



## **Supplementary Information for**

### **Cellular and transcriptional diversity over the course of human lactation**

Sarah K. Nyquist<sup>1,2,3,4</sup>, Patricia Gao<sup>3</sup>, Tessa K. J. Haining<sup>3</sup>, Michael R. Retchin<sup>3</sup>, Yarden Golan<sup>5</sup>, Riley S. Drake<sup>1,3,6</sup>, Kellie Kolb<sup>1,3</sup>, Benjamin E. Mead<sup>1,3</sup>, Nadav Ahituv<sup>5</sup>, Micaela E. Martinez<sup>7</sup>, Alex K. Shalek<sup>1,2,3,6,8,9,10+</sup>, Bonnie Berger<sup>1,4+</sup>, Brittany A. Goods<sup>11,+\*</sup>

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, MA, 02142

<sup>2</sup>Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>3</sup>Department of Chemistry and Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>4</sup>Computer Science and Artificial Intelligence Laboratory and Department of Mathematics, MIT, Cambridge, Massachusetts, 02139

<sup>5</sup>Department of Bioengineering and Therapeutic Sciences and Institute for Human Genetics UCSF, University of California San Francisco, San Francisco, CA 94143

<sup>6</sup>Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139

<sup>7</sup>Department of Biology, Emory University, Atlanta, GA 30322

<sup>8</sup>Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>9</sup>Division of Health Science & Technology, Harvard Medical School, Boston, MA 02115

<sup>10</sup>Department of Immunology, Massachusetts General Hospital, Boston, MA 02114

<sup>11</sup>Thayer School of Engineering and Program in Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH 03755

<sup>+</sup>equal

\*Bonnie Berger, Brittany A. Goods, Alex K. Shalek

**This PDF file includes:**

Supplementary text  
Figures S1 to S6  
Tables S1, S2  
Legends for Datasets S1 to S9  
SI References

**Other supplementary materials for this manuscript include the following:**

Datasets S1 to S9

## **Supplementary Information Text**

### **Supplementary Methods**

#### **Epithelial cell clustering**

Our goals in describing the epithelial cells are to 1) identify distinct functional subclusters involved in different roles during lactation, 2) ask if the proportional abundance of any of these subclusters changed over the course of lactation, and 3) identify any transcriptional changes within any subcluster over the course of lactation which might facilitate an understanding of the functional shifts in these subclusters.

To choose the representation of the data that best answers those questions, multiple Leiden clustering parameters were considered, and the resolution was selected to maximize the number of clusters with distinct marker genes and enriched GO terms and reduce the amount of post-clustering merging required to identify these distinct genes (Fig. S8A,B,J, S9). We note that within the large LC2 celltype, several cluster boundaries were relatively stable at multiple clustering resolutions, but several shifted inconsistently at each resolution, grouping different subsets of cells together (Fig. S7A). Cells in clusters 0, 1, 2, and 3 were unstable across clustering parameters, so these clusters were merged into the secretory lactocyte cluster due to shared expression of various canonical lactation-related genes (*CHRDL2*, *FASN* *CEL*) with few other discriminating markers (Fig. S8G,J). These four clusters changed in abundance over time (milk stages) (Fig. S8E), but because of the lack of discriminating markers, we determined that inference of functional changes over time would be better facilitated with direct regression of gene expression over time across all four clusters instead of by attempting to describe four different subclusters of secretory lactocytes which changed in abundance over time. Despite many shared functions with clusters 0, 1, 2, and 3, cluster 5 was left as its own cluster due to high mitochondrial gene percentage and its consistent separation across clustering resolutions (Fig. S8H,I). Clusters 9, 6 and 8 shared a distinct transcriptional signature and were merged into the LC1 cluster. Clusters 4 and 11 were merged into a single *KRT* high 1 cluster due to cluster 11's specificity to a single donor, and cluster 7 remained as a single *KRT* high 2 cluster. Additionally, these clusters each contained cells from multiple donors and were robust to leave-one-donor-out clustering (Fig. S8F).

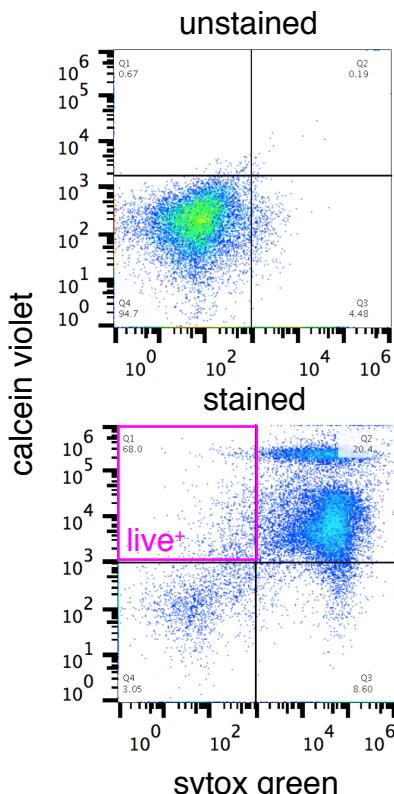
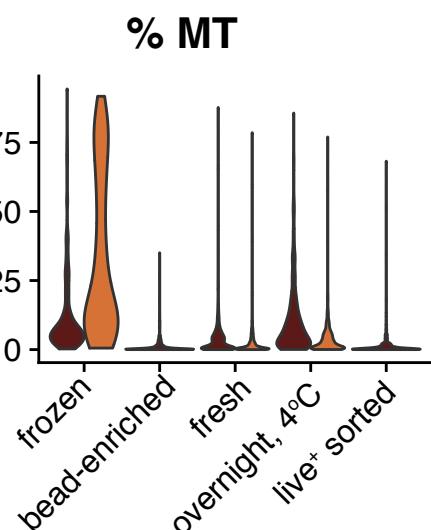
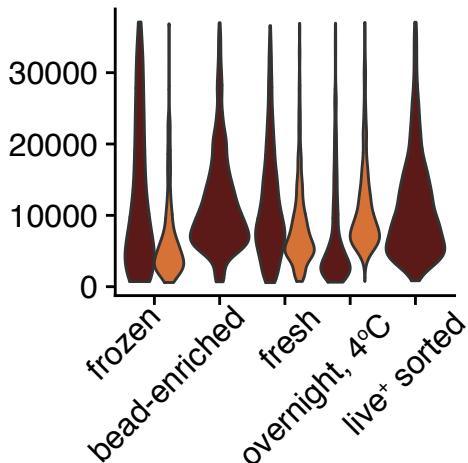
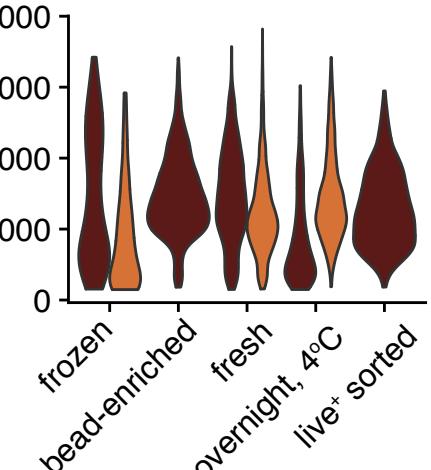
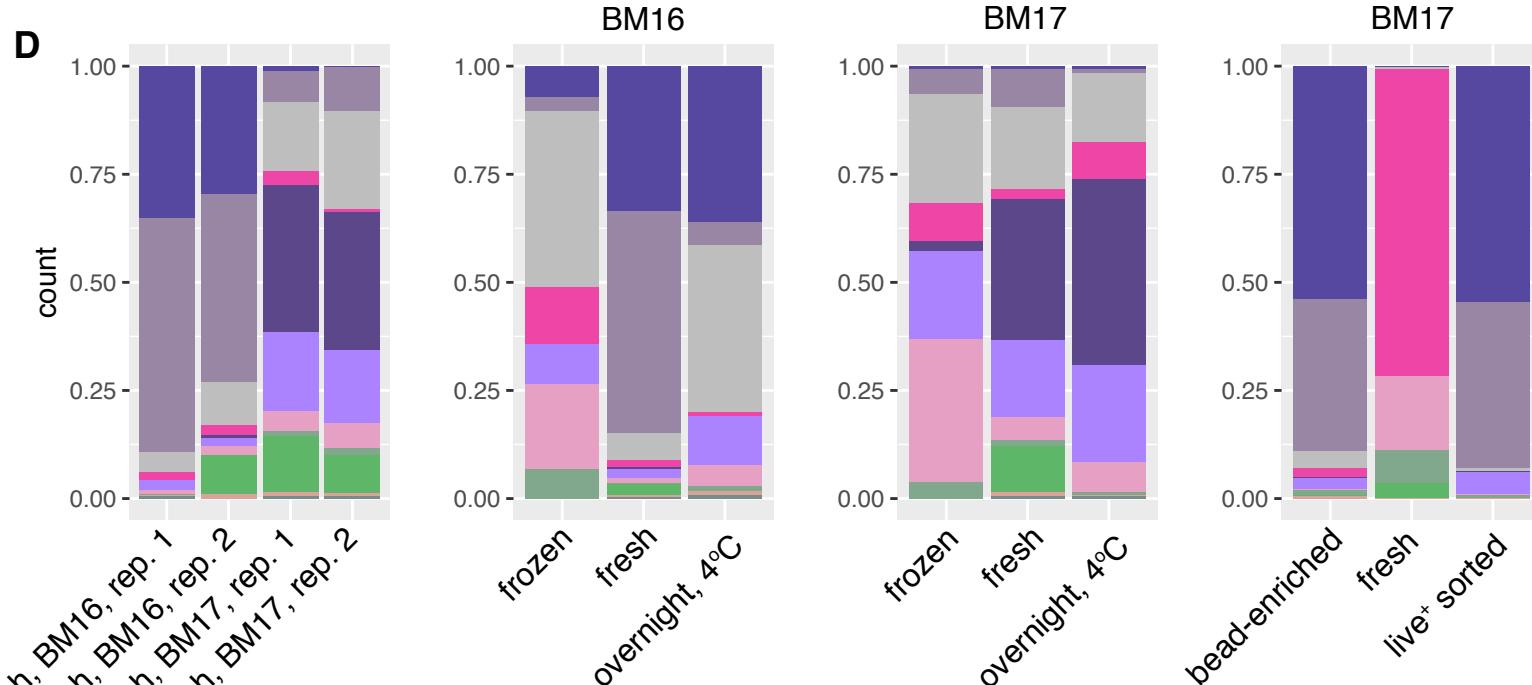
#### **Leave one donor out clustering of macrophage cells**

