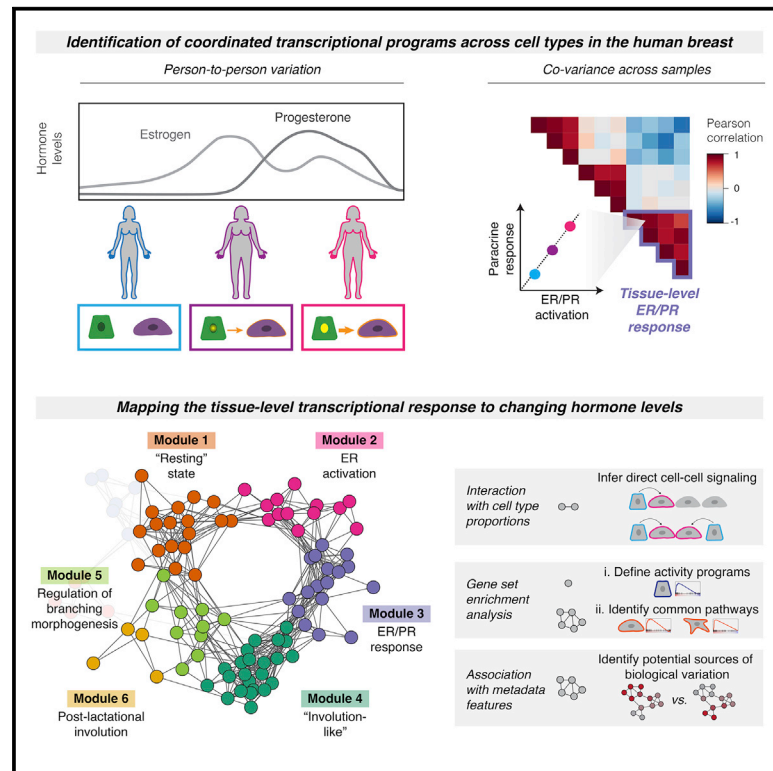


Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution

Graphical abstract



Authors

Lyndsay M. Murrow, Robert J. Weber, Joseph A. Caruso, ..., Matthew Thomson, Thea Tlsty, Zev J. Gartner

Correspondence

lmurrow@gmail.com (L.M.M.), zev.gartner@ucsf.edu (Z.J.G.)

In brief

Estrogen and progesterone regulate breast development and modify cancer risk. Using single-cell analysis and leveraging person-to-person variability to identify gene programs that co-vary across individuals, Murrow et al. map the tissue-level response to ovarian hormones. Prior pregnancy and obesity modify hormone responsiveness in the breast through distinct mechanisms.

Highlights

- Single-cell analysis of the human breast maps the tissue-level response to hormones
- DECIPHER-seq identifies gene programs that co-vary across individuals
- Dependency on cell-type proportions predicts direct cell-cell interactions
- Prior pregnancy and obesity modify hormone responsiveness through distinct mechanisms

Article

Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution

Lyndsay M. Murrow,^{1,8,*} Robert J. Weber,^{1,2,3} Joseph A. Caruso,⁴ Christopher S. McGinnis,^{1,9} Kiet Phong,^{1,3} Philippe Gascard,⁴ Gabrielle Rabadam,^{1,3} Alexander D. Borowsky,⁵ Tejal A. Desai,^{3,10} Matthew Thomson,⁶ Thea Tlsty,⁴ and Zev J. Gartner^{1,7,11,*}

¹Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA

²Medical Scientist Training Program (MSTP), University of California, San Francisco, San Francisco, CA 94158, USA

³Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

⁴Department of Pathology and Helen Diller Cancer Center, University of California, San Francisco, San Francisco, CA 94143, USA

⁵Center for Immunology and Infectious Diseases, Department of Pathology and Laboratory Medicine, University of California, Davis, Davis, CA 95696, USA

⁶Computational Biology, Caltech, Pasadena, CA 91125, USA

⁷Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

⁸Present address: Department of Cellular and Translational Genomics, Genentech, South San Francisco, CA 94080, USA

⁹Present address: Department of Pathology, Stanford University, Stanford, CA 94305, USA

¹⁰Present address: School of Engineering, Brown University, Providence, RI 02912, USA

¹¹Lead contact

*Correspondence: lmurrow@gmail.com (L.M.M.), zev.gartner@ucsf.edu (Z.J.G.)

