer upregulated genes compared to all cell types. These

heatmaps serve as a reference to identify the most transcriptionally distinct cell types relative to ovarian cells.

We validated the pairwise transcriptomic distinctions among cell types by performing gene ontology enrichment analysis on significantly differentially expressed gene sets between two ovarian cell types and their most highly dissimilar reference cell types (**Supplemental Figure 4C&D**). We hypothesized that pairwise cell type comparisons with the most DEGs would result in gene expression signatures (e.g. upregulated genes) for each cell type that correlates with their specific respective biological functions. For instance, ovarian oocytes shared the most DEGs with pancreatic B cells; thus, GO and pathway analysis of genes upregulated in oocytes (relative to pancreatic B cells) should reveal female germ-cell specific functions, while those downregulated in oocytes (upregulated in pancreatic B cells) should reveal adaptive and humoral immune-specific functions. Indeed, the oocyte gene signature resulted in enrichment of GO terms including female gamete generation, pre-implantation embryo, and regulation of reproductive processes, whereas pancreatic B cells resulted in GO terms including adaptive immune response, immunoglobulin mediated immunity, and B cell mediated immunity (**Supplemental Figure 4C**). We further validated this hypothesis using another highly distinct pairwise cell-type comparison, ovarian cumulus cell and brain excitatory neuron (**Supplemental Figure 4D**), and confirmed enrichment of reproductive-specific GO terms among cumulus cells.[43] Characterization of global pairwise gene expression differences among our PreMeno Atlas provides insights into the underlying biological processes and molecular mechanisms that distinguish ovarian cell types from reference cell types and can serve as a tool for ovarian cell gene signature curation in the context of various organ and cell populations.

Use of the PreMeno Atlas to identify genes and pathways that are specific to ovarian theca and stromal cells

As we explored the utility of pairwise differential expression analysis to identify distinctions in ovarian cell type gene expression profiles relative to reference cell types, we found that ovarian theca and stromal cells represented a highly distinct transcriptomic niche in the PreMeno Atlas (**Supplemental Figure 2**). Therefore, we leveraged the PreMeno Atlas as a tool to explore cell type heterogeneity among a given cell population present in diverse organs, focusing on stromal cell types. Stromal cells also are common across many organs in the body, and thus unique genes or pathways in these cells across tissues may reveal insights into their tissue-specific adaptations. We sub-setted 19 stromal cell types from the integrated object and plotted their transcriptomic profiles on a UMAP to enable inspection of the relationships between diverse stromal cells at organ and cell type levels (**Figure 2A**). We observed clear grouping of stromal cell types by organ. Distinct clusters included adipose tissue and uterus, as well as ovarian stromal cell types, though they shared some spatial overlap with a core cluster of organs. The central cluster contains stromal cell types from breast, heart, intestine, kidney, liver, lung, pancreas, skin, and stomach. The top 5 cell-type-specific gene markers were plotted as a heatmap **(Figure 2B**). Stromal gene markers specific to ovarian theca and stromal cells include

GNB2L1, SEPP1, GLTSCR2, WBP5, and TCEB1. These genes showed highly restricted expression in ovarian theca and stromal cells, whereas marker genes from other stromal cell types showed cross over among different organs (e.g. stomach fibroblasts).

We next plotted the intersections of marker genes between ovarian stromal cells and all other stromal cell types in a pairwise Upset plot (**Figure 2C**). Among the 389 ovarian stroma marker genes, 180 marker genes were unique to the ovary. The most significant overlap was with skin stromal cells (99 genes, 25% overlap), followed by skin fibroblasts (80 genes) and pancreatic stellate cells (78 genes). The fewest genes were shared with adipose cells of the endometrium (2 genes) and preadipocytes (3 genes); this is expected based on the distinct clustering of adipose cell types in the UMAP (**Figure 2A**). We conducted gene ontology and pathway analysis on marker gene sets from all stromal cell types and assessed overlap of pathway terms (from GO biological pathway, KEGG, Reactome, and WikiPathway) among cell types (**Figure 2D**). Ovarian theca and stromal cells shared 640 of their 877 enriched pathway terms with skin stromal cells, followed by breast fibroblast (551 terms) and uterus decidual cells (539 terms). Among top pathways enriched by ovarian theca and stromal cell marker genes, protein regulation stands out as a highly expressed biological function, represented by terms including protein refolding, chaperone cofactor-dependent protein refolding, 'de-novo' post-translational protein folding, and eIF3 complex (**Figure 2E**). Heat shock protein activity regulation was also represented by HSF1-dependent transactivation and attenuation phase pathways.

We also evaluated the specificity of genes involved in reproductive pathways enriched in ovarian theca and stromal cells using a heatmap to visualize ovary-specific gene markers across stroma cell types (**Figure 2F**). Ovarian stroma marker genes involved in estrogen signaling were generally specific to the ovary, particularly HSPA6; however, GREB1 and KRT19 were expressed in other stromal cell types to an equal or greater extent than ovarian stromal cells. HSPA2, an ovarian stroma marker gene involved in meiosis in males, was not highly specific to the ovary and shared similar expression levels in skin and lung stromal cells.[44,45] Genes involved in regulation of cell growth, particularly SGK1 and TSPYL2, were largely specific to ovarian stroma, with other genes of that pathway sharing similar expression levels with at least one other stromal cell type.[46,47] No single gene in the reproductive structure development pathway was exclusively expressed in ovarian stromal cells. However, strong expression of ING2 was observed in the ovarian stroma along with endometrial stromal cells and intestine fibroblasts. Lastly, NR4A3 of the corticotropin-releasing hormone pathway [48] was highly expressed by ovarian theca cells, followed by NR4A1 and NR4A2. This comparative approach serves as a tool to understand the expression patterns of ovary-specific genes across different stromal cell types, providing a comprehensive overview of their molecular relationships and potential functional roles.

PreMeno Atlas for contraceptive drug target discovery

Given that we were able to identify well-established ovarian function genes and new genes through our analyses, we next leveraged the PreMeno Atlas for contraceptive discovery. For this, we wanted to identify PreMeno Atlas-defined marker genes that define cell types in the ovary, and cross reference these against other databases to assess their druggability. We first identified top ovarian-cell-specific marker genes at the cell type level and queried them against the Contraceptive Infertility Target Database (CITDBase) (**Figure 3A&B**).[37] Cumulus cells and mural granulosa cells shared similar marker genes, with strong expression of GSTA1, ACSM3, AMH, TNNI3, and SERPINE2 in both cell types, among other markers. Oocytes expressed a distinct set of marker genes that share limited expression among other ovarian cell types. Indeed, known oocyte-specific genes, including OOSP2, DPPA3, PADI6, FIGLA, and TUBB8, appeared among the top 10 marker gene list. Ovarian theca and stromal cells also expressed a distinct set of marker genes that include particularly strong expression of GAS5, STAR, TCF21, PEG3, C7 and OGN. CITDBase leverages mouse knockout (MGI and IMPC) and human clinical (ClinVar, HPO, Monarch) data to report fertility phenotypes for given genes. Here, we note significant representation of oocyte marker genes among female infertility phenotypes in MGI, with PADI6 also reportedly resulting in female infertility in IMPC.[49] GATA4 and INHA, mural granulosa cell marker genes, and FST, a cumulus cell marker gene,[50] also result in female infertility in knockout mice according to MGI. Genes that result in female or non-sex-specific subfertility include GREB1 (theca & stromal cell), HSD17B1 (mural granulosa cell), MFAP2 and MEX3B (cumulus cell). Genes with consistent reports of female infertility in human clinical databases include FIGLA, PADI6, TUBB8, which are all oocyte marker genes, the former two with infertility phenotypes in mouse knockouts. Male infertility phenotypes were reported for mural granulosa cell markers AMH and GATA4 in the HPO, Monarch, and ClinVar (GATA4 only) databases, while ClinVar reported male infertility for the theca cell marker gene STAR.[51]