Because some clusters were dominated by individual donors, we sought to ensure that single no single donor was driving the biological conclusions from our clustering results by performing leave-one-donor-out clustering analysis on the macrophage subset. We repeated the variable gene identification (top 2,000) and clustering at resolution=0.2, as was done on the full macrophage dataset, on the macrophage cells with each donor excluded. We compared clustering assignments of the remaining cells to the clusters identified on all donors using the Adjusted Rand Index, a metric which measures the similarity between two cluster assignments ignoring permutations where a score of 0 indicates a random permutation and a score of 1.0 indicates an exact match. We found that leaving out any donor except donor BM05 resulted in an adjusted Rand index of 0.9-0.98 indicating a high consensus regardless of the inclusion of an individual donor. We further examined the clustering that excluded donor BM05 further to understand the impact of including this donor on the results. BM05 is the donor with the highest number of cells in the macrophage cluster, most of which were assigned to sub-cluster 0 (Fig. S6A). To ensure that no biological conclusions depended on the inclusion of this donor, we repeated the marker gene identification (Fig. S6B) with this donor excluded as well as comparisons of M1 and M2 scores across the 5 clusters identified (Fig. S6C) and found no differences which would impact the biological interpretation of the results.

### **GO Term reduction for functional enrichment of epithelial cells**

Functional enrichment analysis on top epithelial cell subcluster marker genes was performed using Enrichr using the gseapy package with the gene set GO\_Biological\_Processes\_2021 (1,2). Due to the hierarchical structure of the GO database and the overlapping functions of many of the marker genes of the epithelial cell subclusters, representative GO terms were identified through a series of filtering and curating steps. For each subcluster, significantly enriched terms were grouped based on shared marker genes found to be overlapping with the GO term. These grouped terms were further grouped between subclusters based on shared term ID or shared genes. The mean gene set score was calculated for each epithelial cell group and enriched GO term using the Scanpy function “score\_genes” (3). For each group of GO terms, the terms with the highest variance of mean gene scores across epithelial subgroup was chosen such that each epithelial subgroup had between 7 and 15 GO terms for which they had the maximum mean gene score. To avoid redundant terms, GO terms were also merged based on high overlap of genes in the full reference GO term gene list.

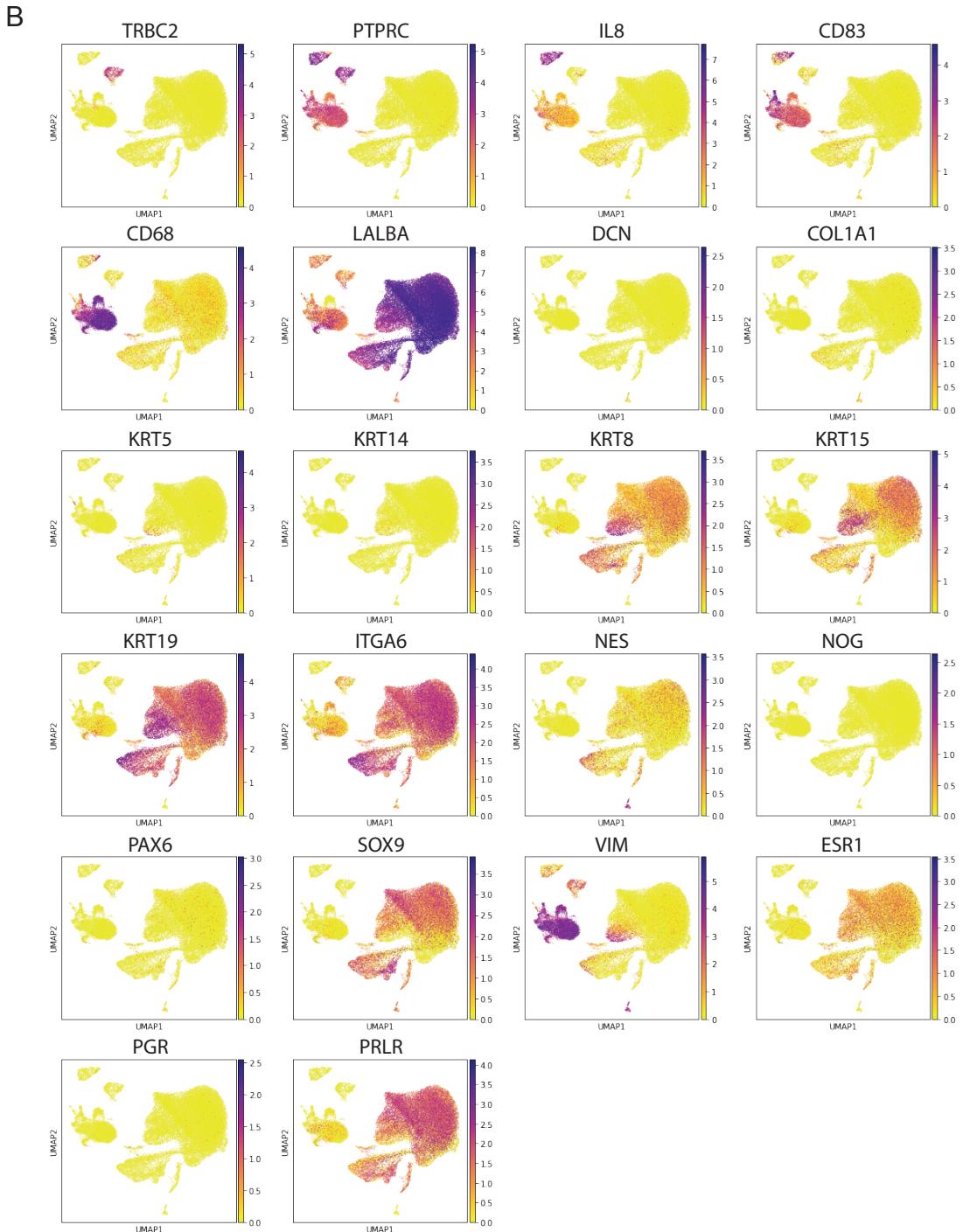
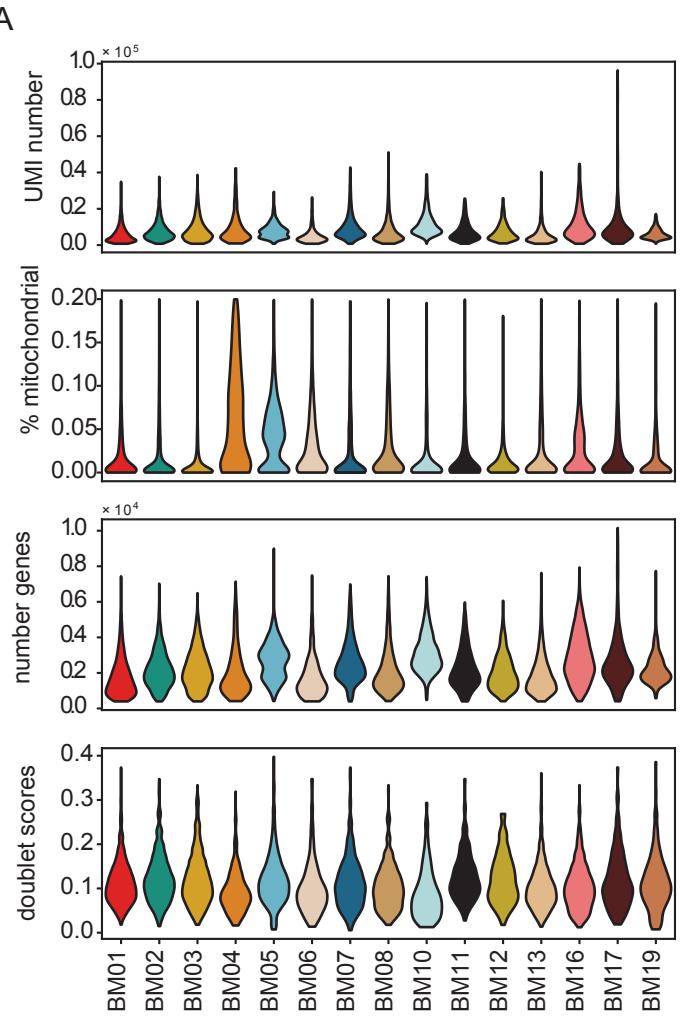
Time-dependent enriched GO terms were identified for genes positively and negatively associated with time postpartum separately and for both LC1 epithelial and secretory lactocyte clusters. These GO terms were similarly curated with an additional filtering step of correlation of the gene set scores over time postpartum in the same direction as the set of differential genes used (e.g. positive correlation for GO terms enriched in the gene list increasing with time).