<https://doi.org/10.1016/j.cels.2022.06.005>

SUMMARY

The rise and fall of estrogen and progesterone across menstrual cycles and during pregnancy regulates breast development and modifies cancer risk. How these hormones impact each cell type in the breast remains poorly understood because they act indirectly through paracrine networks. Using single-cell analysis of premenopausal breast tissue, we reveal a network of coordinated transcriptional programs representing the tissue-level response to changing hormone levels. Our computational approach, DECIPHER-seq, leverages person-to-person variability in breast composition and cell state to uncover programs that co-vary across individuals. We use differences in cell-type proportions to infer a subset of programs that arise from direct cell-cell interactions regulated by hormones. Further, we demonstrate that prior pregnancy and obesity modify hormone responsiveness through distinct mechanisms: obesity reduces the proportion of hormone-responsive cells, whereas pregnancy dampens the direct response of these cells to hormones. Together, these results provide a comprehensive map of the cycling human breast.

INTRODUCTION

Coordinated interactions between cells are essential for the development and maintenance of normal tissue function, and dysregulation of cell-cell interactions is a key driver of disease. In the human breast, fluctuations in the levels of estrogen and progesterone with each menstrual cycle and during pregnancy control cell growth, survival, differentiation, and tissue morphology. The impact of these changes is profound: cumulative lifetime exposure to cycling hormones is a major modifier of breast cancer risk (Collaborative Group on Hormonal Factors in Breast Cancer, 2012), and the majority of breast tumors are estrogen dependent. However, many of the effects of ovarian hormones within the breast are indirect. The estrogen and progesterone receptors (ER/PR) are expressed in only 10%–15% of cells within the epithelium (Clarke et al., 1997). Thus, most of

the changes that occur in response to hormone receptor activation are mediated by a complex cascade of paracrine signaling from hormone-responsive (HR+) cells to other cell types in the breast. Accordingly, cell-cell interactions between HR+ cells and other cell types are key to normal breast morphogenesis. However, due to a number of challenges inherent to hormone signaling and human breast biology, we lack a systems-level understanding of how different cell populations respond to cycling hormone levels.

A first challenge for understanding the tissue-level response to estrogen and progesterone is that there are major differences in glandular architecture and stromal composition and complexity between humans and model organisms like the mouse (Dontu and Ince, 2015; Parmar and Cunha, 2004). For example, while ER expression is restricted to the epithelium in humans, it is also expressed in the stroma in rodents (Mueller et al., 2002;

Palmieri et al., 2004). Therefore, understanding the consequences of cellular crosstalk downstream of estrogen and progesterone requires studying these processes in humans or human models.

A second challenge is that the human breast is both heterogeneous across individuals and characterized by a highly dynamic microenvironment. There is a high degree of variability between individuals in terms of epithelial architecture (Russo et al., 1992), cell composition (Nakshatri et al., 2015; Rosenbluth et al., 2020), and hormone responsiveness (Dunphy et al., 2020; Muenst et al., 2017; Tanos et al., 2013), and these differences likely impact both normal breast function and breast cancer susceptibility. Within individuals, the menstrual cycle and pregnancy/lactation/involution cycle are major drivers of epithelial remodeling, characterized by alternating periods of epithelial expansion and regression in response to changing hormone levels (Anderson et al., 1982; Jindal et al., 2014; Söderqvist et al., 1997; Russo et al., 1992). Histological analyses of paraffin-embedded human tissue sections have also identified cyclical alterations in epithelial architecture and stromal organization across the menstrual cycle (Ramakrishnan et al., 2002; Vogel et al., 1981) and broad remodeling following weaning (Lyons et al., 2011; O'Brien et al., 2010). However, little is known about how this underlying heterogeneity impacts cell state and the intercellular signaling networks that control tissue morphogenesis. As it enables unbiased analysis of cell types within the human mammary gland at single-cell resolution, single-cell RNA sequencing (scRNA-seq) is particularly well suited to investigate this problem.

Here, we use scRNA-seq in a cohort of twenty-eight premenopausal reduction mammoplasty tissue specimens (RM cohort) to trace the transcriptional changes that occur in the human breast downstream of hormone signaling. To provide insight into the cellular interactions that regulate breast tissue homeostasis, we develop DECIPHER-seq: Deconstructing Cell-cell Interactions using Phenotypic Heterogeneity in scRNA-seq data, a systematic computational approach that leverages the high degree of inter-sample transcriptional heterogeneity in the breast to identify coordinated interaction networks across cell types. Our approach was guided by two hypotheses. First, we predicted that if two cell types are acting together—via either direct cell-to-cell signaling or a response to shared microenvironmental/paracrine cues—the transcriptional signatures representing those cell-cell interactions should be correlated across samples. More specifically, since the effects of estrogen and progesterone on other cell types in the breast are controlled by paracrine signaling from HR+ luminal cells, we reasoned that hormone receptor activation in HR+ luminal cells would be correlated with transcriptional changes in other cell types, representing the downstream paracrine response. Second, we predicted that we could infer the types of cell-cell interactions that make up each pairwise correlation by incorporating information about (1) the proportion of each cell type across samples, (2) the similarity of each transcriptional signature, and (3) enrichment of common biological pathways across signatures. Thus, we sought to use the inter-sample transcriptional variability and differences in cell-type proportions present in the dataset as a type of “natural experiment” to understand how the behaviors of different cell types in the breast are coordinated at the tissue level.