To explore concordance between marker gene expression and protein expression, we plotted their expression level using tissue-specific data from Human Protein Atlas (HPA).[52] Protein level, categorized as low, medium, high, and not detected, is reported in 76 cell types in 45 normal tissues. Most ovarian cell marker genes show undetected protein expression in all tissue types, including the ovary. This is not surprising given the wealth of established literature that highlights discrepancies between RNA and protein levels,[53–55] but may also represent some limitations in terms of total numbers of proteins detected in proteomic datasets. Some ovarian cell marker genes, including TCF21, PEG3, GREB1, ACSM3, and GRB14, have low but consistent protein levels across numerous tissue types. Others including CLDN11, GATA4, GSAT1, FDXR, and S100B, have high protein levels in a handful of tissues. These more ubiquitous genes are commonly expressed in the testis, adrenal gland, and prostate.[56] Marker genes with detectable protein expression in the ovaries, according to HPA, include FIGLA, AMH, GSTA1, GREB1, OGN, KLHDC8A, TCF21, and STAR. Several of these are well-established in the context of ovarian biology.[57–59] Taken together, this integrated analysis provides several novel genes and proteins that are associated with ovarian cell types and potential drug targets based on CITDBase results.

GPCRs are historically druggable, and thus could make excellent targets for modulation of ovarian function if there were a specific ovarian GPCR to target.[60,61] To explore this, we plotted the expression of top G-protein coupled receptors (GPCRs) detected in the PreMeno Atlas in all tissue types (**Figure 3C)**. Several of these receptors exhibited heightened expression in the ovary compared to other tissue types, including GPR78, ADRB3, GPR20, and GPR101, whereas MRGPRF and GPR21 show high expression in the uterus as well. CALCRL and GPR146 are expressed highly in the ovary but also other tissues, such as brain, breast, uterus, and heart. Finally, TAS2R7, NTSR2, and UTS2R have relatively low expression in the ovary, despite being among the highest expressed GPCRs in the ovary, due to their higher expression in brain and breast tissue. This GPCR-specific query of PreMeno Atlas coupled with contraceptive-specific database and HPA integration reflects the utility of this tool for drug discovery purposes.

Assessing mouse ovarian gene homology with PreMeno Atlas

Next, we wanted to determine how our ovary-enriched genes compared to those in mouse ovary since this is an important model system for contraceptive and fertility research. We integrated two scRNA-seq ovarian datasets derived from mice (**Supplemental Figure 5**). The first is a temporal ovulation dataset that captures ovulation events after ovarian hyperstimulation.[22] The second is a dataset that captures the ovary across the estrous cycle.[24] We reasoned that these two datasets should provide comprehensive coverage of cell types in the mouse ovary. The two datasets were integrated following standard workflows (see **Methods**). Clusters were annotated based on the original published studies and integrated marker gene identification (**Supplemental Data File 3**), identifying the following cell types: antral mural granulosa, atretic granulosa, cumulus, endothelial, epithelial immune, oocyte, pericyte, preantral granulosa, smooth muscle, steroidogenic theca, stromal, and theca (**Figure 4A**). These cell types are defined by distinct expression patterns of key ovarian genes and are also enriched for pathways that are key to their biological functions (**Supplemental Figure 5**). We plotted the expression of our identified human ovarian cell marker genes from oocytes, mural granulosa cells, cumulus cells, and theca cells (**Figure 3B**) in the mouse ovarian integrated dataset to identify homologous marker genes (**Figure 4B**). Interestingly, we observed significant conservation of all oocyte marker genes and mostly restricted expression in mouse oocytes. Only *Ucma*, homolog of the human oocyte marker UCMA, displayed strongest expression in mouse ovarian epithelial cells. Several homologs of human mural granulosa cell marker genes showed highest expression in antral and preantral mural granulosa cells in mice (*Greb1*, *Phex*, *Tnni3*, *Hsd17b1*, *Inhbb*, *Acsm3*, *Amh*, *Gata4*), while others showed heightened expression in cumulus cells (*Gsta1*), steroidogenic theca cells (*Gsta1*, *Fdxr*), theca cells (*Fdxr*), stromal cells (*Lrat*), and epithelial cells (*Gata4*). Several homologs of human cumulus cell marker genes showed highest expression in cumulus cells in mice (*Inhbb*, *Serpine2*, *Inha*, *Fst*); these genes also commonly shared heightened expression with antral mural granulosa cells. Other homologs of human cumulus cell marker genes were expressed highest (relatively) in smooth muscle cells (*Mfap2*), theca cells (*S100b*), pericytes (*Ndp*), and atretic granulosa cells (*Grb14*, *Mex3b*) in

mice. Lastly, human theca cell marker homologs were highly expressed in cumulus cells (*Cldn11*, *Star*), steroidogenic theca cells (*C7*, *Klhdc8a*), smooth muscle cells (*Ogn*, *Tcf21*), stromal cells (*Tcf21*), pericytes (*Colec11*), and epithelial cells (*Peg3*) in mice.

Finally, we visualized several ovarian cell marker genes in a single cell imaging spatial transcriptomics (iST) dataset of a mouse ovaries taken across an ovulation time course 0, 4 and 12 hours after exposure to ovulatory stimuli (**Figure 4C**). *Amh* is highly expressed in preantral granulosa cells in mice, while its human gene homolog, AMH, is a mural granulosa cell marker.[62] In the MERSCOPE ovary data, *Amh* expression was correlated with small, preantral follicles with stable expression across the ovulation time course in all tissue sections (**Figure 4C, Supplemental Figure 6**). *Inha* and *Serpine2* are expressed in cumulus cells and antral-mural granulosa cells in mice, like their human homologs; spatially, *Inha* and *Serpine2* were distributed among larger, antral follicles in both the antrum and surrounding the oocyte (**Figure 4D&E, Supplemental Figure 7&8**). STAR is a theca cell marker gene in humans; in mice, *Star* showed higher expression in cumulus cells than theca cells, though it's expression in theca cells was still pronounced (**Figure 4B**). *Star* showed strong expression in theca cells throughout the stroma surrounding visible follicles with a distinct increase in expression 12 hours after hCG stimulation in the mouse ovulatory time course (**Supplemental Figure 8**). This spatial data validates and expands upon the integrated mouse ovarian dataset and substantiates marker gene homology among human ovarian cell types.

DISCUSSION

The PreMeno Atlas offers a comprehensive and integrated single-cell reference of healthy, premenopausal female tissues with a focus on integration of the ovary. Ovary-specific scRNA-seq studies have made important contributions to our understanding of ovarian biology in the context of spatial organization of the ovary and cellular heterogeneity,[14,22–24,63] as well as molecular mechanisms of cellular aging.[8,9] These resources serve as valuable contributions to reproductive biology research, and their use for contraceptive drug development parallels tissue-specific atlases that have been applied directly for drug discovery in the context of Alzheimer's disease,[66] inflammatory bowel disease,[67] and colorectal cancer,[68] for instance. However, the aforementioned efforts have not yet been integrated with multi-tissue references; thus, ovarian specificity of a given gene of interest or biological function cannot be directly ascertained from these studies alone. Multi-tissue human cell atlases, which can be leveraged for target identification and tissue specificity analysis, often exclude the ovary[25,28] or include ovarian tissue from post-menopausal donors.[69] Our work directly addresses this gap by incorporating healthy, pre-menopausal reference tissues.