**A****B****nCount RNA****nFeature RNA****D**

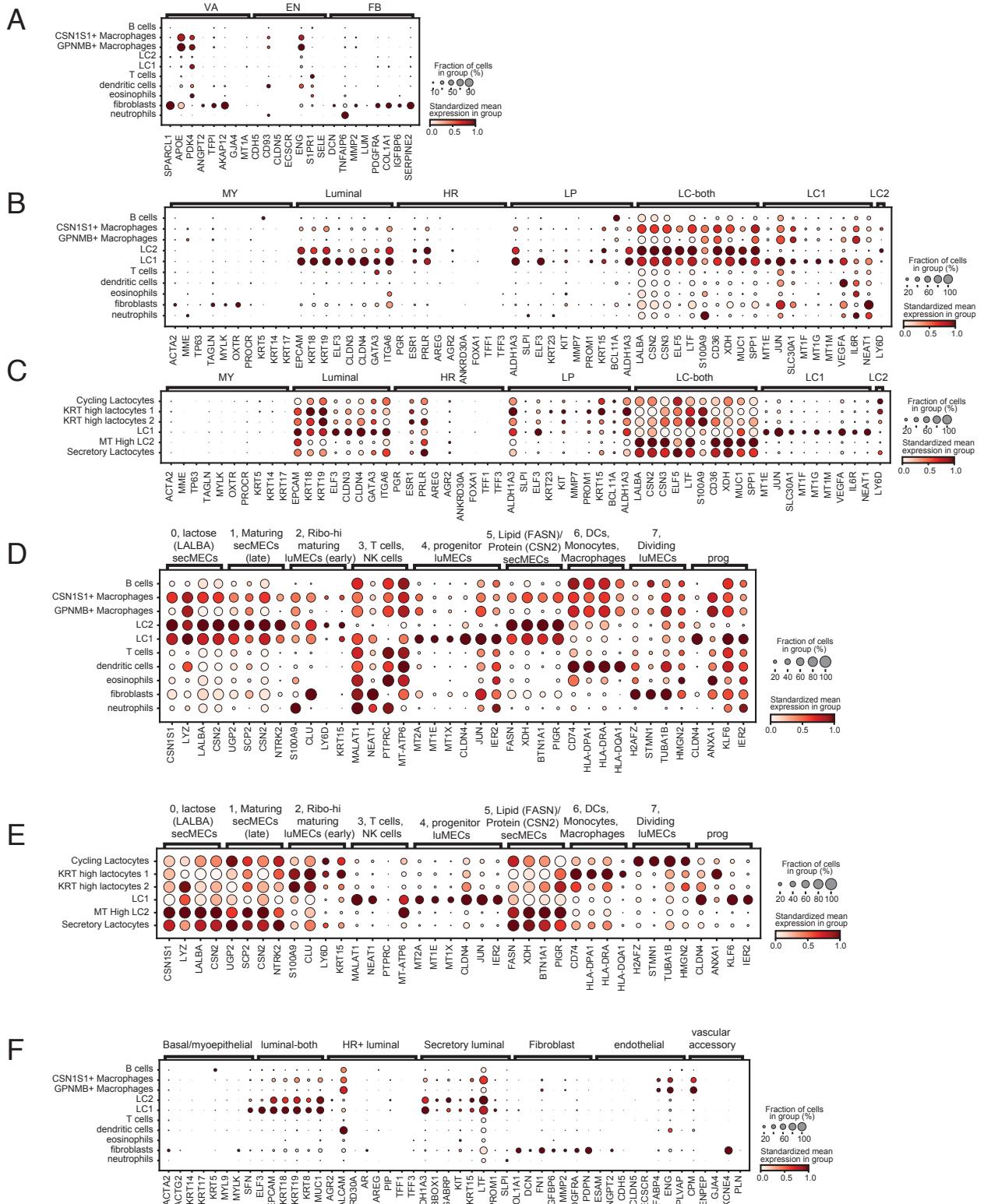
cell\_IDs

- luminal lactocyte 1
- luminal 2
- luminal 2 MHigh
- macrophages
- luminal lactocyte 2
- luminal lactocyte 3
- milk macrophage
- T cells
- fibroblasts
- neutrophils
- luminal 2 proliferating

