A comprehensive single-cell transcriptomic atlas of developing adult *Drosophila* ovary and oogenesis

Deeptiman Chatterjee[1], Allison Jevitt[2], Taylor Otwell[2], Xian-Feng Wang[1,2], Gengqiang Xie[2], Yi-Chun Huang[1,2], Wu-Min Deng[1,2]

[1]Department of Biochemistry and Molecular Biology, Tulane University School of Medicine, New Orleans, LA 70112

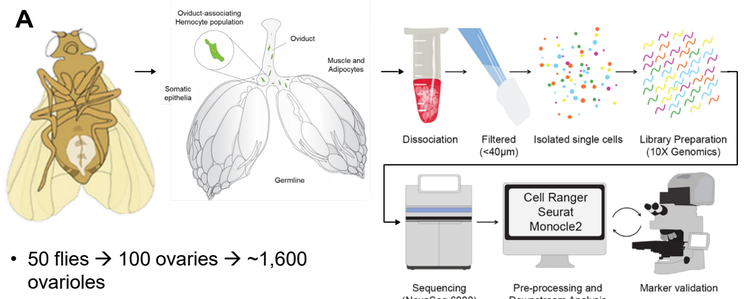
[2]Department of Biological Science, Florida State University, Tallahassee, FL 32306

DOI: 10.1371/journal.pbio.3000538 (In Press)

ABSTRACT

Oogenesis is a complex developmental process that involves spatiotemporally regulated coordination between the germline and supporting, somatic cell populations. This process has been modeled extensively using the *Drosophila* ovary. While different ovarian cell types have been identified through traditional means, the large-scale expression profiles underlying each cell type remain unknown. Using single-cell RNA sequencing technology, we have built a transcriptomic dataset for the adult *Drosophila* ovary and connected tissues. Using this dataset, we identified the transcriptional trajectory of the entire follicle cell population over the course of their development from stem cells to the oogenesis-to-ovulation transition in the corpus luteum. We further identify expression patterns during essential developmental events which take place in somatic and germline cell types such as differentiation, cell-cycle switching, migration, symmetry breaking, nurse cell engulfment, egg-shell formation, and corpus luteum signaling. Extensive experimental validation of unique expression patterns in both ovarian and nearby, non-ovarian cells also led to the identification of many new cell-type- and stage- specific markers. The inclusion of several nearby tissue types in this dataset also led to our identification of functional convergence in expression between distantly related cell types such as the immune-related genes which were similarly expressed in immune cells (hemocytes) and ovarian somatic cells (stretched cells) during their brief phagocytic role in nurse cell engulfment. Taken together, these findings provide new insight into the temporal regulation of genes in a cell-type specific manner during oogenesis and begin to reveal the relatedness in expression between cell and tissues types.

INTRODUCTION AND METHODOLOGY



- 50 flies → 100 ovaries → ~1,600 ovarioles
- Each ovariole is a factory of continuously developing egg chambers
- Egg chambers = Monolayer of follicular epithelium + Germline cells
- Papain digestion to isolate single cells
- Quality Control: Remove cells with > 1% mitochondrial gene expression, outlier values of total number of genes and transcripts (multiplets)
- An all-inclusive approach to identify every possible source of contamination to aid manual removal of contaminant cells (doublet) and ascertain fidelity of cells to fit into unique clusters
- Clusters were validated and doublets were manually removed

FIG 1. (A) Illustration of the overall workflow of scRNA-Seq sample acquisition and data analysis of adult *Drosophila* ovary and interconnecting tissues. **(B)** Schematic for the scRNA-Seq analysis pipeline for dataset pre-processing, and verification analysis.