To this end, we curated scRNA-seq data from 14 reference tissue types derived from existing single and multi-tissue cell atlases by implementing filtering strategies for donor metadata (female, reproductive age, no reported disease phenotypes). Tissue-specific genes recovered from marker gene assessment of the integrated atlas at the organ level correspond

to expected cell types. For instance, brain-specific genes included neural synapse markers NRXNL1, IL1RAPL1, and ADGRB3, all of which are known to be highly expressed in the brain.[70] Among ovary marker genes, we identified many ATP synthase subunit-encoding genes (ATP5G2, ATP5E, and ATP5B), which together are integral to ATP synthase (Complex V) function during oxidative phosphorylation.[71] NGFRAP1 (BEX3), a modulator of neurotropin signaling, is known to be highly expressed in ovarian granulosa cells and is believed to play a role in neuronal growth-factor induced apoptosis.[72] Its role in the ovary has not been explored, but may relate to extensive apoptotic pruning of non-dominant follicles and neurotropic signaling in follicle atresia.[73] Interestingly, we also identified several ovarian tissue marker genes that do not have well described roles with the ovary, highlighting the utility of this resource. For instance, GLTSCR2, a tumor suppressor involved in p53 stability, ribosome biogenesis, and nucleolar stress response, has not been well documented in the ovary but is interesting given the link between p53 and oocyte quality.[74] These findings highlight the use of the PreMeno Atlas to identify novel ovarian-specific genes with potential roles in ovarian biology.

We then used the PreMeno Atlas to explore cell and tissue-specific gene differences, providing pairwise differentially expressed gene lists for all cell types in comparison to ovarian cell types. We further used stromal cells as an example of analyzing shared cell types across organs to identify tissue-specific markers of cell types. Stromal cells are of interest due to their anatomic and donor-to-donor heterogeneity, as well as their potential druggability in relevant pathological conditions, such as stromal-rich cancers.[75,76] Ovarian stroma and theca cells play critical roles in folliculogenesis; indeed, theca cells are recruited at the secondary follicle stage to support developing follicles structurally, mechanically, and molecularly via androgen synthesis.[77,78] Dysfunction of theca cells results in subfertility and infertility due to hyperandrogenism in the case of overactivity and hypoestrogenism in the case of underactivity or functional knockout.[79,80] We identified ovary-specific stromal cell marker genes (GNB2L1, SEPP1, GLTSCR2, WBP5, TCEB1). While these genes have not been extensively characterized in ovarian theca or stromal cells, their known molecular functions align with critical aspects of ovarian physiology. For instance, GNB2L1 (RACK1) is a scaffold protein involved in signal transduction and ribosomal function, which may influence local translation or integrin-mediated adhesion of theca cells in secondary/preantral follicles, where they play a scaffolding and communication role with granulosa cells.[81,82] SEPP1, a selenium transporter and antioxidant, could support oxidative stress buffering required for properly timed follicular activation and oocyte competency.[83] Theca cells are known to express superoxide dismutases (SOD) that sequester reactive oxygen species, but the selenium-mediated glutathione peroxidase (GPX) mechanism of peroxide degradation has not been well characterized in human theca cells; thus, the detection of SEPP1 premises further exploration of this selenium transport mechanism.[83–86] Taken together, this highlights how the PreMeno Atlas can be used to investigate biology and as a tool to drive new hypotheses about cell types in the ovary.

We also use the PreMeno Atlas to nominate putative contraceptive targets based on ovarian cell-type specificity and through integrated analyses with databases and the HPA. We identified the most highly ovarian-specific GPCRs (GPR78, ADRB3, GPR20, and GPR101), and given their reproductive tissue specificity, these receptors could be ovarian-specific contraceptive targets assuming requisite involvement in fertility. ADRB3 is reportedly also associated with male infertility (ClinVar on CITDbase), and is reportedly ovary specific according to Consensus-RNA from Human Protein Atlas.[52] No mouse infertility phenotype is reported as no knock-out has been tested by MGI or IMPC; as such, ADRB3 is a promising druggable contraceptive target that has yet to be explored in the context of female fertility. We also found several ovarian cell marker genes from our analyses have promising druggability profiles in human and mouse infertility phenotype databases. INHA and FST are associated with female infertility phenotypes in mice and have roles in regulating follicle stimulating hormone (FSH) production and receptivity.[87–89] However, due to their central roles in gonadotropin signaling, they may be poor candidates for nonhormonal contraceptives. Several oocyte-specific genes display both mouse and/or human infertility phenotypes, including PADI6, FIGLA, SOHLH2, TUBB8, DPPA3 and LHX8. These genes are predominantly expressed in oocytes and play essential roles in oocyte maturation, meiotic progression, and early embryonic development.[90–92] For example, PADI6 is critical for the formation of the subcortical maternal complex, and its disruption leads to embryonic arrest at the 2-cell stage in mice; because of its essential role post-fertilization, targeting it could pose risks if conception occurs despite contraceptive use.[93] FIGLA and SOHLH2 are transcription factors required for primordial follicle formation and oocyte differentiation, while TUBB8 encodes a beta-tubulin isoform specific to oocytes and is essential for proper spindle formation during meiosis.[92] Importantly, several of these genes have bioactive compounds or inhibitors reported in drug-target databases, and their oocyte-restricted expression patterns suggest they could be targeted without systemic hormonal disruption. Further exploration of these genes and ovary-specific GPCRs could therefore advance ovarian contraceptive discovery.

We also performed an integrated analysis of several identified targets through CITDBase with data from the Human Protein Atlas. Not surprisingly, we observed limited overlap between ovarian cell-specific marker genes and ovarian protein level information curated from Human Protein Atlas given that proteomics data has many fewer features than transcriptomics data. We did find high expression of several ovarian genes, including FIGLA, INHA, and FDXR. While this may be due to true biological attenuation of mRNA to protein, it is more likely the product of limited primary data on the ovarian proteome and the low abundance of oocytes. Only a handful of studies have provided reference maps of the human ovarian proteome[94,95] and matrisome.[96] While valuable, these studies report data from small sample sizes (2 to 5 women per age group) and offer little cell-specific information. Broader and more precise proteomic references for ovarian cell types and tissue would facilitate integration of multi-omic ovarian atlases such as

ours for more robust drug target identification. More generally, a female-specific protein expression reference database may prove a more precise tool for contraceptive target validation.

Finally, to better understand cross-species ovarian biology and assess the conservation of key cellular markers, we integrated scRNA-seq data from prior mouse ovary datasets, harmonized cell-type annotations, and overlaid human cell-type-specific markers to identify homologous expression patterns. The strong conservation of oocyte markers and their spatial restriction to mouse oocytes supports the utility of these genes as robust identifiers across species.[63,97] Interestingly, mural granulosa and cumulus cell markers demonstrated partially overlapping expression between humans and mice, with some markers broadly expressed across granulosa subtypes and others localized to non-granulosa cell types such as smooth muscle and stromal cells. These differences may reflect species-specific divergence or underscore plasticity in granulosa cell identities across follicle stages.[98,99] Our analysis reveals that while many human ovarian markers retain conserved expression in mouse orthologs, their distribution across cell types may vary slightly, emphasizing the need for careful interpretation when translating findings between species.