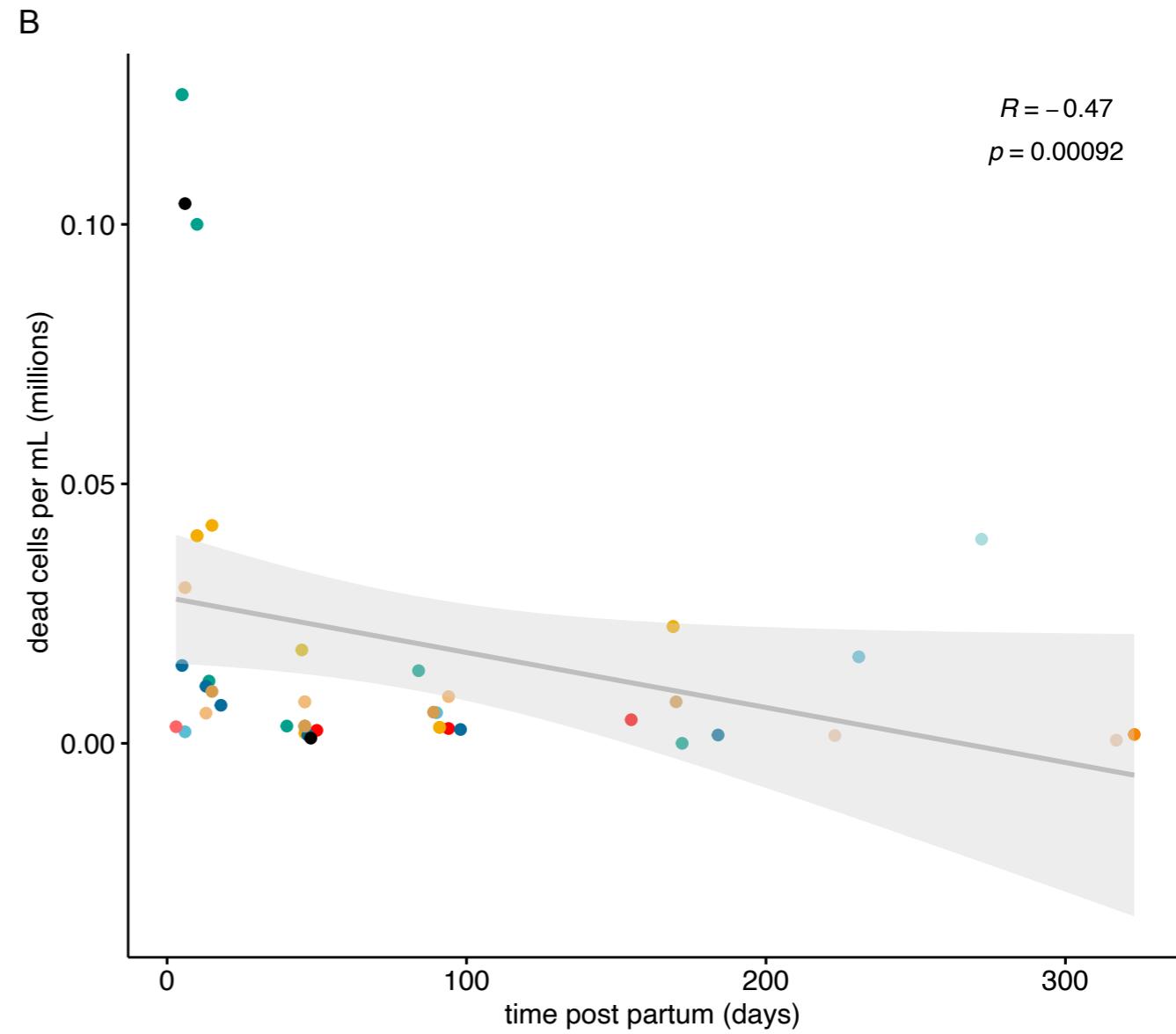
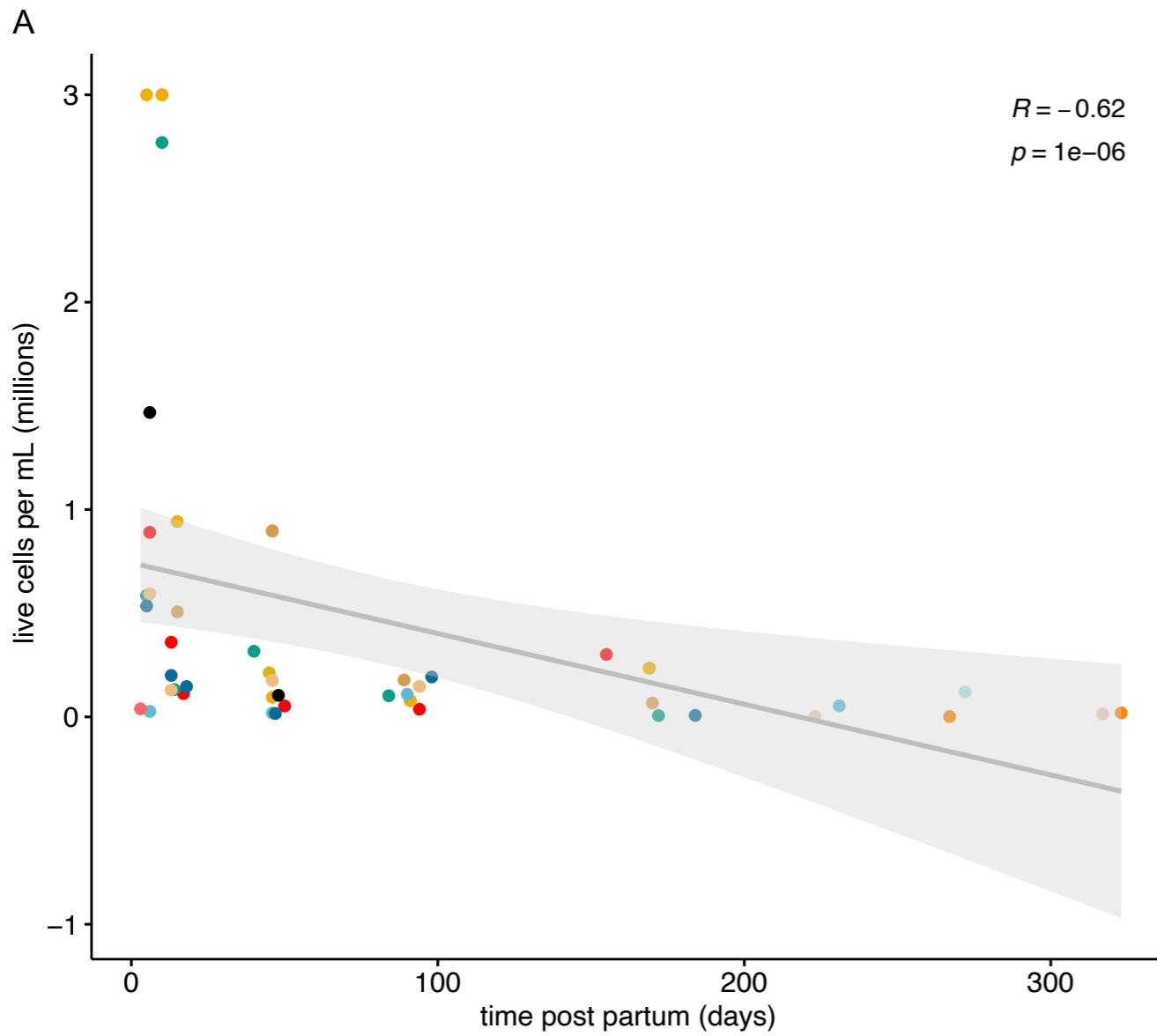
**Fig. S1. Comparison of cell isolation methods from human breast milk.** **A.** Representative flow sorting plot for live cells (calcein violet+sytox-). **B.** Violin plots of quality control metrics (%MT, nCount RNA, nFeature RNA) for each processing condition across two donors. **C.** UMAPs of combined data colored by processing method, donor, and cell identity. **D.** Stacked bar charts showing relative frequencies of cell types for each indicated condition.



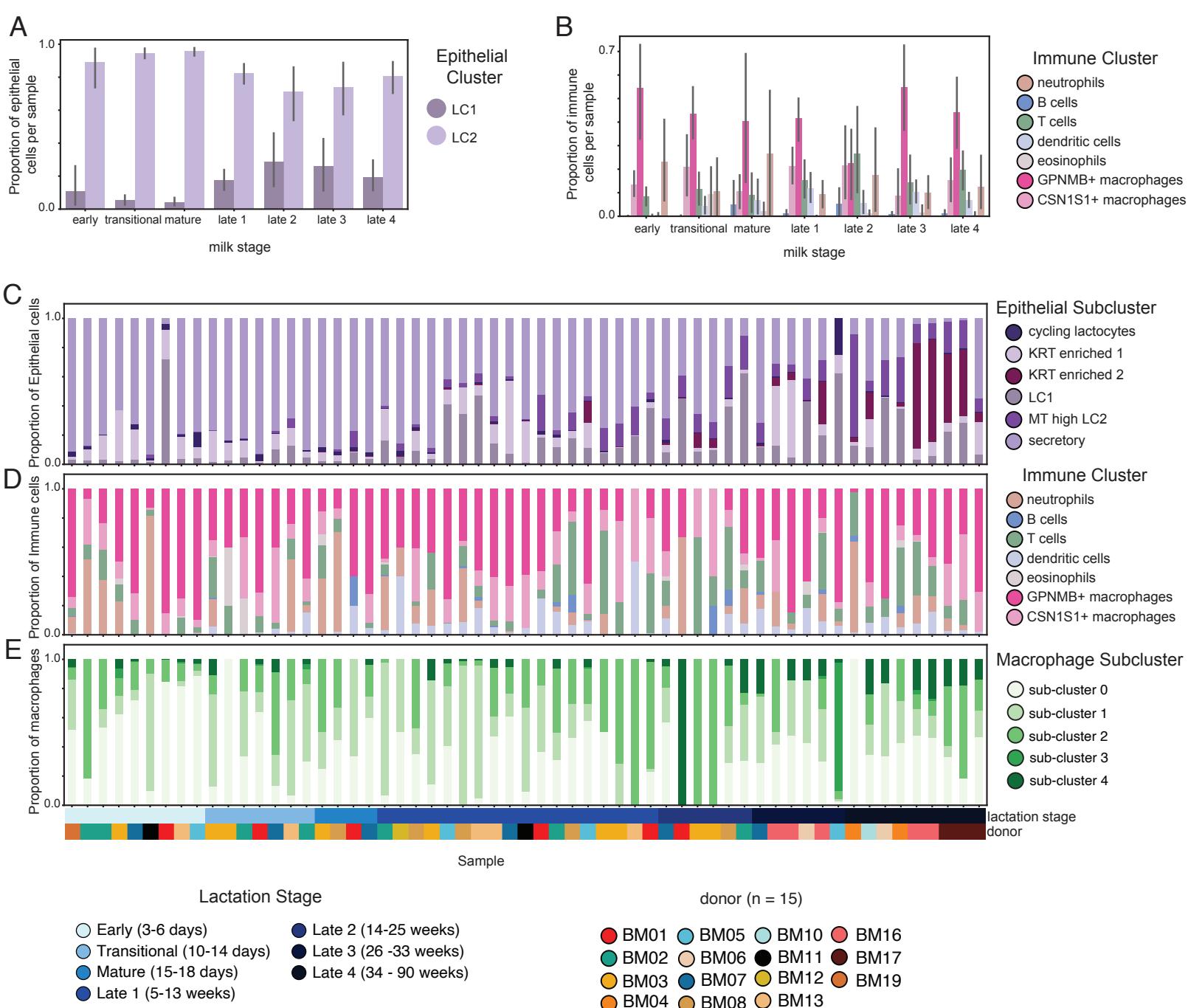
**Fig. S2.** Sequencing quality control and feature plots. **A.** Violin plots of quality control metrics for each donor in the cohort. **B.** Feature plots showing expression of select genes.



