Profiling of mature stage human breastmilk cells identifies host-defense lactocyte sub-populations

John P. Gleeson,¹ Namit Chaudhary,¹ Rose Doerfler,¹ Katherine C. Fein,¹ Trish Hredzak-Showalter,¹ Kathryn A. Whitehead^{1,2*}

¹ Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

² Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*Corresponding author:

Kathryn A. Whitehead, kawwhite@cmu.edu

Abstract

Breastmilk is chock-full of nutrients, immunological factors, and cells that aid infant development. Maternal cells are the least studied breastmilk component, and their unique properties are difficult to identify using traditional techniques. Here, we characterized the cells in mature stage breastmilk from healthy donors at the protein, gene, and transcriptome levels. Holistic analysis of flow cytometry, qPCR, and single cell RNA sequencing data identified the predominant cell population as epithelial with smaller populations of macrophages and T cells. Two percent of epithelial cells expressed four stem cell markers: SOX2, TRA-1-60, NANOG, and SSEA4. Furthermore, milk contained six distinct epithelial lactocyte sub-populations, including three previously unidentified sub-populations programmed towards host-defense and intestinal development. Pseudotime analysis delineated the differentiation pathways of epithelial progenitors. Together, these data define healthy human maternal breastmilk cells and provide a basis for their application in maternal and infant medicine.

37 Introduction

38 Although human infants are born immature (1, 2), their diet of breastmilk enables
 39 continued organ growth and development postpartum, particularly within the intestine (3).
 40 Human milk contains growth factors, immunomodulatory components, and bacterial and
 41 maternal cells. A substantial body of research has delineated the role of many nutritional
 42 components of milk, including proteins, fats, and sugars, with a primary goal of creating
 43 formula that recapitulates breastmilk (4). There is also considerable interest in bacterial
 44 cells present in the maternal milk microbiome, which help to establish the infant gut
 45 microflora (5). By contrast, despite their abundance ($\sim 10^6$ cells per ml), the biology of
 46 breastmilk cells and their potential role in infant development remains poorly understood.

47 The first step in determining the physiological relevance of breastmilk cells is to identify
 48 and characterize them. Various breastmilk cell types have been described to some degree
 49 in the literature, including lactocytes, immune cells, myoepithelial cells, and stem cells
 50 (6). Immune cells are the dominant population in early-stage milk (colostrum; <1 week
 51 after birth). Whereas, in mature stage milk from healthy mothers (>2 weeks postpartum),
 52 the majority of cells are lactocytes, which are terminally differentiated milk-producing
 53 epithelial cells (7). Multipotent stem cells have also been identified in breastmilk based on
 54 the expression of several stem cell gene markers, including p63 and KLF4 (8, 9). These
 55 cells were also differentiated into all three germ layers (10), a unique characteristic of
 56 pluripotent stem cells (11).

57 Although the roles of the maternal cells in breastmilk are not yet clear, there is early
 58 evidence that they may participate in several vital functions. For example, bulk RNA
 59 sequencing studies showed that the maternal cells in colostrum express a high level of
 60 non-nutritive genes (iron-binding, triglyceride catabolism, and protein digestion).
 61 Dominant gene expression then shifts in mature stage milk to the two genes responsible
 62 for producing milk proteins – α -lactalbumin and β -casein – in mature stage milk (12).
 63 Establishing baseline concentrations and gene expression profiles for distinct breastmilk
 64 cell types might also aid the development of diagnostics for the detection of cancer in the
 65 breastfeeding person (13).

66 Additionally, there is a report that maternal stem and immune cells exit the GI tract of the
 67 infant and functionally integrate into the infant's organs, similarly to immune cells (14).
 68 These data open the possibility that maternal cells play a role in organ development and
 69 passive immunity. For example, necrotizing enterocolitis, an intestinal disease associated
 70 with the use of formula, affects $\sim 7\%$ of premature infants and leads to 1,000-2,000 infant
 71 deaths a year in the US (15, 16).

72 Given this mounting evidence that maternal breastmilk cells are involved in infant
 73 development and may have properties amenable to cellular therapies and diagnostics (17),
 74 it is imperative that milk cells are clearly and comprehensively characterized. To that end,
 75 we were motivated to conduct a three-pronged analysis of milk cells with the broadest
 76 applicability to nursing people worldwide – fresh, mature-stage milk cells from healthy
 77 donors. The experiments reported here were designed to address the limitations of
 78 previous milk cell studies.

79 Specifically, numerous studies relied exclusively on flow cytometry to identify distinct
 80 milk cell populations (10). This is problematic because of variable fluorescent antibody
 81 quality and unavailability of antibodies for novel cell types in milk. Another limitation is

that maternal breastmilk cells are sometimes frozen after collection and before analysis (13). This can be problematic because some populations of breastmilk cells do not survive the freezing process; thus, analysis biases against cell populations that are not sufficiently robust.

Through this work, we have circumvented these challenges by analyzing fresh maternal breastmilk cells by flow cytometry, quantitative real time PCR (qRT-PCR), and single cell RNA sequencing (scRNA-seq) simultaneously. Although there is one other study that characterized breastmilk cells using scRNA-seq, it was performed on frozen breastmilk samples from women with gestational diabetes (13). To our knowledge, our study is the first of its kind, as it holistically describes the complete population of maternal breastmilk cells from healthy women. Because most lactating mothers are classified as healthy, these findings provide a broadly applicable reference dataset for maternal breastmilk cells and lay the foundation for future studies on the role of these cells in infant development and their potential as therapeutic interventions.