Based on this approach, we identify a network of coordinated activity programs in HR+ cells and other cell types that represent the dynamic tissue-level response of the human breast to changing hormone levels. Using differences in cell-type proportions across samples, we infer a subset of activity programs that depend on direct cell-to-cell signaling and find that these direct interactions primarily comprise signaling from HR+ cells to other cell types. Using these data, we generate hypotheses about how person-to-person variation at the tissue level is linked to specific biological mechanisms at the cellular level, and directly test these hypotheses using flow cytometry and immunostaining in an expanded cohort of samples. We find that paracrine signaling from HR+ cells to neighboring cell types depends on both the magnitude of the ER/PR transcriptional response and the overall abundance of HR+ cells in the tissue. Accordingly, we demonstrate that prior pregnancy and obesity both lead to decreased hormone responsiveness in the breast, but act through distinct mechanisms: pregnancy influences the magnitude of the ER/PR signaling response in HR+ luminal cells, whereas obesity reduces the proportion of HR+ cells and therefore downstream paracrine signaling. These changes are consistent with the protective effect of prior pregnancy and high body mass index (BMI) against premenopausal breast cancer. Overall, these results provide a comprehensive map of the cycling human breast and the dynamic cell-cell interactions that underlie normal breast function and breast cancer risk. More broadly, we describe a systematic approach to unravel the functional significance of person-to-person variability in the human breast at the tissue level, by linking individual cell types' transcriptional signatures to higher-order modules of cell-cell interactions.

RESULTS

Person-to-person variability in transcriptional cell state in the premenopausal human breast

To identify inter-individual differences in transcriptional cell state in the human breast, we performed scRNA-seq on 86,136 cells collected from 28 healthy premenopausal donors who underwent reduction mammoplasty (Figures 1A and S1A; Table S1). To obtain an unbiased snapshot of the epithelium and stroma, we collected live (DAPI-negative) singlet cells from all samples by fluorescence-activated cell sorting (FACS) (Figures S1A and S1B; Table S2). For a subset of samples, we also collected purified epithelial cells or purified luminal and basal/myoepithelial cells (Figures S1A and S1B; Table S2). We used MULTI-seq lipid-based barcoding and *in silico* genotyping for sample multiplexing to minimize technical variability between samples (Figure S1C; Table S2; STAR Methods; Heaton et al., 2020; McGinnis et al., 2019b).

Sorted basal and luminal cell populations were well resolved by Uniform Manifold Approximation and Projection (UMAP) (Figure S1D). Unsupervised clustering identified one basal/myoepithelial cluster, two luminal clusters, and six stromal clusters (Figure 1B). Based on the expression of known markers, the two luminal clusters were annotated as HR+ and secretory luminal cells, and the six stromal clusters were annotated as fibroblasts, vascular endothelial cells, lymphatic endothelial cells (“lymphatic”), smooth muscle cells/pericytes (“vascular accessory”), lymphocytes, and macrophages (Figures 1B, S1E, and



(legend on next page)

S1F). The luminal populations described here closely match those identified as “HR+/mature luminal” and “secretory/luminal progenitor” in previous flow cytometry and scRNA-seq analyses of the human breast (Lim et al., 2010; Bhat-Nakshatri et al., 2021; Nguyen et al., 2018). Here, we use the nomenclature “hormone-responsive/HR+” and “secretory” to refer to these two luminal cell types. The HR+ cluster was enriched for the hormone receptors ESR1 and PGR (Figure S1G), and other known markers such as ANKRD30A (Figures S1E and S1F; Nguyen et al., 2018). Consistent with previous studies demonstrating variable hormone receptor expression across the menstrual cycle (Battersby et al., 1992), expression of ESR1 and PGR transcripts were sporadic and often non-overlapping. Within the HR+ luminal cluster, 22% of the cells had detectable levels of ESR1 or PGR, with only 2% of HR+ cells expressing both transcripts (Figure S1H).