**Fig. S3: Comparisons to existing MEC dataset cluster marker genes:** Each plot shows gene expression of cells and clusters from this study (rows) using marker gene and cluster labels from previous studies (columns) **A.** Dotplot of marker genes identified in stromal cell clusters in non-lactating breast tissue in (4) grouped by major celltype cluster. VA: vascular accessory, EN: endothelial, FB: fibroblast. **B.** Dotplot of marker genes identified in epithelial cell clusters in both non-lactating breast tissue and hBM in (4) grouped by major celltype cluster (**B**) and epithelial subcluster (**C**). **D.** Dotplot of marker genes for clusters identified in scRNA-Seq in hBM in (5) plotted across all major celltypes (**D**) and epithelial subclusters (**E**). **F.** Dotplot of marker genes identified in *EPCAM*<sup>+</sup> cells from non-lactating breast tissue in (6) plotted across major celltype clusters.

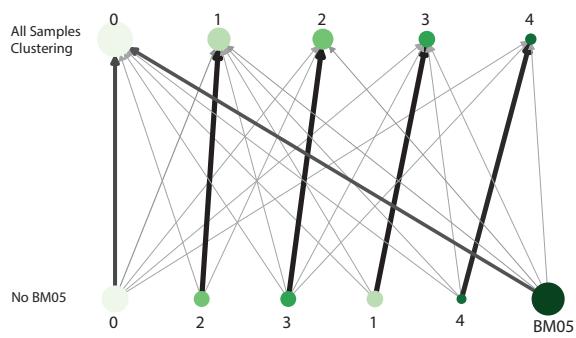


**Fig. S4. Cell counts over time in collected hBM samples. A.** Live cells per mL and **B.** dead cells per mL by time postpartum in the cohort.

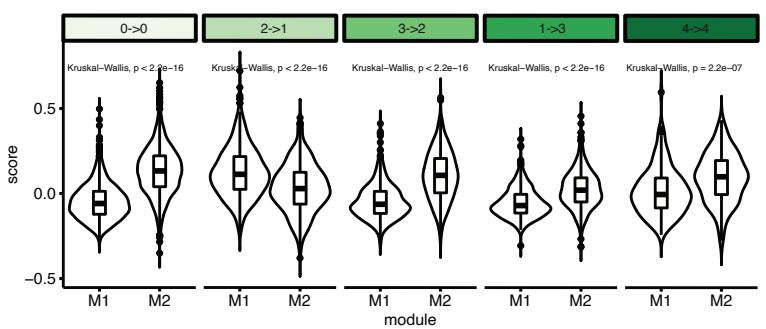


**Fig. S5: Celltype subset compositions across time and samples.** **A,B.** Proportions of each major epithelial cluster (**A**) and immune cluster (**B**) per sample, split by milk stage. Error bars show standard deviation. **C-E.** Frequency of cell types identified for each sample within epithelial subclusters (**C**), immune cell subclusters (**D**), and macrophage subclusters (**E**). Lactation stage and donor number are represented by colors below the stacked bar plots

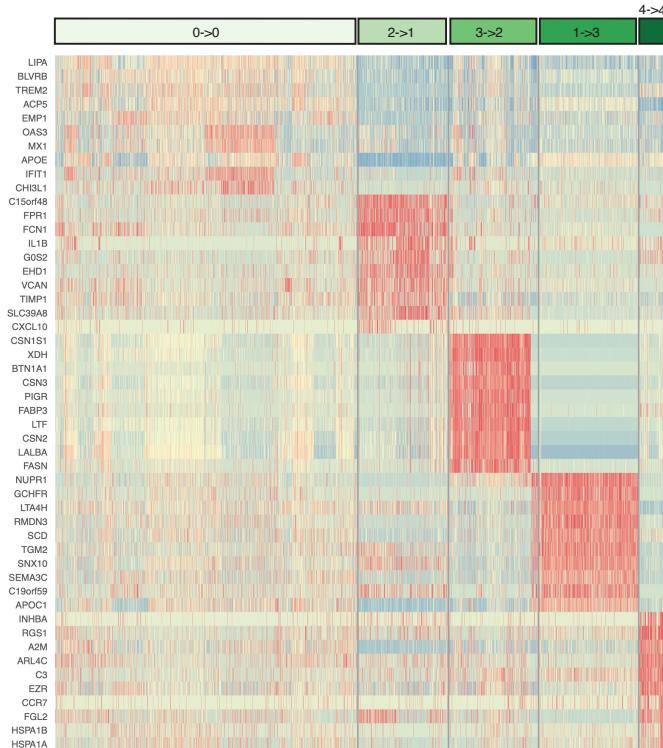
A.



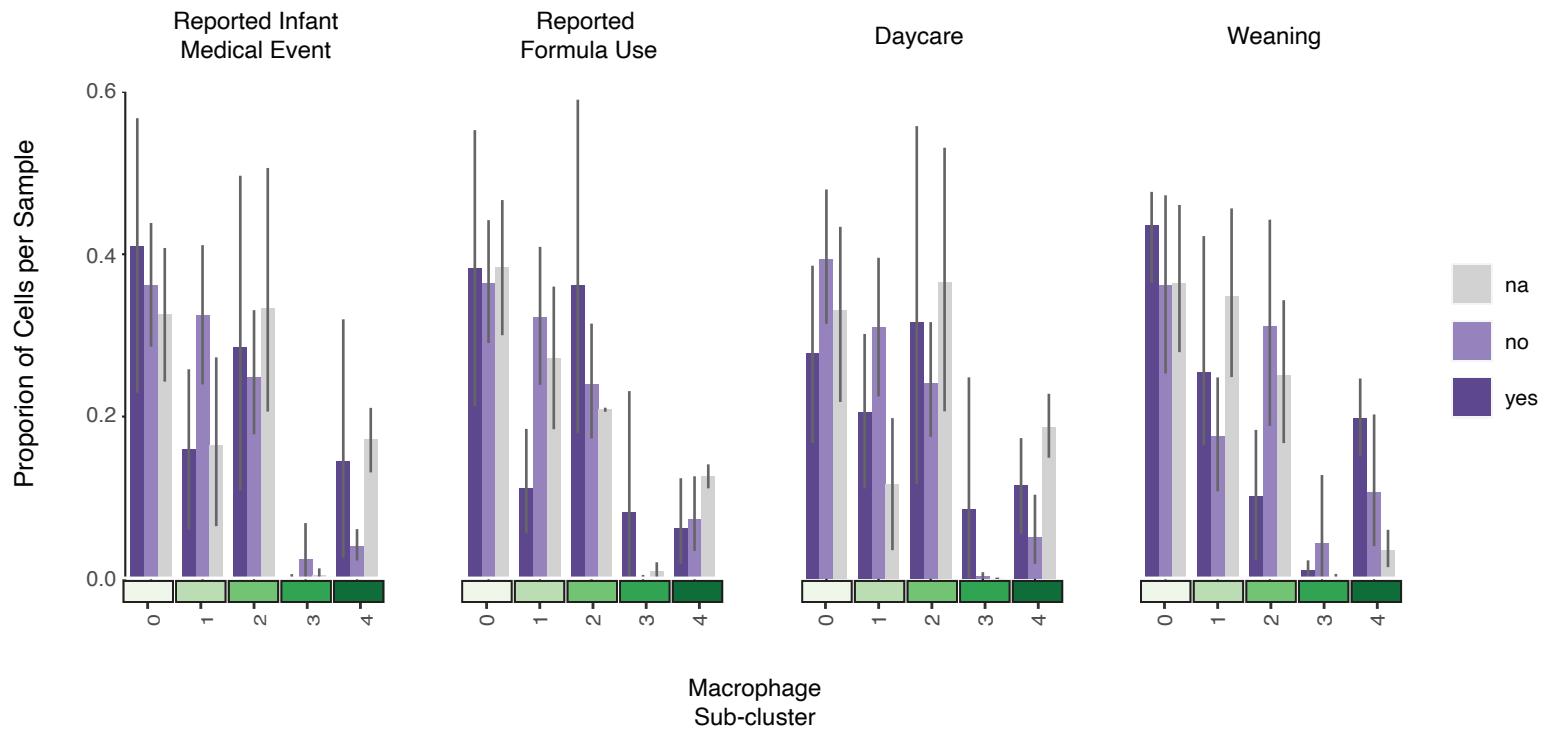
C.