Results

Epithelial lactocytes are the most abundant cell population in breastmilk

For our initial analysis, we collected freshly expressed breastmilk from donors to characterize the maternal cell populations using flow cytometry and RT-qPCR (Fig. 1). Although recruitment was open to all donors regardless of postpartum age, we received donations only of mature stage milk (≥ 4 weeks, averaging 32 weeks post-partum; Table S1) (12). Following the cell isolation process as previously reported (10), we assessed viability by flow cytometry using a fluorescent live-dead stain and observed cell viability of $79.2\% \pm 6.9\%$ (Fig. S1; $N = 10$). The traditional Trypan Blue cell viability assay reported a high degree of false positives due to the presence of milk fat globules; this assay should be avoided with breastmilk cells.

Fig. 1 Maternal cells in breastmilk were identified by protein, gene, and transcriptome analyses. Mature-stage breastmilk from healthy donors was assessed for

111 expression of cell marker proteins (e.g. EpCAM, CD45) in freshly isolated samples by flow
 112 cytometry. Then, RNA was isolated and used to conduct complementary gene expression analysis
 113 by qRT-PCR. Single cell transcriptomics added depth to the analysis and defined the lactocyte
 114 population in breastmilk.

115
 116 To understand the cell populations present, we selected surface markers for flow
 117 cytometry analysis based on previous studies: epithelial cells (EpCAM+), immune cells
 118 (CD45+), and mesenchymal cells (VIM+) (6, 10, 12). We also assayed for lactocytes,
 119 which derive from epithelial cells, and thus are EpCAM+ in addition to being CK18+. The
 120 selected histograms indicate that the vast majority of breastmilk cells are of epithelial
 121 origin (93%), with more modest populations of CK18+, CD45+, and VIM+ cells from
 122 healthy donors ($N=13$) (Fig. 2A). Immune cell numbers increased from ~7% in milk from
 123 healthy donor-infant dyads to ~63% and ~36% when the infant ($N=1$) or the mother
 124 ($N=3$), respectively, had been sick (Fig. 2B). In the case of illness, the number of donors
 125 was too low to run statistical analysis. These flow cytometry data are shown for all donors
 126 in (Fig. 2C) and agree with previous reports, aside from the low percentage of cells
 127 expressing CK18 (10, 13). However, mRNA expression analysis shows reasonably high
 128 expression of *KRT18*, the gene associated with CK18 and lactocytes (Fig. 2D).
 129 Accordingly, we concluded that the lactocyte population observed by flow cytometry was
 130 artificially low, likely due to poor binding affinity of the antibody. Gene expression of an
 131 immune cell gene, *PTPRC*, and the mesenchymal gene marker, *VIM*, were significantly
 132 lower in expression, consistent with the flow cytometry data.

133

Fig. 2 Mature-stage breastmilk from donors contained epithelial lactocytes and a smaller population of immune cells. (A) Flow cytometry quantified expression of cell-specific markers: EpCAM+ (epithelial cells), CD45+ (immune cells), CK18+ (lactocytes), and VIM+ (mesenchymal cells). These data suggest that the CK18 antibody bound poorly to lactocytes, and likely underestimates true lactocyte percentages. (B) The relative cellular composition in breastmilk was affected by the health status of the mother and infant. Representative density plots are shown of EpCAM vs CD45 cell populations in milk from healthy donors ($N = 10$), donors with sick infants ($N = 1$), or donors with recent mastitis ($N = 3$). (C) Flow cytometry analysis of protein markers ($N = 13$) and (D) mRNA expression of gene markers corresponding to proteins in (C) indicate the presence of epithelial lactocytes ($N = 6-7$). Data are shown as mean $\pm$ SEM. One-way ANOVA with Tukey's multiple comparison test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Stem-like transcription factors are highly expressed in 2% of breastmilk cells

There has been significant interest in breastmilk as a novel source of stem cells. Stem cells can be identified by the transcription factors SOX2, NANOG, and OCT4, which are considered the master regulators in stem cells, and by REX1, SSEA4, TRA-1-60, and KLF4, which are co-regulators of pluripotency (18). As such, we asked whether the unknown fraction of epithelial cells in breastmilk were stem-like using flow cytometry and qRT-PCR.

For flow cytometry, we gated on EpCAM+ cells and then measured the percentage of cells expressing SOX2, TRA-1-60, NANOG, and SSEA4 (Fig. 3A). Although we wanted to assess OCT4 as well, we were limited by the constraints of our flow panel. Ultimately, we deemed it more informative to have two master regulator markers and two pluripotency markers in our analysis. These experiments indicated that most epithelial cells were SOX2+ and TRA-1-60+, but significantly fewer were NANOG+ and SSEA1+ (Fig. 3B).

We also conducted gene expression analysis by RT-qPCR for the genes encoding SOX2, TRA-1-60, NANOG, and SSEA4: *SOX2*, *PODXL*, *NANOG*, and *FUT4*, respectively. We quantified expression levels relative to the housekeeping gene, *EEF1A*, which was highly and uniformly expressed in all breastmilk cells. Epithelial cells expressed 1,000-fold greater *FUT4* mRNA than the other assessed genes, with decreasing expression of *PODXL* > *NANOG* > *SOX2* (Fig. 3C). The higher *FUT4* expression is likely due to its role in human milk oligosaccharide synthesis, so expression here may not be solely pluripotency related (19). When considered together, these data indicate that ~2% of the epithelial cells were positive for all four transcription factors assessed.