The PreMeno Atlas has a few limitations and challenges that we would like to note. Variability in data quality across different scRNA-seq datasets posed challenges for integration and harmonization, and ultimately given our very stringent quality control filters, we ended up losing many cells for the sake of higher quality data. Larger Atlas efforts that prioritize very high-quality cells and under-represented tissues in younger female donors would help to mitigate this challenge for future studies. The datasets we used were also derived from multiple sources with varying levels of metadata annotation, leading to challenges in data preprocessing and metadata harmonization. We addressed this with extensive manual curation and the use of publicly available tools for cell type naming harmonization, but the lack of automated protocol for this process can make the addition of future data to this atlas challenging. The lack of comprehensive metadata for non-ovarian tissues, such as donor age or disease history, complicated standardization and integration processes. Our study suggests that targeted approaches for the development of oocyte data specifically would be very valuable for drug discovery efforts given that this cell type is often poorly represented in scRNA-seq datasets. Finally, through our curation efforts we identified several tissues, including the stomach and heart, that had poor representation of healthy, female-derived data. Future efforts could prioritize the generation of these healthy references in pre-menopausal populations if possible.

CONCLUSION

We show that there are clear marker genes that define cell types in the ovary when compared to the rest of the human body, including cumulus cells, oocytes, and ovarian theca/stromal cells. We used the PreMeno Atlas to explore differences in stromal cells across the body, highlighting shared and unique pathways that are enriched in these cell types. Finally,

we show that this resource can be used to find ovary-specific genes that can serve as possible therapeutic targets for contraceptive development or fertility enhancement. Taken together, our PreMeno Atlas provides a critical resource for identifying ovary-specific genes, greatly enhancing the capacity for drug target discovery using human datasets. While our efforts focused on highlighting identified genes for possible contraception, this resource and genes identified can also be used for identifying novel genes involved in ovarian function, guide marker selection for ovarian biology, and nominating possible genes for fertility enhancement. Our analyses suggest that future work should seek to bolster scRNA-seq Atlas efforts from underrepresented tissues in female donors, expand the ovarian proteome to bolster cross-dataset comparisons, and validate several identified targets experimentally, particularly in cases where information is lacking from existing databases.

Data Availability

Source data are available in the public datasets and databases in **Supplemental Data File 1**. All code for analysis of data is available on GitHub (https://github.com/Goods-Lab).

Acknowledgments

We would like to thank Dr. Margie Ackerman Lab and Dr. Jiwon Lee, and their lab members for helpful discussions throughout this project. We would also like to thank Dr. Shuo Xiao, Dr. Jeff Pea, and Dr. Stephen Ward for helpful discussions. This work was supported by the Bill & Melinda Gates Foundation [INV-003385 and INV-040475]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. B.A.G is supported in part through the Geisel School of Medicine at Dartmouth's Center for Quantitative Biology through a grant from the National Institute of General Medical Sciences (NIGMS, P20GM130454) of the NIH.

Author Contributions

B.A.G. conceptualized the work, supervised all aspects of the study, generated figures, and wrote the manuscript. O.K. supervised data curation and integration. Data curation was performed by Z.N., R.H., and O.K. H.V., Z.N., A.O., O.K., C.K., I.G. and B.A.G performed analyses and generated figures. Original paper drafting was performed by H.V., O.K., and B.A.G.; all authors contributed to editing of the manuscript. Funding was secured by B.A.G. and F.D. All authors participated in discussions regarding this manuscript and approved its final submission.

Conflicts of Interest

The authors declare no competing interests.

References:

1. Contraceptive Use in the United States by Demographics | Guttmacher Institute. https://www.guttmacher.org/fact-sheet/contraceptive-use-united-states (2020).
2. Cooper, D. B. & Patel, P. Oral Contraceptive Pills. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
3. Moreira, L. R., Ewerling, F., Barros, A. J. D. & Silveira, M. F. Reasons for nonuse of contraceptive methods by women with demand for contraception not satisfied: an assessment of low and middle-income countries using demographic and health surveys. *Reproductive Health* **16**, 148 (2019).
4. Johnston, D. S. & Kopf, G. S. The urgent need for innovation in contraception. *Biology of Reproduction* **108**, 519–521 (2023).
5. Van de Sande, B. *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* **22**, 496–520 (2023).
6. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993–996 (2006).
7. Radoux, C. J., Vianello, F., McGreig, J., Desai, N. & Bradley, A. R. The druggable genome: Twenty years later. *Front. Bioinform.* **2**, (2022).
8. Edson, M. A., Nagaraja, A. K. & Matzuk, M. M. The Mammalian Ovary from Genesis to Revelation. *Endocrine Reviews* **30**, 624–712 (2009).
9. Matzuk, M. M. & Burns, K. H. Genetics of mammalian reproduction: modeling the end of the germline. *Annu Rev Physiol* **74**, 503–528 (2012).
10. The GTEx Consortium atlas of genetic regulatory effects across human tissues | Science. https://www.science.org/doi/10.1126/science.aaz1776.
11. Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P. & Hayashizaki, Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci Data* **4**, 170113 (2017).
12. Babayev, E. *et al.* Cumulus expansion is impaired with advanced reproductive age due to loss of matrix integrity and reduced hyaluronan. *Aging Cell* **22**, e14004 (2023).
13. Balough, J. L., Dipali, S. S., Velez, K., Kumar, T. R. & Duncan, F. E. Hallmarks of female reproductive aging in physiologic aging mice. *Nat Aging* **4**, 1711–1730 (2024).
14. Wu, M. *et al.* Spatiotemporal transcriptomic changes of human ovarian aging and the regulatory role of FOXP1. *Nat Aging* **4**, 527–545 (2024).
15. Ikeda, K., Horie-Inoue, K. & Inoue, S. Identification of estrogen-responsive genes based on the DNA binding properties of estrogen receptors using high-throughput sequencing technology. *Acta Pharmacol Sin* **36**, 24–31 (2015).
16. Ayehunie, S. *et al.* Characterization of a Hormone-Responsive Organotypic Human Vaginal Tissue Model. *Reprod Sci* **22**, 980–990 (2015).
17. Cicatiello, L. *et al.* Gene expression monitoring in hormone-responsive human breast cancer cells during estrogen-induced cell cycle progression. *Nat Genet* **27**, 95–95 (2001).
18. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 1–14 (2018).
19. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631–656 (2019).