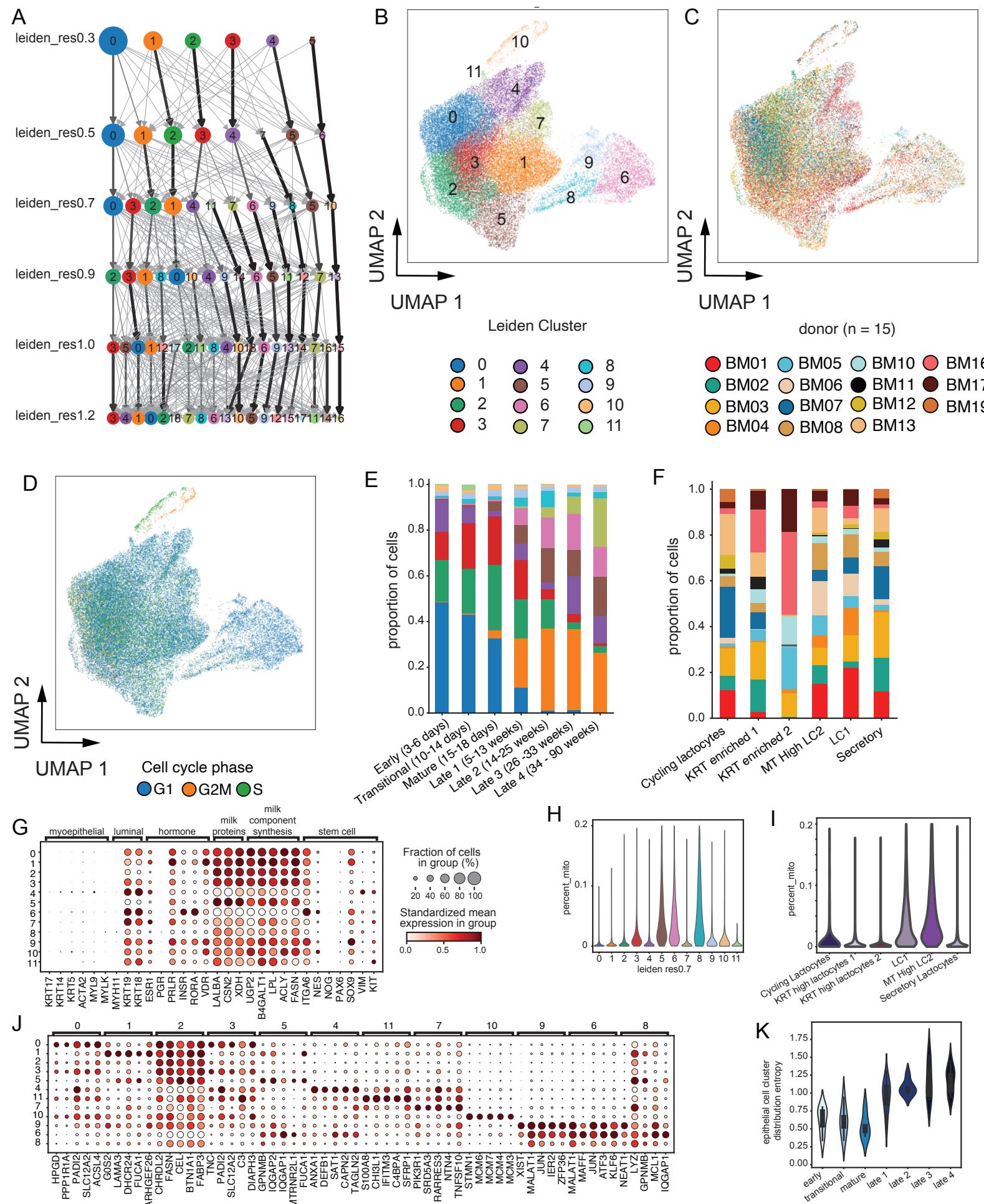
B.



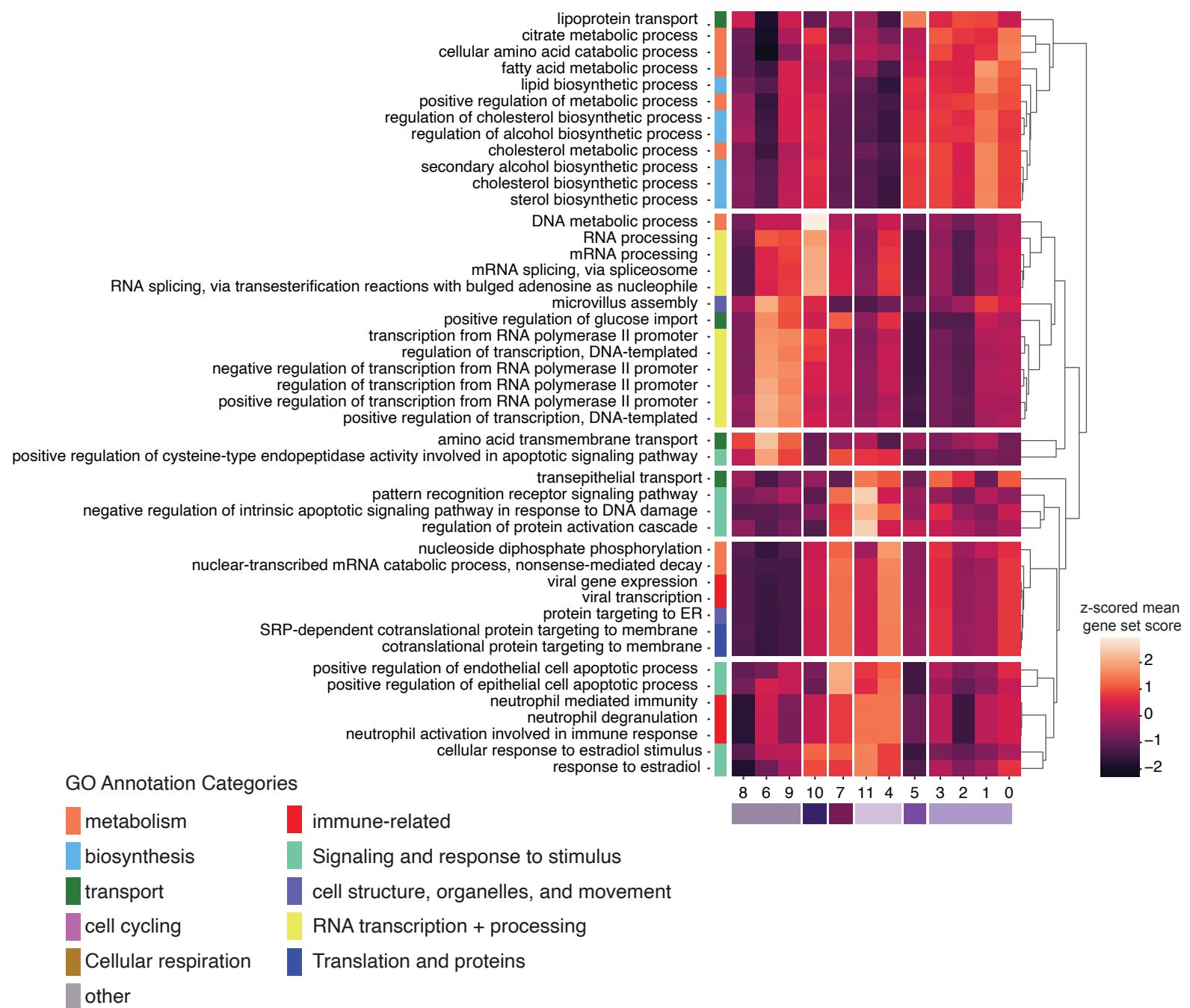
**Fig. S6: Macrophage sub-clustering with donor BM05 removed:** **A.** Cell cluster reassignment by upon re-clustering without donor BM05. Circle sizes are proportional to proportion of cells in each cluster. Top row represents clusters computed on macrophages from all donors, bottom row represents clusters computed on macrophages from all donors except BM05 with cells from donor BM05 represented as the final circle. Edge weights represent the proportion of cells from BM05-excluded clusters which are assigned to each of the all-sample clusters. **B.** Top marker genes for each BM05-removed cluster, colored by the all-sample cluster which contains the highest proportion of those cells and for which the marker genes are most similar. Cell clusters labeled along the top of heatmap labeled with the format “BM05 removed cluster name”->“All samples cluster name”. **C.** M1 and M2 module scores on cells from each BM05 removed clusters, labeled as in **(B)**.