Fig. 3 Stem-like transcription factors were expressed in 2% of breastmilk cells. Four stem-like transcription factors were assessed in breastmilk epithelial cells. (A) From flow cytometry analysis, representative histograms of SOX2+, TRA-1-60+, NANOG+, and SSEA4+ stained and unstained samples are shown. (B) Flow cytometry analysis of EpCAM+ cells ($N = 15$)

indicated high expression of SOX2 and TRA-1-60 with moderate to low expression of NANOG and SSEA4, respectively. (C) mRNA expression relative to a housekeeping gene was quantified for the genes corresponding to the proteins in (B) ($N = 7$). Data are means $\pm$ SEM. One-way ANOVA with Tukey's multiple comparison test; * $P < 0.05$, **** $P < 0.0001$.

Single-cell transcriptomic analysis reveals unique lactocyte sub-types

Our flow cytometry and mRNA expression data suggest that breastmilk cells are primarily epithelial with a moderate number of lactocytes and a small number of stem-like cells. However, only limited conclusions can be drawn from these two techniques, as the quality of flow data may have been influenced by antibody quality, and bulk gene expression data cannot delineate gene expression patterns in varied cell populations.

To overcome the latter shortcoming, we performed single-cell RNA sequencing to robustly quantify and classify the maternal cell types in milk. For these experiments, we identified six donors with similar cellular profiles by flow cytometry as determined by a first donation. These donors provided samples a second time, and we selected three of those donors (mean 28 weeks postpartum) to advance to scRNA-seq experiments based on sample consistency and cell yield as determined by flow cytometry (Table S1). Upon their third donation, we submitted samples from these three donors within 1-2 hours of expression for scRNA-seq analysis. In total, the transcriptomes of 26,922 maternal breastmilk cells were sequenced, with a median of 1,057 genes per cell.

Gene expression across all three donor samples was dominated by two genes, *LALBA* and *CSN2*, that accounted for $> 40\%$ of gene counts (Fig. 4A). These two genes encode the proteins alpha-lactalbumin and beta-casein, respectively, which are two most abundant proteins in human milk (20). The genes identified by scRNA-seq were confirmed in a larger donor population by qRT-PCR ($N = 7$), high expression of *CSN2* and *LALBA* with decreasing expression of *CSN1S1* $>$ *FTH1* $>$ *LYZ*, genes encoding the proteins alpha-S1-casein, ferritin heavy chain 1, and lysozyme, respectively (Fig. 4B).

To identify the main cell types within the breastmilk cell population that are generalizable across individuals, we performed a combined analysis of the scRNA-seq data using the Seurat and Monocle pipeline (21). This analysis identified one large group of cells comprising the contiguous clusters 1 – 7 as well as two smaller independent clusters (8 and 9) based on Uniform Manifold Approximation and Projection (UMAP) clustering (22) (Fig. 4C). Comparison across donors showed that their cluster mappings were reasonably homogenous (Fig. 4D).

We then examined the differential expression of genes within each cluster (Fig. 4E), which identified two immune cell clusters, cluster 8 as macrophages (*CD68+*, *ITGAX+*) and cluster 9 as T-cells (*PTPRC+*, *CD3E+*). Although the major cluster (1 – 7) was epithelial (*LALBA+*, *EPCAM+*), there were unique sub-populations that we believe to be previously unidentified in human breastmilk: *FGFBP1+* lactocytes for intestinal development (cluster 2), fatty-acid synthesizing and insulin-resistant *PTPRF+* lactocytes (cluster 6), and secretory/chemotactic epithelial cells (cluster 7). Cluster 5 contained the cells that had undergone the least differentiation. These included a small population of progenitor-like cells, defined by high expression of *CD55* and *CLDN4* (23), and a larger

population of differentiating cells, which are the precursors of epithelial cells and
lactocytes. Finally, cluster 4 comprised apoptotic epithelial cells, and it is unclear what
cluster these cells belonged to prior to degradation.

Fig. 4 Single-cell transcriptome analysis identified breastmilk cells as primarily epithelial-derived lactocytes, including three novel sub-populations involved in host-defense. (A) Fifty percent of the transcripts produced by mature breastmilk cells from healthy donors corresponded to the milk-producing genes for lactalbumin (*LALBA*) and three types of caseins (*CSN2*, *CSN1S1*, and *CSN3*), N = 3. (B) RT-qPCR quantified mRNA expression relative to

229 *EEF1A* expression for several top genes ($N = 6$). (C) Single cell RNAseq data generated a UMAP
 230 plot that identified one major population of cells (Clusters 1 – 7) and two minor populations
 231 (Clusters 8 and 9). (D) The UMAP cluster map is consistent across all donors ($N = 3$). (E) Each
 232 cluster had unique expression patterns of the top ten marker genes. These patterns were used to
 233 identify specific cell types, including three novel types of epithelial lactocytes: *FGFBP1*+
 234 lactocytes, *PTPRF*+ lactocytes, and chemotactic epithelial cells.

235

236 To determine the pathway of differentiation of progenitor-like epithelial cells, we
 237 performed pseudotime trajectory analysis (Fig. 5). Fig. 5A shows the differentiation
 238 pathways of the breastmilk cells, originating from the most progenitor-like cells, which
 239 are part of cluster 5. The starting node is found in the top, deep purple portion of cluster 5.
 240 As pseudotime progresses, the cells differentiate along several distinct pathways, with
 241 partially and terminally differentiated cells appearing in reddish-orange and yellow,
 242 respectively. Fig. 5B shows the spatiotemporal gene expression profiles of nine selected
 243 cell markers. High expression of genes associated with progenitor cells or stemness (e.g.
 244 *CD44*, *SOD2*) occurred earliest in pseudotime and corresponded with cells in cluster 5
 245 (24-26), with *SOD2* peaking in cluster 1. These progenitor cells briefly expressed
 246 *CALML5*, which was exclusively expressed by differentiating epithelial cells, before they
 247 diverged into several cell types. Differentiated cells, which appeared in late pseudotime,
 248 included mature lactocytes in clusters 2 and 3, lipid and fatty-acid synthesizing cells in
 249 cluster 6, and chemotactic/secretory epithelial cells in clusters 1 and 7 (Fig. 5C).