Beyond identifying the major cell types, single-cell analysis resolved a high degree of person-to-person transcriptional variability in the human breast. Following batch-correction (Figure S2A) (Butler et al., 2018), cells from different individuals were represented across all cell-type clusters (cluster entropy = 0.93, STAR Methods) (Figure S1B). However, despite this mixing across cell types, individuals displayed distinct transcriptional signatures within individual cell-type clusters (Figures 1C and S1C). Because we used MULTI-seq to multiplex samples, we were able to confirm that this variation in cell state was not due to technical variation, as we directly compared cells from different samples that were run in the same batch to cells from matched samples that were run across multiple batches. Cells from the same sample were more similar to each other than cells from different samples, regardless of the batch/day of processing (Figures S2D and S2E; Table S2; STAR Methods).

Inferring shared transcriptional responses and direct cell-to-cell signaling interactions in the human breast

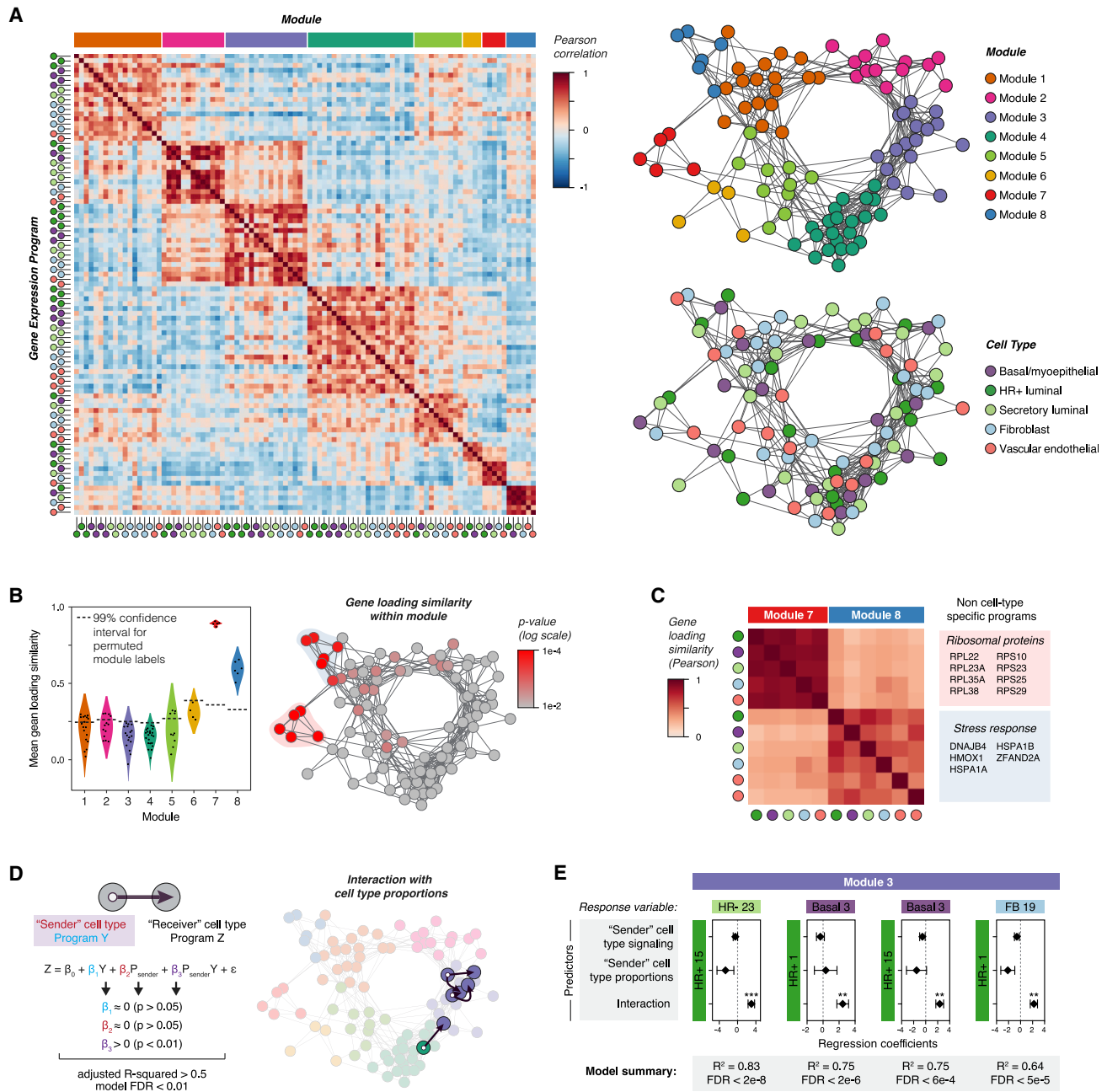
Since estrogen and progesterone are master regulators of breast development, and the levels of these hormones fluctuate across the menstrual cycle, we predicted that ER/PR signaling and the downstream paracrine response would be a major source of transcriptional heterogeneity across samples in our dataset. Based on random sampling across the menstrual cycle and differences in hormonal contraceptive use, we would expect to