METHODS - MANUAL DOUBLET REMOVAL

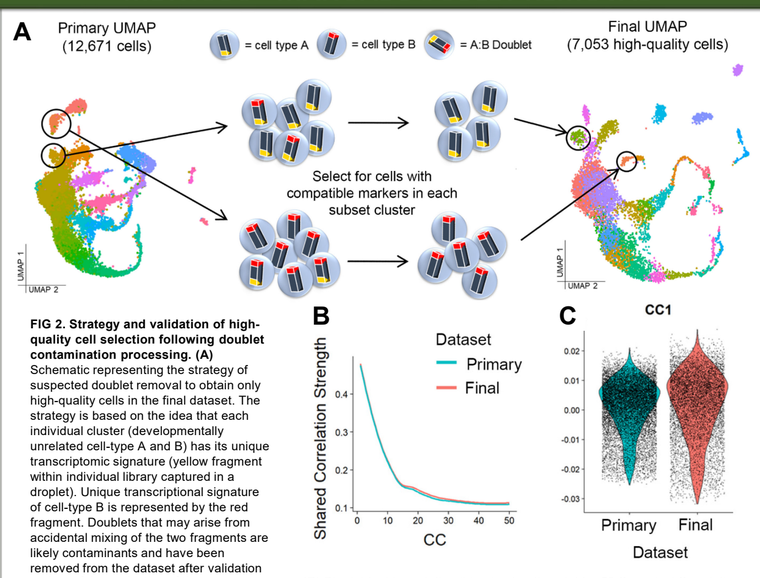


FIG 2. Strategy and validation of high-quality cell selection following doublet contamination processing. (A) Schematic representing the strategy of suspected doublet removal to obtain only high-quality cells in the final dataset. The strategy is based on the idea that each individual cluster (developmentally unrelated cell-type A and B) has its unique transcriptomic signature (yellow fragment within individual library captured in a droplet). Unique transcriptional signature of cell-type B is represented by the red fragment. Doublets that may arise from accidental mixing of the two fragments are likely contaminants and have been removed from the dataset after validation and careful examination of the genes. Individual clusters (or group of similar clusters) were selected and were cleaned for contaminating markers using this strategy, to obtain high-quality cells. **(B)** Biweight midcorrelation (bicor) saturation plot for 50 Canonical Correlation vectors (CCs) that were used to align the final and primary datasets. The two datasets are highly correlated even after stringent cleanup, indicating that no data was lost. **(C)** Violin plot to show the distribution of the canonical correlation projection vector (CC1) across the primary and the final datasets.

RESULTS - ATLAS OF *DROSOPHILA* OOGENESIS

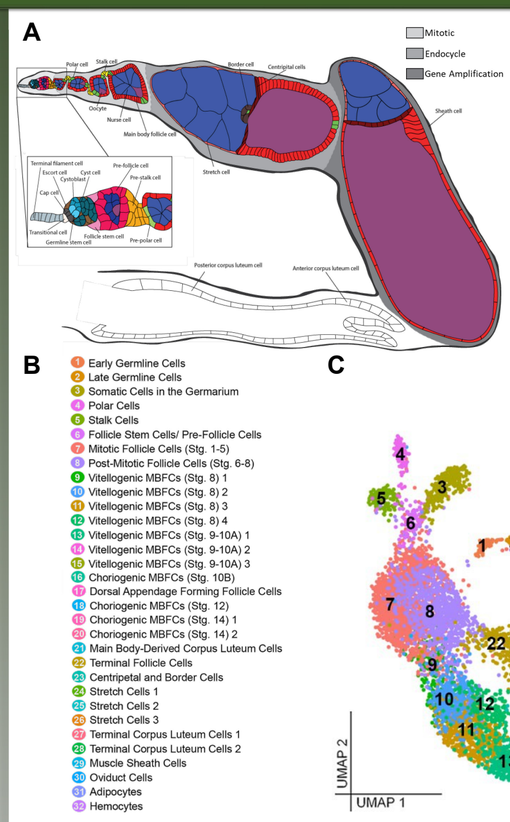


FIG 3. (A) Cell-type diversity in a single *Drosophila* ovariole (see figure 1 for reference). The entire ovariole is divided into 14 stages of development that the egg chamber undergoes. These 14 stages are grouped together into 3 major cell cycle stages, namely, the Mitosis, Endocycle and Gene Amplification stages (colored in grayscale). The anterior-most tip of the ovariole (called the germarium) contains the stem cell niche for both the somatic and germline cells. The diversity of cells in this region is showed in the inset. (B) Cell types/states assigned to different clusters. 7,053 high quality cells, clustered into 32 unique cell type identities. (C) UMAP plot of clusters of all the cells in the ovary. 429,855,892 total reads were obtained, with 28,995 mean reads per cell. Approx. 7,100 UMI and ~1,300 genes were obtained per cell. Note how cells of the same lineage are connected.