250 Finally, we examined the expression homogeneity of selected markers across seven
 251 donors (Fig. S2), two of which were included in our scRNAseq (Table S1). *SOD2* and
 252 *CALML5* mRNA were moderately and homogenously expressed (Fig. S2A/B), suggesting
 253 progenitor and epithelial cell populations are consistent across donors. By contrast, the
 254 chemotactic epithelial marker *FDCSP* had 30-fold and 100-fold higher expression in two
 255 donors (Fig. S2D). These data suggest that lactocytes are homogenous across donors,
 256 while the specialized sub-types may be enriched for a specific need of the feeding infant,
 257 such as intestinal stem cell development (*FGFBP1*+ lactocytes; cluster 2), insulin
 258 resistance (*PTPRF*+ lactocytes; cluster 6), or intestinal immune defense (chemotactic
 259 epithelial cells; cluster 7).

Fig. 5 Pseudotime trajectory analysis identified the pathways by which breastmilk epithelial cells differentiate into lactocytes. (A) Pseudotime trajectory analysis determined the pathway of differentiation of progenitor-like epithelial cells (starting node is located in the dark purple region). (B) As pseudotime progressed, the expression of progenitor genes subsided as cells differentiated into mature phenotypes. (C) Based on the pseudotime trajectories and gene expression kinetics, milk cells differentiated along three pathways into chemotactic/secretory epithelial cells, fatty-acid synthesizing cells, and mature lactocytes. Each pathway was associated with overexpression of key genes compared to the other cell types, including *KRT15* (cluster 1), *FDCSP* (cluster 7), *FASN* (cluster 6), *LALBA* (cluster 2), and *FGFBP1* (cluster 3).

Discussion

Through this work, we have provided a detailed characterization of the cellular population in healthy mature-stage breastmilk. We confirmed that most cells were epithelial lactocytes and that ~2% of breastmilk cells are stem-like. Single-cell transcriptomic analysis identified novel sub-populations of lactocytes programmed towards host defense and intestinal development.

Most research on human breastmilk cells has focused on the characterization of stem-like cells present in the milk (14, 17, 27, 28). The most thorough of these studies isolated stem-like cells expressing multiple stem-transcription factors (OCT4+, SOX2+, and NANOG+) and differentiated them into all three germ layers (10). Unfortunately, this study was unable to determine the percentage of cells positive for more than one marker because the antibodies used in flow cytometry experiments were all FITC-labelled. To our knowledge, we are the first to report that the percentage population of stem-like cells in breastmilk is 2% (SOX2+/TRA-1-60+/NANOG+/SSEA4+) (Fig. 3). Interestingly, single-cell transcriptomic analysis did not identify a cluster of stem-like cells within our breastmilk cell population. This is most likely due to the difference between gene expression and protein translation. However, *SOD2* was expressed well in our samples (Fig. 5B), and this expression has been linked to pluripotency transcription in stem cells (25). Further studies are needed to understand the transcriptomic and proteomic profiles of the stem-like cells in breastmilk to determine if they are as useful for regenerative medicine applications, as many authors suggest. Recently, Habowski *et al.* used combined transcriptomics and proteomics to delineate the rapid differentiation of colonic stem cells (29). This same type of analysis could potentially determine the stemness and pluripotency of breastmilk-derived cells at a gene and protein level.

Our findings are generally aligned with those reported from the scRNAseq analysis of two donors with gestational diabetes (13). Differences observed might be due to the greater number of cells (6x) analyzed in this study, our use of an additional donor, and/or our use of fresh, rather than frozen samples. The use of fresh cells assures that results are not biased towards more robust milk cells that survive freeze/thaw cycles. Further, freeze/thaw processes have been linked to altered gene expression and translation in human primary cells (30, 31).

These methodological differences enabled our identification of novel sub-populations: *FGFBP1*+ lactocytes (cluster 2), *PTPRF*+ lactocytes (cluster 6), and chemotactic epithelial cells (cluster 7; Fig. 4E). Regarding cluster 2, *FGFBP1* is a soluble carrier protein that binds to fibroblast growth factor (FGF), which plays a key role in intestinal development and disease regulation (32, 33). Cluster 2 also expressed *TNFRSF11B*, which activates the Wnt/ β -catenin pathway (essential for intestinal stem cell proliferation) (34) and regulates the differentiation of immune sampling M-cells (35).

The *PTPRF*+ lactocytes highly expressed *MUC1*, *TGFBR3*, and *PTPRF*. *PTPRF* overexpression in breastmilk cells might be linked to insulin resistant mothers (12). *MUC1* secreted in breastmilk binds to the lectin domain of dendritic cells found in infant intestines, preventing pathogenic interactions (36). *TGFBR3* plays a key role in activation and inhibition of T and B cells respectively (37). In summary, this population may contribute to maternal insulin-resistance signaling and host-defense in the infant.