20. Mantri, M., Zhang, H. H., Spanos, E., Ren, Y. A. & De Vlaminck, I. A spatiotemporal molecular atlas of the ovulating mouse ovary. *Proceedings of the National Academy of Sciences* **121**, e2317418121 (2024).
21. Marečková, M. *et al.* An integrated single-cell reference atlas of the human endometrium. *Nat Genet* **56**, 1925–1937 (2024).
22. Huang, R. *et al.* Single-cell and spatiotemporal profile of ovulation in the mouse ovary. *bioRxiv* 2024.05.20.594719 (2024) doi:10.1101/2024.05.20.594719.
23. Jones, A. S. K. *et al.* Cellular atlas of the human ovary using morphologically guided spatial transcriptomics and single-cell sequencing. *Science Advances* **10**, eadm7506 (2024).
24. Morris, M. E. *et al.* A single-cell atlas of the cycling murine ovary. *Elife* **11**, e77239 (2022).
25. The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
26. HTCA: a database with an in-depth characterization of the single-cell human transcriptome | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/51/D1/D1019/6709243?login=false.
27. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041.
28. Pan, L. *et al.* Single Cell Atlas: a single-cell multi-omics human cell encyclopedia. *Genome Biology* **25**, 104 (2024).
29. CZI Cell Science Program *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research* **53**, D886–D900 (2025).
30. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995 (2013).
31. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184–1191 (2009).
32. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics* **7**, 44 (2016).
33. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
34. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* **8**, 329-337.e4 (2019).
35. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
37. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).
38. Clustering trees: a visualization for evaluating clusterings at multiple resolutions | GigaScience | Oxford Academic. https://academic.oup.com/gigascience/article/7/7/giy083/5052205?login=false.
39. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
40. Sinha, S. *et al.* Contraceptive and Infertility Target DataBase: a contraceptive drug development tool for targeting and analysis of human reproductive specific tissues†. *Biology of Reproduction* **105**, 1366–1374 (2021).

41. Bomba-Warczak, E. K. *et al.* Exceptional longevity of mammalian ovarian and oocyte macromolecules throughout the reproductive lifespan. *eLife* **13**, RP93172 (2024).
42. Nagashima, T. *et al.* Connective tissue growth factor is required for normal follicle development and ovulation. *Mol Endocrinol* **25**, 1740–1759 (2011).
43. Ma, R.-J. *et al.* Role of miRNAs in glucose metabolism of mouse cumulus cells†. *Biol Reprod* **110**, 895–907 (2024).
44. Son, W. Y. *et al.* Specific expression of heat shock protein HspA2 in human male germ cells. *Mol Hum Reprod* **5**, 1122–1126 (1999).
45. Harasimov, K. *et al.* The maintenance of oocytes in the mammalian ovary involves extreme protein longevity. *Nat Cell Biol* **26**, 1124–1138 (2024).
46. PubChem. SGK1 - serum/glucocorticoid regulated kinase 1 (human). https://pubchem.ncbi.nlm.nih.gov/gene/SGK1/human.
47. TSPYL2 TSPY like 2 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/64061.
48. PubChem. NR4A3 - nuclear receptor subfamily 4 group A member 3 (human). https://pubchem.ncbi.nlm.nih.gov/gene/NR4A3/human.
49. Austin, C. P. *et al.* The Knockout Mouse Project. *Nat Genet* **36**, 921–924 (2004).
50. Fullerton, P. T., Monsivais, D., Kommagani, R. & Matzuk, M. M. Follistatin is critical for mouse uterine receptivity and decidualization. *Proceedings of the National Academy of Sciences* **114**, E4772–E4781 (2017).
51. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
52. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci* **27**, 233–244 (2018).
53. Ghazalpour, A. *et al.* Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLoS Genet* **7**, e1001393 (2011).
54. Weerakoon, H. *et al.* Integrative temporal multi-omics reveals uncoupling of transcriptome and proteome during human T cell activation. *npj Syst Biol Appl* **10**, 1–13 (2024).
55. Veerla, J. P., Saurav, J. R., Robben, M. & Luber, J. M. Analyzing Lack of Concordance Between the Proteome and Transcriptome in Paired scRNA-Seq and Multiplexed Spatial Proteomics. Preprint at https://doi.org/10.48550/arXiv.2307.00635 (2023).
56. Pignatti, E. *et al. FDXR* variants cause adrenal insufficiency and atypical sexual development. *JCI Insight* **9**, (2024).
57. Dewailly, D. & Laven, J. AMH as the primary marker for fertility. *Eur J Endocrinol* **181**, D45–D51 (2019).
58. Wang, Z., Liu, C.-Y., Zhao, Y. & Dean, J. FIGLA, LHX8 and SOHLH1 transcription factor networks regulate mouse oocyte growth and differentiation. *Nucleic Acids Research* **48**, 3525–3541 (2020).
59. Ronen-Fuhrmann, T. *et al.* Spatio-temporal expression patterns of steroidogenic acute regulatory protein (StAR) during follicular development in the rat ovary. *Endocrinology* **139**, 303–315 (1998).
60. Lorente, J. S. *et al.* GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* 1–22 (2025) doi:10.1038/s41573-025-01139-y.
61. Sriram, K. & Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol Pharmacol* **93**, 251–258 (2018).

62. Tanimoto, R. *et al.* Blocking estrogen-induced AMH expression is crucial for normal follicle formation. *Development* **148**, dev197459 (2021).
63. Fan, X. *et al.* Single-cell reconstruction of follicular remodeling in the human adult ovary. *Nat Commun* **10**, 3164 (2019).
64. Zhou, C. *et al.* Single-Cell Atlas of Human Ovaries Reveals The Role Of The Pyroptotic Macrophage in Ovarian Aging. *Advanced Science* **11**, 2305175 (2024).
65. Isola, J. V. V. *et al.* A single-cell atlas of the aging mouse ovary. *Nat Aging* **4**, 145–162 (2024).
66. The Alzheimer's Cell Atlas (TACA): A single-cell molecular map for translational therapeutics accelerator in Alzheimer's disease - PubMed. https://pubmed.ncbi.nlm.nih.gov/36254161/.
67. DOP132 An integrated single-cell atlas enables the discovery and validation of a novel candidate therapeutic target for Inflammatory Bowel Disease | Journal of Crohn's and Colitis | Oxford Academic. https://academic.oup.com/ecco-jcc/article/19/Supplement_1/i325/7966930.
68. Single-cell atlas defines distinct immune escape mechanisms in colorectal cancer | Nature Cancer. https://www.nature.com/articles/s43018-024-00809-x.
69. Pan, L. *et al.* HTCA: a database with an in-depth characterization of the single-cell human transcriptome. *Nucleic Acids Research* **51**, D1019–D1028 (2023).
70. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci Rep* **8**, 8868 (2018).
71. Walker, J. E. The ATP synthase: the understood, the uncertain and the unknown. *Biochem Soc Trans* **41**, 1–16 (2013).
72. Jiang, J.-Y. *et al.* Mural granulosa cell gene expression associated with oocyte developmental competence. *J Ovarian Res* **3**, 6 (2010).
73. Sidhar, H. & Giri, R. K. Induction of Bex genes by curcumin is associated with apoptosis and activation of p53 in N2a neuroblastoma cells. *Sci Rep* **7**, 41420 (2017).
74. Hu, W., Zheng, T. & Wang, J. Regulation of Fertility by the p53 Family Members. *Genes Cancer* **2**, 420–430 (2011).
75. Kozlova, N., Grossman, J. E., Iwanicki, M. P. & Muranen, T. The Interplay of the Extracellular Matrix and Stromal Cells as a Drug Target in Stroma-Rich Cancers. *Trends in Pharmacological Sciences* **41**, 183–198 (2020).
76. Stromal heterogeneity may explain increased incidence of metaplastic breast cancer in women of African descent | Nature Communications. https://www.nature.com/articles/s41467-023-41473-6.
77. Theca: the forgotten cell of the ovarian follicle in: Reproduction Volume 140 Issue 4 (2010). https://rep.bioscientifica.com/view/journals/rep/140/4/489.xml.
78. Biswas, A. *et al.* Theca cell mechanics and tissue pressure regulate mammalian ovarian folliculogenesis. 2024.05.06.592641 Preprint at https://doi.org/10.1101/2024.05.06.592641 (2024).
79. Ma, Y. *et al.* Androgen Receptor in the Ovary Theca Cells Plays a Critical Role in Androgen-Induced Reproductive Dysfunction. *Endocrinology* **158**, 98–108 (2017).
80. Magoffin, D. A. Ovarian theca cell. *The International Journal of Biochemistry & Cell Biology* **37**, 1344–1349 (2005).
81. Itami, S., Tamotsu, S., Sakai, A. & Yasuda, K. The Roles of THY1 and Integrin Beta3 in Cell Adhesion During Theca Cell Layer Formation and the Effect of Follicle-Stimulating Hormone on THY1 and Integrin Beta3 Localization in Mouse Ovarian Follicles. *Biology of Reproduction* **84**, 986–995 (2011).