**Fig. S7: Macrophage sub-cluster metadata comparisons:** Composition of each sub-cluster as a function of infant medical events, formula use, daycare, weaning status.

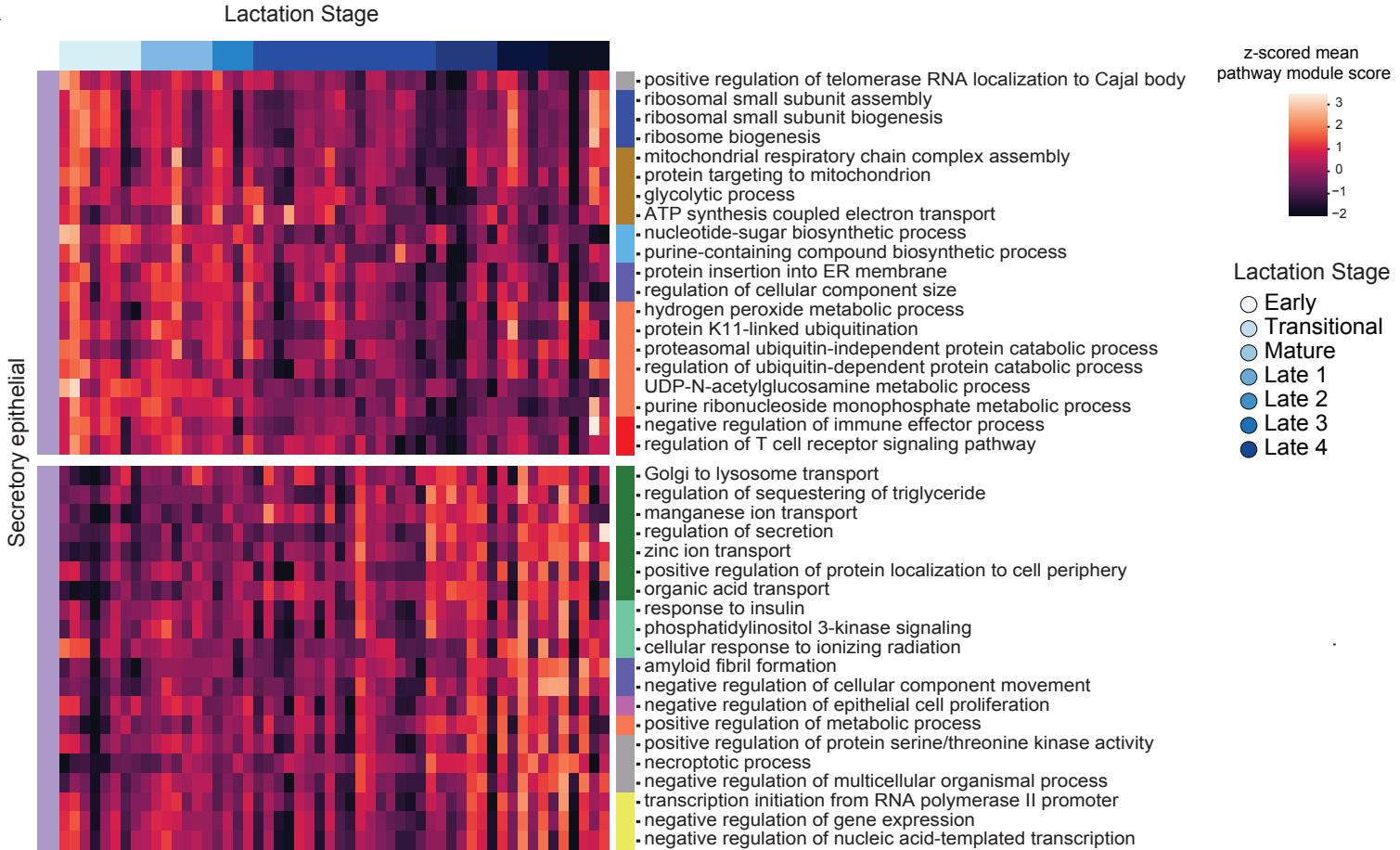


**Fig. S8. Epithelial sub-clustering** **A.** Cell cluster reassignment of different Leiden clustering resolutions (labeled on left). Circle sizes are proportional to proportion of cells in each cluster. Edge weights and opacity represent the proportion of cells from lower clustering resolution which are assigned clusters in higher clustering resolution. **B.** UMAP visualizing epithelial cells colored by Leiden (resolution =0.7) clusters **C.** Donor **D.** and computationally defined cell cycle score. **E.** Proportion of cells during each milk stage colored by Leiden cluster. **F.** Proportion of cells in each epithelial sub-group originating from each donor. **G.** Mean expression in cell subset standardized within genes (color) and percent of cells expression (dot size) of canonical mammary epithelial marker genes and **J.** pseudobulk marker genes for each Leiden cluster **H.** Percent mitochondrial gene expression of each Leiden cluster and consensus cluster (**I**). **K.** Entropy of epithelial cell distribution per sample across each milk stage.