Finally, chemotactic epithelial cells (cluster 7) had high expression of *FDCSP*, *IL19*, *SAA1*, *WFDC2*, *TCN1*, and *LCN2*. *FDCSP* is considered an ancestral precursor gene to *CSN3*, which produces the milk protein κ -casein (38, 39). The other genes with high expression in this cluster appear to confer host-defense, chemokine signaling, and intestinal immune maturation. For example, IL-19 expression was increased in the milk cells of lactating women with mastitis that had been treated with oral probiotics (40). The bioactive protein product of *SAA1* in human milk confers protection to the neonate in part by intestinal host-defense and Th17 activation (41). Additionally, *WFDC2* contributes to

innate immune defenses of the lung, nasal, and oral cavities (42, 43), and *LCN2* expression occurs in response to small bowel villous disruption (44). Bach and colleagues suggested that mammary epithelial cells should be conceptualized as being part of a continuous spectrum of differentiation, and our results support this (45). Overall, our data suggest that the unique lactocyte sub-types identified here by scRNAseq are temporally and uniquely suited to the health and development of mother-infant dyads and strengthen the infants' immune system during early development.

These findings have provided a genetic fingerprint for the cells in healthy, mature-stage breastmilk and a framework for examining the functionality of these cell types in biologic, prophylactic, and therapeutic contexts. Future work will determine the role of the sub-populations identified here in healthy and sick infants and apply those learnings to the development of novel, oral therapeutic interventions for gastrointestinal and other infantile diseases.

Materials and Methods

Breastmilk sample collection and cell isolation

This study was approved by Carnegie Mellon University Institutional Review Board (IRB) under the protocol number – STUDY2019_00000084. Participants were recruited via flyers and social media advertisements in mother and infant groups in the greater Pittsburgh area. All participants provided informed written consent. Pump-expressed mature breastmilk was obtained from each participant at lactation suites at Carnegie Mellon University, placed on ice, and transported to the laboratory immediately after expression for analysis. Breastmilk samples were processed as previously reported (10). Briefly, breastmilk samples were diluted 1:1 with sterile phosphate buffered saline (PBS; Life Technologies), centrifuged, washed, and resuspended in PBS prior to centrifugation three times. The resultant breastmilk cell pellet was then processed for single-cell RNA sequencing (scRNAseq), quantitative real-time PCR (RT-qPCR), or flow cytometry.

Cell viability and flow cytometry analysis

Cell pellets were resuspended in 100 µl of 10% fetal bovine serum (FBS; VWR) and incubated with either Trypan Blue (Invitrogen) or LIVE/DEAD™ Fixable Yellow stain (Invitrogen) for 5 minutes at room temperature and 30 minutes at 37°C, respectively. Fixable Yellow samples were subsequently fixed using Flow Cytometry Fixation Buffer (R&D Systems), centrifuged, and resuspended in 10% FBS in PBS. Trypan Blue cell viability was measured using Countess™ II Automated Cell Counter (ThermoFisher). For flow cytometry, samples were fixed in Fixation Buffer for 30 minutes at 4°C, washed with Permeabilization Buffer (R&D Systems), centrifuged, and incubated with fluorescent antibodies (Table S2) in Permeabilization Buffer at 4°C for 1 hour. Samples were then washed with 10% FBS in PBS, centrifuged, and resuspended in 10% FBS in PBS. Appropriate negative internal controls were used. Data acquisition was done with a NovoCyte® 3000 (ACEA Biosciences) and data analysis with NovoExpress.

Gene expression analysis by RT-qPCR

Cell pellets were lysed in Buffer RLT, mixed with an equal volume of 100% ethanol, and the RNA was extracted with RNeasy mini kit (Qiagen). cDNA was generated from 1500 ng of RNA using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). RT-qPCR was performed using SYBRTM Select Master Mix (Applied Biosystems) on a ViiA 7 Real-Time PCR System (Applied Biosystems). Primer sequences are presented in Table S3.

Single-cell RNA sequencing

Cells were processed with Chromium Next GEM Single Cell 3' GEM Library and Gel Bead Kit v 3.1 (10X Genomics) according to manufacturer instructions. Briefly, cells were diluted in master mix containing reverse transcription reagents and primer and were transferred to the ChromiumChip with gel beads and partitioning oil for preparation of nanoliter-scale Gel Bead-In-Emulsions (GEMs). Reverse transcription produced cDNA with a cellular 10X barcode and unique molecular identifier (UMI), and the cDNA was recovered with Dynabeads MyOne SILANE magnetic beads. The cDNA was then subjected to 8-14 cycles of amplification before clean-up using SPRIselect Reagent (Beckman). Total cDNA yield was calculated following quality assessment (High Sensitivity D5000 ScreenTape, Agilent) and fluorometric quantitation (Qubit, ThermoFisher). Libraries were constructed by enzymatic fragmentation, end repair and A-tailing according to the manufacturer's instructions. Two rounds of clean-up were performed: First, a double sided SPRIselect Reagent (Beckman) prepared samples for adapter ligation. Then, another round of magnetic bead clean-up was performed before sample indexing PCR for 5-16 cycles depending on the cDNA input for library construction. A further double sided size selection produced libraries of 400-600 bp, which were quantified for paired end sequencing (26 bp x 98 bp). The 10X single cell RNA libraries quality was checked using Fragment Analyzer (NGS Fragment Kit, Agilent) and they were quantified by qPCR (Kapa qPCR quantification kit, Kapa biosystem). The libraries were then normalized, pooled, and sequenced on NovaSeq6000 platform (Illumina) with S1 100 cycles kit (read1: 28 bp; read2: 91 bp).