82. RACK1 receptor for activated C kinase 1 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/10399.
83. Qin, S. *et al.* Oocytes maintain low ROS levels to support the dormancy of primordial follicles. *Aging Cell* **24**, e14338 (2025).
84. SUGINO, N. Reactive oxygen species in ovarian physiology. *Reprod Med Biol* **4**, 31–44 (2005).
85. Lava Kumar, S. *et al.* Glutathione peroxidase (GPX1) - Selenocysteine metabolism preserves the follicular fluid's (FF) redox homeostasis via IGF-1- NMD cascade in follicular ovarian cysts (FOCs). *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1870**, 167235 (2024).
86. Hummitzsch, K. *et al.* Expression levels of the selenium-uptake receptor LRP8, the antioxidant selenoprotein GPX1 and steroidogenic enzymes correlate in granulosa cells. *Reprod Fertil* **5**, e230074 (2024).
87. PubChem. FST - follistatin (human). https://pubchem.ncbi.nlm.nih.gov/gene/FST/human.
88. Nagaraja, A. K. *et al.* Defective Gonadotropin-Dependent Ovarian Folliculogenesis and Granulosa Cell Gene Expression in Inhibin-Deficient Mice. *Endocrinology* **151**, 4994–5006 (2010).
89. Meinsohn, M.-C. *et al.* Single-cell sequencing reveals suppressive transcriptional programs regulated by MIS/AMH in neonatal ovaries. *Proc Natl Acad Sci U S A* **118**, e2100920118 (2021).
90. Wagner, M. *et al.* Single-cell analysis of human ovarian cortex identifies distinct cell populations but no oogonial stem cells. *Nat Commun* **11**, 1147 (2020).
91. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Molecular Cell* **72**, 1021-1034.e4 (2018).
92. Chen, Y.-Y. *et al.* Single-cell transcriptomics of staged oocytes and somatic cells reveal novel regulators of follicle activation. *Reproduction* **164**, 55–70 (2022).
93. Williams, J. P. C. & Walport, L. J. PADI6: What we know about the elusive fifth member of the peptidyl arginine deiminase family. *Philos Trans R Soc Lond B Biol Sci* **378**, 20220242.
94. Wang, L. *et al.* A two-dimensional electrophoresis reference map of human ovary. *J Mol Med (Berl)* **83**, 812–821 (2005).
95. Ouni, E., Vertommen, D., Chiti, M. C., Dolmans, M.-M. & Amorim, C. A. A Draft Map of the Human Ovarian Proteome for Tissue Engineering and Clinical Applications. *Mol Cell Proteomics* **18**, S159–S173 (2019).
96. Ouni, E. *et al.* Proteome-wide and matrisome-specific atlas of the human ovary computes fertility biomarker candidates and open the way for precision oncofertility. *Matrix Biol* **109**, 91–120 (2022).
97. Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858-873.e4 (2017).
98. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Mol Cell* **72**, 1021-1034.e4 (2018).
99. Poulsen, L. C. *et al.* Two waves of transcriptomic changes in periovulatory human granulosa cells. *Hum Reprod* **35**, 1230–1245 (2020).

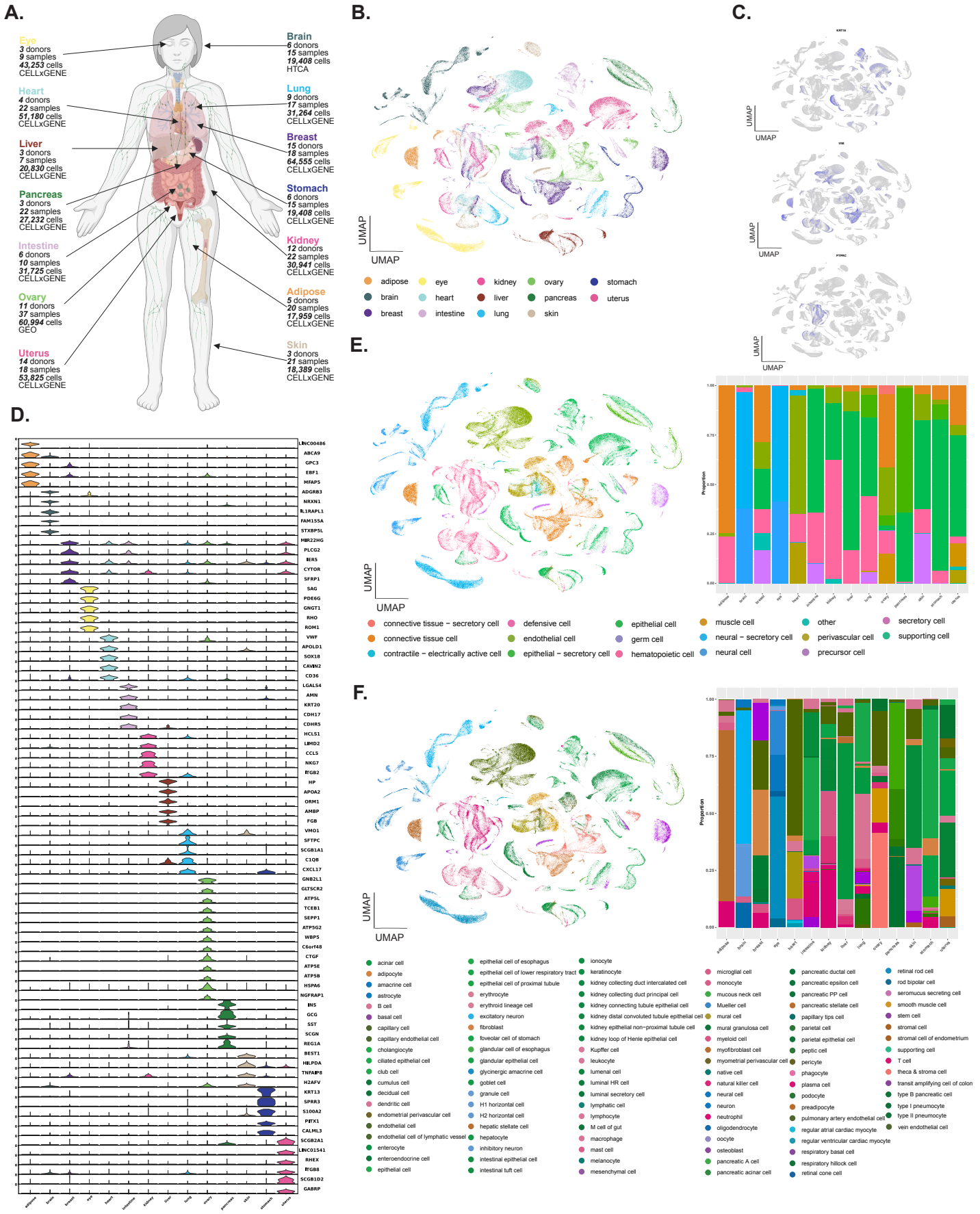

Figure 1. Overview of the data curation and integration process for the PreMeno Atlas. A. Diagram of Atlas composition by organ, donor number, sample number, and data source. **B.** Uniform