METHODS/RESULTS - MARKER VALIDATION

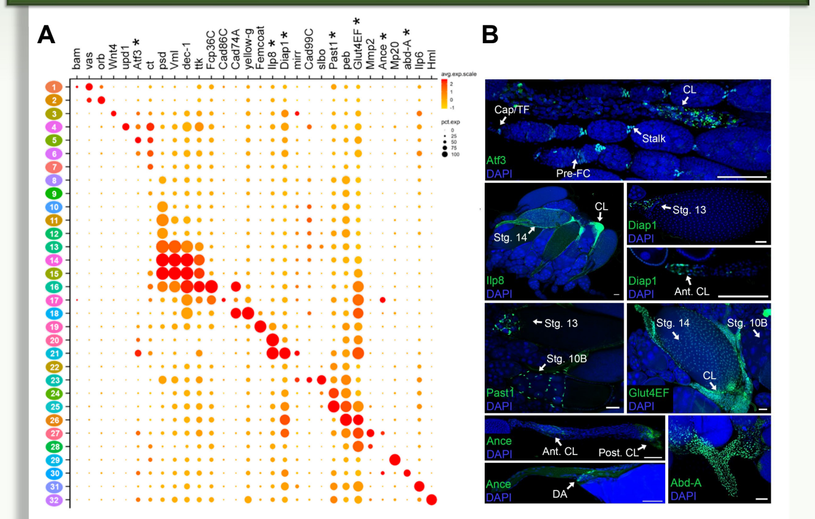


FIG 4. (A) Dot plot of identifying established marker genes for each 32 clusters. Newly identified marker genes are indicated with an asterisk (*). **(B)** Experimental validation of the 7 new marker genes shown in figure A. All expression (green) is marked using GFP-tagged proteins under endogenous control except *Ance*, marked using RFP under T2A-Gal4 control. All images are z-projections. Additional cell-type and egg chamber stage-information is indicated (Cap/TF= Cap and Terminal Filament Cells, Pre-FC= Pre-Follicle Cells, Stalk= Stalk Cells, CL= Corpus Luteum Cells, Stg.=Stage, Ant. CL= Anterior Corpus Luteum Cells, Post. CL= Posterior Corpus Luteum Cells, DA= Dorsal Appendage Forming Follicle Cells). DAPI marks nuclei. Scale bar = 50 µm.

- Known markers were used to establish cell type identities.
- Clusters without known expression data were assessed for novel markers and expression pattern for these markers was identified using GFP-tagged fluorescent markers.

RESULTS - TRANSCRIPTOMIC 'TOOL-KITS'

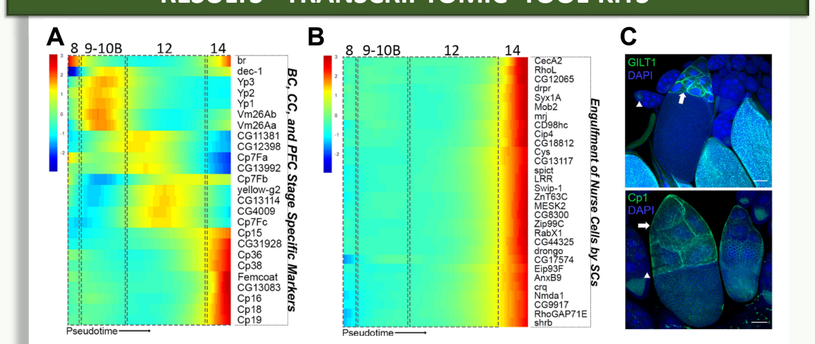


FIG 5. (A) Pseudotime-ordered heatmap of stage 8-14 specific markers from post-mitotic follicle cells (clusters 10-16, 18-19, 22-23, 26). Estimated stage boundaries (dotted boxes) are superimposed on the heatmap for reference. **(B)** Pseudotime-ordered heatmap of genes (clusters 22-23, 26) highlighting 30 (out of 79) genes that are also expressed in the hemocyte cluster. (Minimum expression in 50 cells; q < 0.05). **(C)** Experimental validation for two highly expressed genes in stretched cells (not shown in the heatmap in B) using GFP-tagged proteins under endogenous control. Arrows point to stretched cells and arrowheads point to additional expression in oocytes. All images are a single z-slice through the center of egg chambers. DAPI marks nuclei. Scale bar = 20 µm.

RESULTS - TRANCRIPTOMIC CONVERGENCE

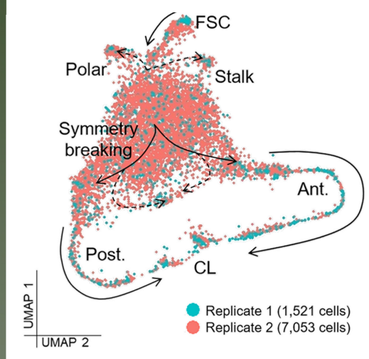


FIG 6. UMAP of somatic/follicle cell clusters from a dataset containing 2 replicates exhibit transcriptomic convergence of distant cell types, based on functional relatedness.

- Somatic follicle cells originate from the stem cell (FSC) cluster (indicated by the solid arrow) and assume polar and stalk cell fate (indicated by the dashed arrow).
- The remaining cells assume mature follicle cell fate that split into two distinct transcriptomic states (solid arrow): the anterior (Ant.) and posterior (Post.) cell types, during follicular symmetry breaking.
- Cells in the resulting Ant. and Post. trajectories partially converge (dashed arrow) to form the migratory cells, and eventually converge terminally into the corpus luteum (CL) clusters. This convergence occurs due to functional relatedness.