scRNAseq analysis

All reads were processed using the 10X cellranger pipeline. Briefly, reads were demultiplexed, using the cellranger mkfastq function. Demultiplexed reads from all lanes were aligned to the GRCh38 genome, followed by filtering, barcode counting, and UMI counting using the cellranger count function. After constructing the gene expression matrix, cells were processed using Seurat and Monocle packages. Cells were filtered to remove low quality cells (< 400 genes), doublets (> 5,000 genes), and cells with high mitochondrial fraction (> 25%). Cells were assigned cell cycle scores based on the presence of S, G1, and G2M genes. The filtered subset was normalized and scaled using Seurat's SCT pipeline. Mitochondrial percentage, cell cycle scores, gene counts, and UMI counts were regressed out. Samples from all three donors were integrated using Seurat's SCTransform integration pipeline. Highly variable genes from all 26,922 cells were used as features for Principal Component Analysis (PCA) using the RunPCA function. The first 50 principal components were used for UMAP dimensionality reduction using the RunUMAP function. The integrated RNA counts from the Seurat object along with the first 50 principal components and the UMAP coordinates were transferred to a cell_data_set for clustering and trajectory inference using monocle3. Cells were clustered using the Leiden algorithm in monocle using the cluster_cells function. For trajectory

inference, the principal graph was learned using the learn_graph function. Cells were ordered in pseudotime by choosing the root node using the order_cells function. A node in the epithelial cell cluster with high expression of KRT7 and KRT18 was chosen as the root node. Gene dynamics as a function of pseudotime were analyzed using the plot_genes_in_pseudotime function. Finally, the clusters from monocle3 were added to the Seurat object's metadata. Differentially expressed genes were identified using the MAST algorithm with the FindAllMarkers function.

Statistical analysis

All data are represented as mean $\pm$ standard error of the mean (SEM). Statistical analysis was carried out using Prism 8 software (GraphPad) using Student's t-test and one-way ANOVA with Bonferroni's post hoc tests. A significant difference was defined as $P < 0.05$.

H2: Supplementary Materials

Fig. S1 – Cells isolated from the breastmilk of 12 unique donors were 40 to 95% viable with an average viability of 70%. Viability was measured using flow cytometry LIVE/DEAD™ Fixable Yellow stain ($79.2\% \pm 6.9\%$; $N = 10$).

Fig. S2 –Cluster marker mRNA expression in a larger donor pool is relatively consistent (A) UMAP localization and corresponding mRNA expression of epithelial cell marker, progenitor cells (*SOD2*; cluster 1/5) and differentiating cell marker (*CALML5*; cluster 5) are consistently expressed across multiple donor samples. Markers for specialized subtypes are enriched in specific donors (C; *PTPRF*+ lactocytes) and (D; chemotactic epithelial cells). UMAP analysis ($N = 3$) and RT-qPCR ($N = 7$).

Table S1. Donor characteristics and experimental groups

Donor ID	Weeks Postpartum	Health status	Viability assay	Cell type flow cytometry	Cell type qPCR	Stem cell flow cytometry	Stem cell qPCR	scRNAseq	RNAseq qPCR
2.1	18	H		Y	Y	Y	Y		Y
2.2	28	H		Y		Y			
2.3	31	H	Y					Y	
4	38	H		Y		Y			
5	83	H	Y	Y		Y			
6.1	12	H		Y	Y	Y	Y		Y
6.2	19	H		Y		Y			
6.3	20	H	Y					Y	
8.1	39	H	Y	Y		Y			
8.2	44	H		Y		Y			Y
9	21	M	Y	Y		Y			

11	11	M	Y	Y	Y	Y	Y		Y
12.1	29	H		Y		Y			
12.2	32	H		Y		Y			
12.3	33	H	Y					Y	
13	92	H	Y	Y	Y	Y	Y		Y
17	6	S	Y						
18	5	H		Y	Y	Y	Y		Y
19	36	M							
21	40	H	Y	Y	Y	Y	Y		Y

Health status: H = healthy mother and infant, M = recent mastitis infection, and S = infant currently sick.

Table S2. Flow cytometry antibodies used in analysis

Marker	Fluorophore	Company	Cat. No
EpCAM	PE-Cy7	eBioscience	25-9326-42
CD45	Alexa Fluor® 488	R&D Systems	FAB1430G
Vimentin	APC	Invitrogen	MA528601
CK18	FITC	Invitrogen	MA110327
aSMA	Alexa Fluor® 594	R&D Systems	IC1420T-100
SOX2	eFluor® 660	eBioscience	50-9811-82
TRA-1-60	DyLight® 488	Invitrogen	MA1023D488X
Nanog	PE	Invitrogen	PA546891
SSEA4	PerCP-eFluor® 710	eBioscience	46-8843-42
CD49f	eFluor® 450	eBioscience	48-0495-82

Table S3. Primer sequences for RT-qPCR

Gene	Forward	Reverse
<i>EEF1A</i>	ATCTCAGGCTGACTGTGCTG	AGTGTGTAAGCCAGAAGGGC
<i>EPCAM</i>	GCTGGCCGTAAACTGCTTTG	ACATTTGGCAGCCAGCTTTG
<i>PTPRC</i>	TGAAATGTGTATGCACCTATTG AAA	TGGGGCCTGTAAAAGTGTCC
<i>VIM</i>	CGCACATTCGAGCAAAGACA	ACAAGAGCGCCCCTAAGTTT
<i>KRT18</i>	AAAGCCTGAGTCCTGTCCTTTC	GGTGAAGCTCATGCTGTCCG
<i>FUT4</i>	GGGGGTTCTTCCTCACCTTG	ATATGGCCTGTGGCAGATGG
<i>PODXL</i>	TCAGACCGTGGTCGTCAAAG	TTCATGTCACTGACCCCTGC
<i>SOX2</i>	TACAGCATGATGCAGGACCA	CCGTTTCATGTAGGTCTGCGA
<i>NANO G</i>	ACATGAGTACTGCTTTAGTTGG T	TCCACCCCAACCAAAAATTAA CA
<i>LALBA</i>	AGCCAGGTCCCTCAGTCAA	ATGGGCCAACCAGTAGTCAA
<i>CSN2</i>	GCTCTTGCAAGGGAGACCAT	TCTGTAATAGATTCTCTACTGC TTG
<i>FTH1</i>	AGAACTACCACCAGGACTCAG A	GTCAAAGTAGTAAGACATGGA CAG
<i>SPP1</i>	AACGCCGACCAAGGAAAAC	TGTTGTGGAGGGGTAGGTACA
<i>LYZ</i>	AGGTGTGAGTTGGCCAGAAC	ACACATCCAGTTTGCTAGGC
<i>TMSB1 0</i>	ACGAGACTGCACGGATTGTTT	TCCGCTTCTCCTGCTCAATG