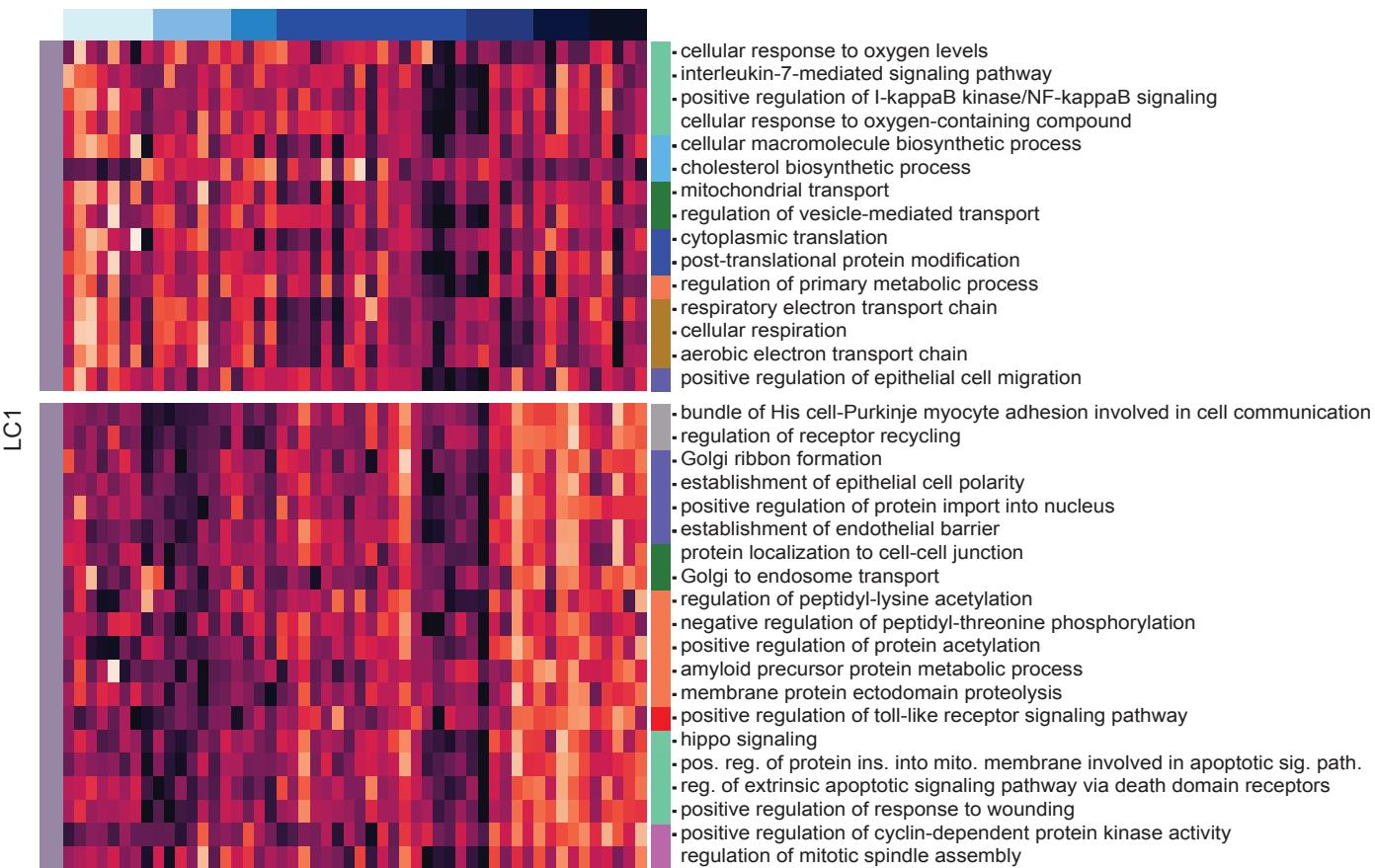


**Fig. S9. Epithelial sub-cluster GO heatmap.** Reduced top Enrichr results from the gene ontology biological processes 2021 database on the marker genes for each Leiden clusters, colored by the mean gene set score for all genes in that pathway on cells in that subgroup, scaled by a z-score across subgroups. Leiden clusters ordered by the sub-group which they were merged into with colors corresponding to sub-group identity.

A



B



**Fig. S10.** **A.** Reduced top Enrichr GO biological process results on genes changing over time in secretory epithelial cluster and **B.** LC1 epithelial cluster heatmaps represent sample means of gene set scores of each pathway (rows) z-scored across samples (columns), samples ordered by increasing time postpartum. Pathways colored by curated related GO term classifications, full results in Dataset S9.

**Table S1:** Donor information

<b>Donor</b>	<b>Age</b>	<b>Week Delivered</b>	<b>Labor Induced</b>	<b>Antibiotics During Delivery</b>	<b>Pregnancy Number</b>	<b>Delivery Mode</b>	<b>Race</b>	<b>height (from forms)</b>	<b>height (inches)</b>	<b>height (m)</b>	<b>estimated weight over the course of the study (lbs)</b>	<b>estimated pre-pregnancy weight (lbs)</b>	<b>estimated weight over the course of the study (kg)</b>	<b>estimated pre-pregnancy weight (kg)</b>	<b>BMI (over course of study)</b>	<b>BMI (pre-pregnancy)</b>
BM01	31	42	N	N	1	vaginal	caucasian	5'2"	62	1.57	130	121	59.04	54.96	23.81	22.16
BM02	31	42	Y	N	1	vaginal	caucasian	5'7"	67	1.70	130	126	58.82	57.30	20.31	19.79
BM03	31	37	N	Y	1	vaginal	caucasian	172 cm	68	1.72	154	134	70.00	61.00	23.66	20.62
BM04	35	42	N	Y	1	vaginal	caucasian	5'3"	63	1.60	188	192	85.28	87.09	33.30	34.01
BM05	31	39	Y	Y	1	vaginal	asian	5'9"	69	1.75	170	135	77.25	61.23	25.15	19.94
BM06	32	41	N	Y	1	vaginal	caucasian	5'3"	63	1.60	107	106	48.38	48.23	18.89	18.84
BM07	34	39	N	Y	3	C section	european	5'5"	65	1.65	170	155	77.26	70.31	28.34	25.79
BM08	31	38	N	N	1	C section	caucasian	5'7"	67	1.70	147	142	66.68	64.41	23.02	22.24
BM09	33	40	Y	Y	3	C section	caucasian	5'1"	61	1.55	195	187	88.45	84.82	36.84	35.33
BM10	31	41	Y	N	1	C section	caucasian	5'0"	60	1.52	152	125	69.10	56.70	29.75	24.41
BM11	33	39	N	N	1	vaginal	caucasian	5'8"	68	1.73	145	145	65.62	65.77	22.00	22.05
BM12	34	39	Y	N	1	vaginal	caucasian	5'7"	67	1.70	170	178	76.96	80.74	26.57	27.88
BM13	33	42	Y	Y	1	C section	hispanic/white	5'6"	66	1.68	154	152	69.85	68.95	24.86	24.53
BM15	34	37	Y	Y	1	vaginal	caucasian	5'3"	63	1.60	140	160	63.50	72.57	24.80	28.34
BM16	30	40	Y	Y	1	vaginal	caucasian	5'5"	65	1.65	135	135	61.23	61.23	22.46	22.46
BM17	25	42	Y	N	1	vaginal	caucasian	na	na	na	na	na	na	na	na	na
BM19	29	42	N	N	2	vaginal	caucasian	5'6"	66	1.68	185	155	83.91	70.31	29.86	25.02

**Table S2:** Generalized additive model (GAM) results for celltype proportion association with sample metadata

p-values of association of each celltype proportion with metadata using generalized additive model with formula proportion ~ donor + <metadata_column>							
	time post partum	any_formula_Yes	daycare_Years	hormonal_birthcontrol_Yes	hormonal_birthcontrol_no	vaccines_no	
B Cells	-	-	-	-	-	-	
T Cells	0.00234	-	-	-	-	-	
Dendritic Cells	-	-	-	-	-	0.01705	0.03
Eosinophils	-	-	-	-	-	-	
LC1	0.000189	-	0.03521	0.00532	-	-	
LC2	-	-	-	-	-	-	
GPNMB+ Macrophage	-	0.026	-	-	-	-	
CSN1S1+ Macrophage	-	-	-	-	-	-	
Neutrophils	-	-	-	-	-	-	
Number of samples included in comparison	56	55	55	53	53	53	47



p-values of association of each immune celltype proportion with metadata using generalized additive model with formula proportion ~ donor + <metadata column>

	time postpartum	any_formula_no	Infant_sick_yes	vaccines_no
B Cells	-	-	-	-
T Cells	0.0541	-	-	-
Dendritic Cells	-	-	0.0117	-
Eosinophils	-	0.0120	-	-
GPNMB+ Macrophage	-	-	-	0.0302
CSN1S1+ Macrophage	-	-	-	-
Neutrophils	-	-	-	-
<b>Number of samples included in comparison</b>	56	55	53	47

p-values of association of each celltype proportion with each donor using generalized additive model with formula proportion ~ donor+s(time post partum)

p-values of association of each epithelial celltype proportion with metadata using generalized additive model with formula proportion ~ donor + <metadata_column>			
	time post partum	daycare_Ye s	hormonal_birthcontrol_y es
Cycling lactocytes	-	-	-
KRT high 1	-	-	-
KRT high 2	0.049	-	-
LC1	0.00335	-	-
MT High LC2	-	0.0139	0.00835
Secretory lactocytes	-	-	-
Number of samples included in comparison	55	54	52

**Dataset S1** (separate file): Sample-level metadata

**Dataset S2** (separate file): All cells marker genes

**Dataset S3** (separate file): Macrophage sub-cluster marker genes

**Dataset S4** (separate file): Macrophage Reactome results

**Dataset S5** (separate file): Lactocyte and monocyte gene modules

**Dataset S6** (separate file): Epithelial cell marker genes

**Dataset S7** (separate file): Time-associated genes for each epithelial cell subcluster

**Dataset S8** (separate file): Universal epithelial varying genes and genes which change in opposite directions with time in LC1 epithelial cells and secretory epithelial cells

**Dataset S9** (separate file): Enrichr results for time-associated genes in epithelial cells

## SI References

1. Z. Fang, *GSEApY: Gene Set Enrichment Analysis in Python* (Zenodo, 2020) <https://doi.org/10.5281/ZENODO.3748085> (September 16, 2021).
2. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
3. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
4. A.-J. Twigger, *et al.*, Transcriptional changes in the mammary gland during lactation revealed by single cell sequencing of cells from human milk. *Nat. Commun.* **562** (2022).
5. J. F. Martin Carli, *et al.*, Single Cell RNA Sequencing of Human Milk-Derived Cells Reveals Sub-Populations of Mammary Epithelial Cells with Molecular Signatures of Progenitor and Mature States: a Novel, Non-invasive Framework for Investigating Human Lactation Physiology. *J. Mammary Gland Biol. Neoplasia* **25**, 367–387 (2020).
6. L. M. Murrow, *et al.*, “Changes in epithelial proportions and transcriptional state underlie major premenopausal breast cancer risks” (bioRxiv, 2020) <https://doi.org/10.1101/430611> (August 23, 2021).