Manifold Approximation Projection (UMAP) along dimensions 1 and 2 for all data colored by organ of origin (n = 511,365 cells). **C.** Feature plots of epithelial (KRT18, top), mesenchymal (VIM, middle), and hematopoietic (PTPRC, bottom) cell marker genes. **D.** Stacked violin plot of expression level of top marker genes identified per organ. **E.** UMAP of data colored by cell population and corresponding stacked bar chart of cell population proportion per organ. **F.** UMAP of data colored by cell type and corresponding stacked bar chart of cell type proportion per organ.

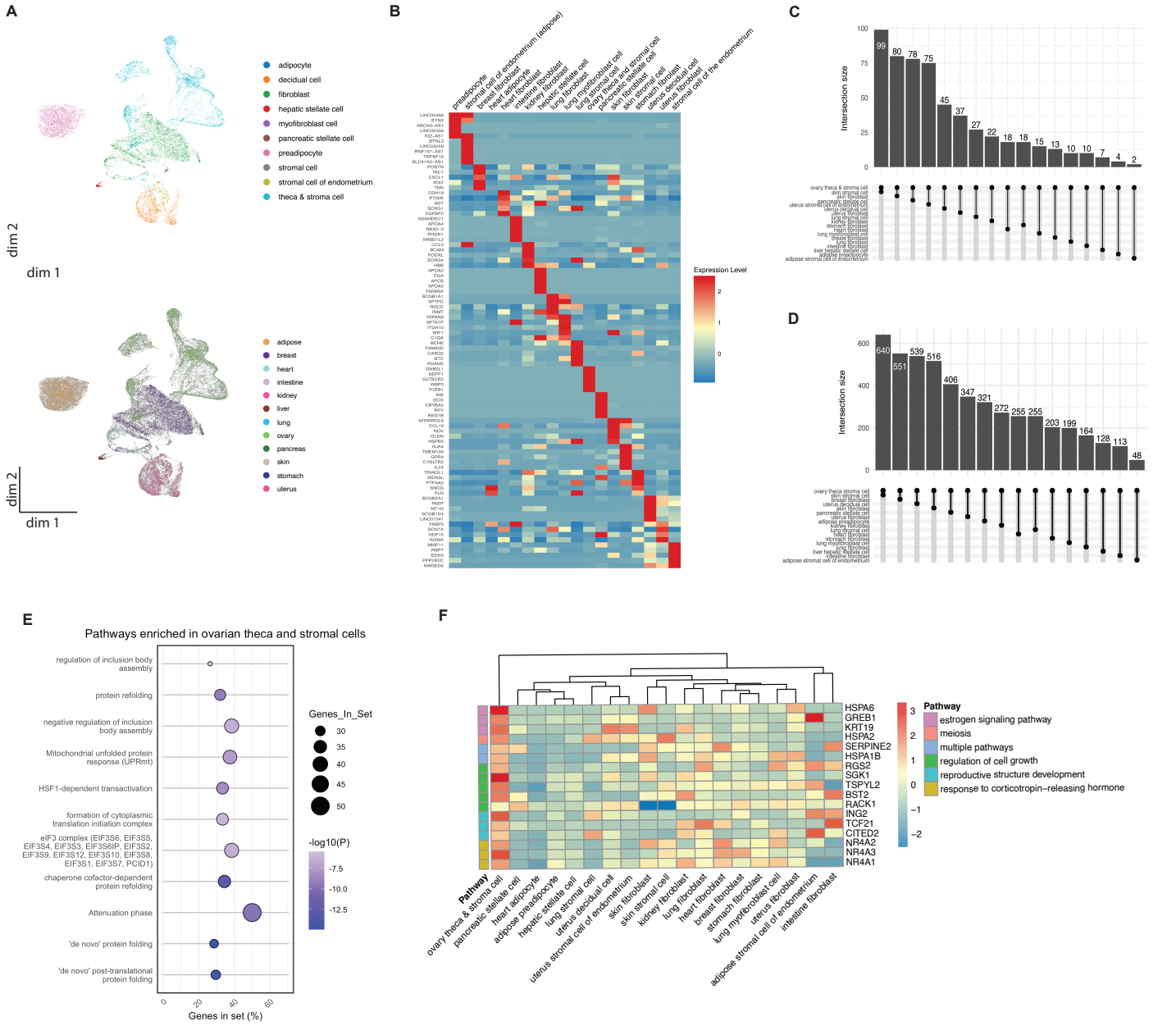

Figure 2. Characterization of stromal cell types across the PreMeno Atlas. A. UMAP of stromal cells colored by cell type (top) and organ (bottom) (n = 72,675 cells). B. Heatmap of expression of top 5 marker genes per stromal cell type. C. Upset plot of reference stromal cell marker gene intersections with ovarian theca and stromal cells. D. Upset plot of reference stromal cell GO ontology term intersections with ovarian theca and stromal cells. E. Enrichment analysis of ovarian theca and stromal cell marker genes. F. Heatmap of marker genes from select ovarian stromal pathway results plotted with expression across all stromal cell types.