<i>FABP3</i>	ACTACATGAAGTCACTCGCTCA TA	TGAGGGTAGGGGGAAGGTTA
<i>FGFBP 1</i>	CAGCCTGGCTCCTGTTGAAT	GGCGTTCACCTTGTCTGAG
<i>CALML 5</i>	TACGAGGAGTTCGCGAGGAT	AGAGTCCCAGCACAAAAGCA
<i>FDCSP</i>	TTTCGCCCCACTTCCACCAAT	AGGTGACCAGGTTTATCGTGAC
<i>SOD2</i>	AAACCTCAGCCCTAACGGTG	CACGTTTGATGGCTTCCAGC
<i>TGFBR 3</i>	CACTAAGCCCCTTGCTGTGA	CCAGACCATGGAAAATTGGTG G
<i>CSN1S1</i>	CTCACCTGTCTTGTGGCTGT	TGTTTCTCTGCCTGTTTCATACC

Acknowledgments

Acknowledgments: This project used the University of Pittsburgh HSCRF Genomics Research Core for RNAseq prep and library prep, and we thank H. Monroe and D. Hollingshead for their assistance. This work was supported by the UPMC Genome Center with funding from UPMC's Immunotherapy and Transplant Center for single cell sequencing, and we thank Y. Pan for their assistance. **General:** We thank our donors for providing breastmilk samples and making this work possible. **Funding:** NIH New Directors Fund NICHD DP2-HD098860. **Author contributions:** J.P.G and K.A.W conceived and designed the experiments. R.D. and T.H.S. recruited donors. J.P.G., N.C. and K.C.F. performed the experiments. J.P.G., N.C. and K.A.W. analyzed the data. J.P.G., N.C., R.D., K.C.F., T.H.S., and K.A.W. contributed to writing and editing the manuscript. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper may be requested from the authors.

1. A. G. Cummins, F. M. Thompson, Postnatal Changes in mucosal immune response: A physiological perspective of breast feeding and weaning. *Immunology & Cell Biology* **75**, 419-429 (1997).
2. J. P. Gleeson *et al.*, The enhanced intestinal permeability of infant mice enables oral protein and macromolecular absorption without delivery technology. *Int. J. Pharm.* **593**, 120120 (2021).
3. A. S. Goldman, Modulation of the Gastrointestinal Tract of Infants by Human Milk. Interfaces and Interactions. An Evolutionary Perspective. *The Journal of nutrition* **130**, 426S-431S (2000).
4. D. K. Dror, L. H. Allen, Overview of Nutrients in Human Milk. *Advances in Nutrition* **9**, 278S-294S (2018).
5. K. Kordy *et al.*, Contributions to human breast milk microbiome and enteromammary transfer of *Bifidobacterium breve*. *PloS one* **15**, e0219633-e0219633 (2020).
6. M. Witkowska-Zimny, E. Kaminska-El-Hassan, Cells of human breast milk. *Cell. Mol. Biol. Lett.* **22**, 11-11 (2017).
7. F. Hassiotou *et al.*, Maternal and infant infections stimulate a rapid leukocyte response in breastmilk. *Clin Transl Immunology* **2**, e3-e3 (2013).
8. M. D. Cregan *et al.*, Identification of nestin-positive putative mammary stem cells in human breastmilk. *Cell Tissue Res.* **329**, 129-136 (2007).