RESULTS - TRANSCRIPTOMIC DIVERGENCE

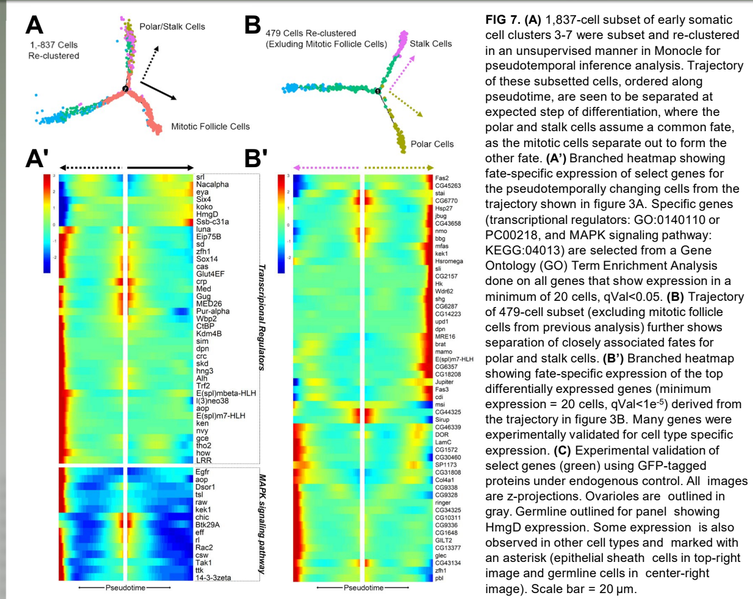


FIG 7. (A) 1,837-cell subset of early somatic cell clusters 3-7 were subset and re-clustered in an unsupervised manner in Monocle for pseudotemporal inference analysis. Trajectory of these subsetted cells, ordered along pseudotime, are seen to be separated at expected step of differentiation, where the polar and stalk cells assume a common fate, as the mitotic cells separate out to form the other fate. **(A')** Branched heatmap showing fate-specific expression of select genes for the pseudotemporally changing cells from the trajectory shown in figure 3A. Specific genes (transcriptional regulators: GO:0140110 or PC00218, and MAPK signaling pathway: KEGG:04013) are selected from a Gene Ontology (GO) Term Enrichment Analysis done on all genes that show expression in a minimum of 20 cells, qVal<0.05. **(B)** Trajectory of 479-cell subset (excluding mitotic follicle cells from previous analysis) further shows separation of closely associated fates for polar and stalk cells. **(B')** Branched heatmap showing fate-specific expression of the top differentially expressed genes (minimum expression = 20 cells, qVal<1e^{-5}) derived from the trajectory in figure 3B. Many genes were experimentally validated for cell type specific expression. **(C)** Experimental validation of select genes (green) using GFP-tagged proteins under endogenous control. All images are z-projections. Ovarioles are outlined in gray. Germline outlined for panel showing HmgD expression. Some expression is also observed in other cell types and marked with an asterisk (epithelial sheath cells in top-right image and germline cells in center-right image). Scale bar = 20 µm.

- Pseudotemporal Inference allowed us to investigate known fate change events for novel transcriptomic divergence.
- Transcriptomic convergence and divergence is driven by cellular function.

HIGHLIGHTS

- Identification of the transcriptomic basis for early differentiation of polar and stalk cells from the main body follicle cells, mitosis-to-endocycle switch, and follicular symmetry breaking.
- Identification of transcriptomic signatures of different follicle cell groups that carry out important developmental functions such as migration, engulfment of nurse cells, and eggshell formation.
- Identification of novel ovulation-related markers in late-stage follicle cells (termed pre-corpus luteum) as they undergo the developmental switch from oogenesis to ovulation regulation.
- Functional convergence of distant cell types. For example, nurse-cell engulfing stretched cells express genes that are shared by the hemocytes, and the transcriptomic signature of cells in the corpus luteum overlaps with that of the cells in the oviduct and the hemocytes.
- Assembly of a high-quality *Drosophila* oogenesis transcriptomic dataset.

ACKNOWLEDGEMENTS

- Dr. Julia Wang (FSU Med School)
- Dr. Michelle Arbeitman (FSU Med School)
- Dr. Brian Oliver (NIH)

FSU Translational Core
FSU FlyMasters
FSU The CMB Graduate Association (TCGA)

National Institute of General Medical Sciences
National Institutes of Health