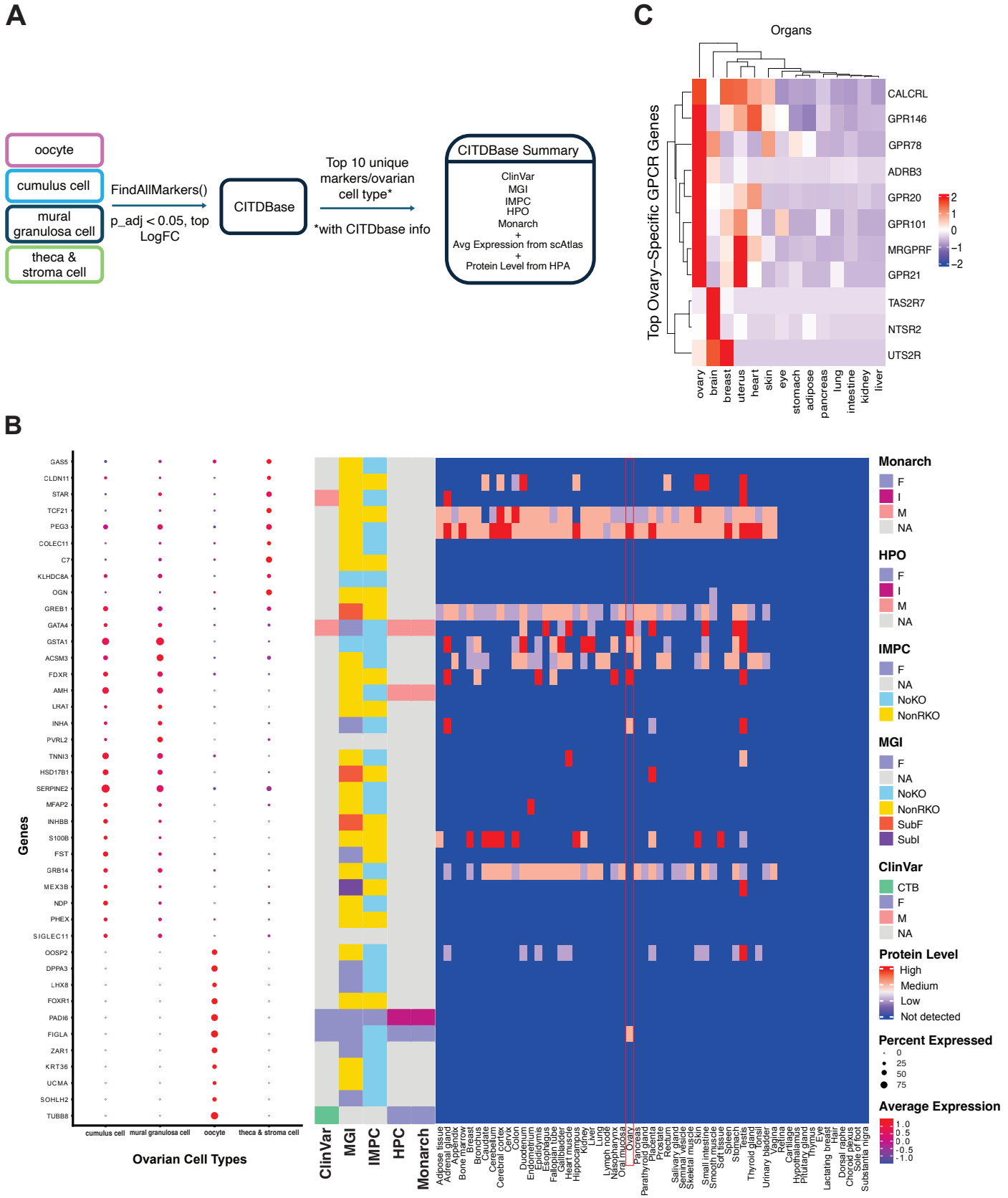

Figure 3. Contraceptive target discovery and proteomic validation using ovarian cell marker genes. A. Flow chart of ovarian cell marker gene identification and cross-database analysis. B. (Left)

Dotplot of ovarian cell marker gene expression across ovarian cell types. (Middle) Heatmap of mouse and human fertility phenotypes in CITDBase. (Right) Protein level expression of ovarian cell marker genes across 45 tissue types in Human Protein Atlas. In figure legend, CITDBase database labels are defined as follows. Monarch, HPO, and ClinVar: F = female infertility, I = non-sex specific infertility, M = male infertility, CTB = infertility in both sexes, NA = unreported. IMPC and MGI: F = female infertility, NA = unreported, NoKO = genes with no knockout, NonRKO = non-reproductive genes with a knockout, SubF = female subfertility, SubI = non-sex specific subfertility. C. Expression of ovary-enriched GPCRs across all organs.

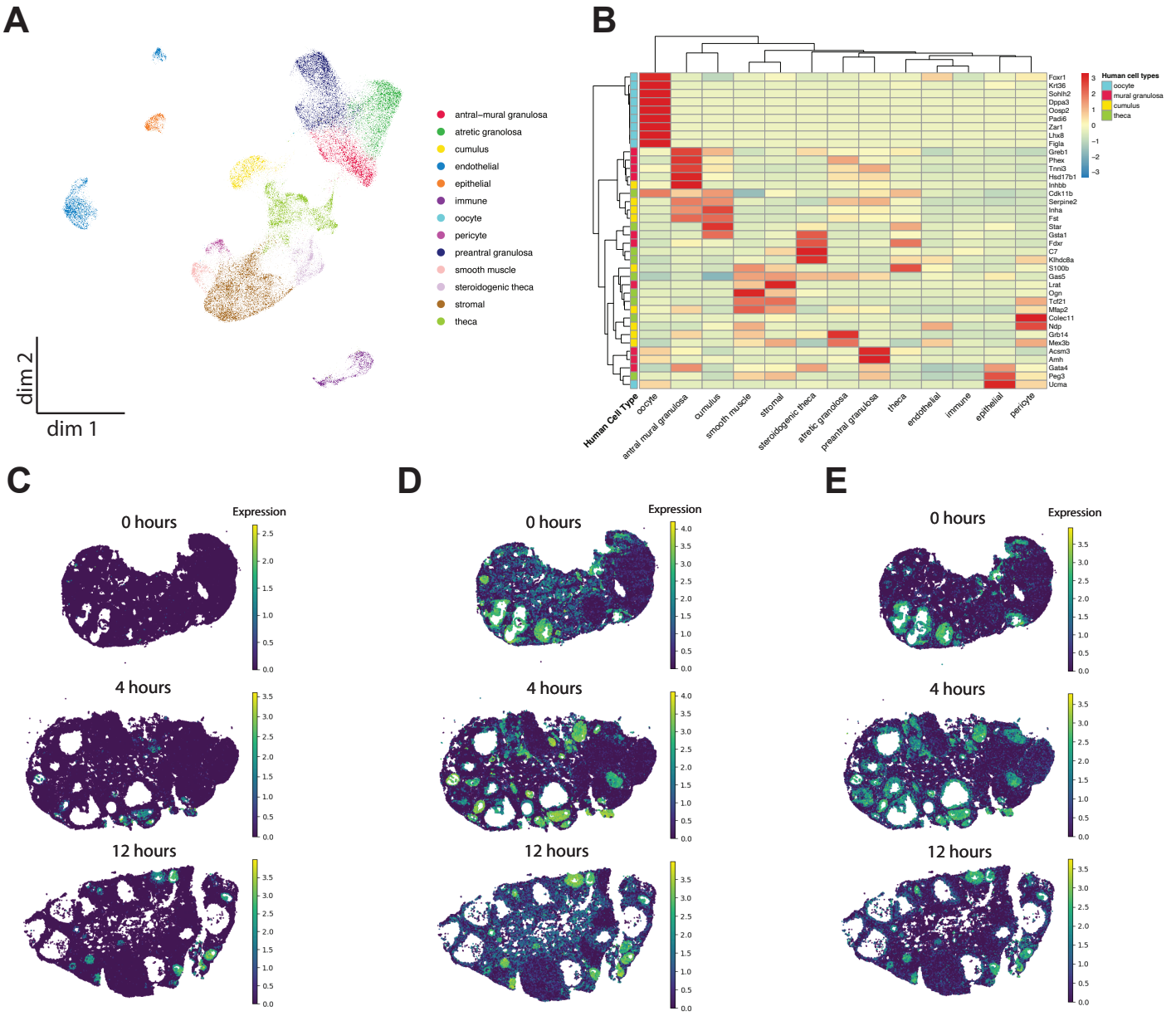

Figure 4. PreMeno Atlas comparison with harmonized single-cell mouse ovary dataset. A. UMAP of harmonized cell types in mouse ovary dataset. **B.** Expression of ovarian cell marker genes from human PreMeno Atlas across mouse ovarian cell types. Normalized expression in spatial imaging RNA-seq ovarian tissue section across an ovulation time course after stimulation with hCG (0 hours – top, 4 hours – middle, 12 hours – bottom) for **C.** *Amh*, **D.** *Inha*, and **E.** *Serpine2*. Remainder of ovary sections are shown in **Supplemental Figures 6-9**.