9. E. Thomas, N. Zeps, M. Cregan, P. Hartmann, T. Martin, 14-3-3 σ (sigma) regulates proliferation and differentiation of multipotent p63-positive cells isolated from human breastmilk. *Cell Cycle* **10**, 278-284 (2011).
10. F. Hassiotou *et al.*, Breastmilk is a novel source of stem cells with multilineage differentiation potential. *Stem Cells* **30**, 2164-2174 (2012).
11. I. International Stem Cell, Assessment of established techniques to determine developmental and malignant potential of human pluripotent stem cells. *Nature communications* **9**, 1925-1925 (2018).
12. D. G. Lemay *et al.*, RNA sequencing of the human milk fat layer transcriptome reveals distinct gene expression profiles at three stages of lactation. *PloS one* **8**, e67531-e67531 (2013).
13. 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. *Journal of Mammary Gland Biology and Neoplasia* **25**, 367-387 (2020).
14. J.-P. Molès *et al.*, Breastmilk cell trafficking induces microchimerism-mediated immune system maturation in the infant. *Pediatric Allergy and Immunology* **29**, 133-143 (2018).
15. J. R. Robinson *et al.*, Surgical necrotizing enterocolitis. *Semin Perinatol* **41**, 70-79 (2017).
16. S. C. Fitzgibbons *et al.*, Mortality of necrotizing enterocolitis expressed by birth weight categories. *J. Pediatr. Surg.* **44**, 1072-1076 (2009).
17. N. Ninkina *et al.*, Stem cells in human breast milk. *Human Cell* **32**, 223-230 (2019).
18. T. Seymour, A.-J. Twigger, F. Kakulas, Pluripotency Genes and Their Functions in the Normal and Aberrant Breast and Brain. *International journal of molecular sciences* **16**, 27288-27301 (2015).
19. D. S. Newburg *et al.*, Innate protection conferred by fucosylated oligosaccharides of human milk against diarrhea in breastfed infants. *Glycobiology* **14**, 253-263 (2003).
20. M. J. C. van Herwijnen *et al.*, Comprehensive Proteomic Analysis of Human Milk-derived Extracellular Vesicles Unveils a Novel Functional Proteome Distinct from Other Milk Components. *Molecular & cellular proteomics : MCP* **15**, 3412-3423 (2016).
21. C. Sonesson, M. D. Robinson, Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255-261 (2018).
22. M. D. Luecken, F. J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
23. P. Wang *et al.*, A molecular signature for purified definitive endoderm guides differentiation and isolation of endoderm from mouse and human embryonic stem cells. *Stem cells and development* **21**, 2273-2287 (2012).
24. C. He *et al.*, SOD2 acetylation on lysine 68 promotes stem cell reprogramming in breast cancer. *Proceedings of the National Academy of Sciences* **116**, 23534-23541 (2019).
25. C. Solari *et al.*, Manganese Superoxide Dismutase Gene Expression Is Induced by Nanog and Oct4, Essential Pluripotent Stem Cells' Transcription Factors. *PloS one* **10**, e0144336-e0144336 (2015).
26. H. Ghebeh *et al.*, Profiling of normal and malignant breast tissue show CD44^{high}/CD24^{low} phenotype as a predominant stem/progenitor marker when used in combination with Ep-CAM/CD49f markers. *BMC Cancer* **13**, 289 (2013).
27. M. Ş. Aydın, E. N. Yiğit, E. Vatandaşlar, E. Erdoğan, G. Öztürk, Transfer and Integration of Breast Milk Stem Cells to the Brain of Suckling Pups. *Scientific Reports* **8**, 14289 (2018).
28. S. Li *et al.*, Characterization of stem cells and immune cells in preterm and term mother's milk. *Journal of Human Lactation* **35**, 528-534 (2019).

29. A. N. Habowski *et al.*, Transcriptomic and proteomic signatures of stemness and differentiation in the colon crypt. *Communications Biology* **3**, 453 (2020).
30. X. Fu *et al.*, Effects of cryopreservation and long-term culture on biological characteristics and proteomic profiles of human umbilical cord-derived mesenchymal stem cells. *Clin Proteomics* **17**, 15-15 (2020).
31. E. Kadić, R. J. Moniz, Y. Huo, A. Chi, I. Kariv, Effect of cryopreservation on delineation of immune cell subpopulations in tumor specimens as determined by multiparametric single cell mass cytometry analysis. *BMC Immunol.* **18**, 6 (2017).
32. C. W. Houchen, R. J. George, M. A. Sturmoski, S. M. Cohn, FGF-2 enhances intestinal stem cell survival and its expression is induced after radiation injury. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **276**, G249-G258 (1999).
33. S. Kanazawa *et al.*, VEGF, basic-FGF, and TGF-beta in Crohn's disease and ulcerative colitis: a novel mechanism of chronic intestinal inflammation. *Am J Gastroenterol* **96**, 822-828 (2001).
34. F. Luan *et al.*, TNFRSF11B activates Wnt/ β -catenin signaling and promotes gastric cancer progression. *Int J Biol Sci* **16**, 1956-1971 (2020).
35. S. Kimura *et al.*, Osteoprotegerin-dependent M cell self-regulation balances gut infection and immunity. *Nature Communications* **11**, 234 (2020).
36. N. Koning *et al.*, Human Milk Blocks DC-SIGN-Pathogen Interaction via MUC1. *Frontiers in immunology* **6**, 112-112 (2015).
37. Y. S. Jang *et al.*, Lactoferrin causes IgA and IgG2b isotype switching through betaglycan binding and activation of canonical TGF- β signaling. *Mucosal Immunology* **8**, 906-917 (2015).
38. C. HOLT, J. A. CARVER, Darwinian transformation of a 'scarcely nutritious fluid' into milk. *J. Evol. Biol.* **25**, 1253-1263 (2012).
39. K. Kawasaki, A.-G. Lafont, J.-Y. Sire, The Evolution of Milk Casein Genes from Tooth Genes before the Origin of Mammals. *Mol. Biol. Evol.* **28**, 2053-2061 (2011).
40. J. de Andrés, E. Jiménez, I. Espinosa-Martos, J. M. Rodríguez, M.-T. García-Conesa, An Exploratory Search for Potential Molecular Targets Responsive to the Probiotic *Lactobacillus salivarius* PS2 in Women With Mastitis: Gene Expression Profiling vs. Interindividual Variability. *Front Microbiol* **9**, 2166-2166 (2018).
41. G. H. Sack, Jr. *et al.*, Serum amyloid A1 (SAA1) protein in human colostrum. *FEBS Open Bio* **8**, 435-441 (2018).
42. L. Bingle *et al.*, WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. *Respir Res* **7**, 61-61 (2006).
43. K. M. Morris *et al.*, Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. *Scientific Reports* **6**, 35011 (2016).
44. M. d. M. G. Prata *et al.*, Comparisons between myeloperoxidase, lactoferrin, calprotectin and lipocalin-2, as fecal biomarkers of intestinal inflammation in malnourished children. *J Transl Sci* **2**, 134-139 (2016).
45. K. Bach *et al.*, Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications* **8**, 2128 (2017).