identify samples with varying levels of ER/PR activation in HR+ luminal cells (Figure 1D). If these HR+ cells are signaling to other cell types, such as basal cells, we would further expect to see a second activity program in those cells representing the downstream paracrine response. Finally, this “paracrine response” activity program should co-vary with the level of ER/PR activation across different samples (Figure 1D). Thus, we developed a computational pipeline, DECIPHER-seq, based on the hypothesis that inter-sample transcriptional variation contains meaningful information about how the behaviors of different cell types in the breast are coordinated at the tissue level, and that transcriptional signatures (“activity programs”) representing interactions between two cell types should correlate across samples. DECIPHER-seq uses individual pairwise correlations between activity programs to build a higher-order network map of coordinated cell state changes in the human breast (Figure 1E).

The activities of two cell types can be coordinated in multiple ways. In the premenopausal breast, we expect the tissue-level response to hormones to lead to at least two types of coordinated interactions: direct cell-to-cell signaling interactions between HR+ cells and other cell types, and more complex downstream interactions involving cell-type-specific responses to a shared microenvironment. We predict that the first type of interaction would depend on the proportion of HR+ cells in the breast, whereas the second type of interaction would involve cell-type-specific (e.g., transcriptionally distinct) activity programs that may be enriched for similar biological processes. Therefore, in downstream analyses, we infer modules that are enriched for direct cell-cell signaling interactions (i.e., modules containing links that depend on the proportion of one cell type across samples), and exclude modules driven by non-cell-type-specific responses (i.e., modules containing transcriptionally similar activity programs) (Figure 1E). We also define individual activity programs and modules by performing gene set enrichment analysis, which allows us to infer higher-order functional interactions between multiple cell types. Finally, we uncover associations between annotated metadata features and sets of activity programs to infer potential sources of biological variation (Figure 1E). Known biology associated with paracrine signaling downstream of ER/PR activation provides a powerful “proof of concept” to establish that correlated changes in cell state can be used to identify biologically relevant cell-cell interactions.

Figure 1. Sample-to-sample variability in transcriptional cell state in the premenopausal human breast

- (A) Single-cell transcriptional analysis links biological variables with person-to-person heterogeneity in transcriptional cell state. scRNA-seq workflow: reduction mammoplasty samples were processed to epithelial-enriched tissue fragments, then to single cells, followed by MULTI-seq sample barcoding, library preparation using the 10X Chromium system, and sequencing.
- (B) The major epithelial and stromal cell types in the breast were identified and visualized by UMAP dimensionality reduction and unsupervised clustering of twenty-eight samples reduction mammoplasty samples (GEO: GSE198732, Table S1).
- (C) Density plots (arbitrary units, linear scale) highlighting the transcriptional cell state of hormone-responsive (HR+) luminal cells and basal/myoepithelial cells from each sample.
- (D) Overview of conceptual approach: we hypothesized that hormone receptor activation would represent a major source of transcriptional variability in our dataset and that hormone receptor activation in hormone-responsive (HR+) luminal cells would correlate with transcriptional changes in other cell types—representing the downstream paracrine response. Based on the differences in hormone levels due to menstrual cycling (depicted, left) or hormonal contraceptive use, we predicted that gene expression programs representing ER/PR signaling in HR+ luminal cells and the downstream signaling response in other cell types would co-vary across samples (right).
- (E) Using individual pairwise correlations between cell activities, DECIPHER-seq builds a tissue-level map of the cell-cell interactions present in the healthy human breast and identifies modules of transcriptional states that co-occur across the same sets of samples. In downstream analyses, we exclude modules driven by non-cell-type-specific responses to shared signals and uncover modules enriched for putative direct cell-cell signaling interactions. We define activity programs using gene set enrichment analysis, identify common pathways enriched across activity programs in a module, and uncover potential sources of biological variation by testing association with annotated metadata features.



(legend continued on next page)

To identify activity programs within cell types in the premenopausal breast, we performed non-negative matrix factorization (NMF) on each of the major cell-type clusters in our dataset (Figure S3A). A similar approach was recently used by Pelka et al. to identify multicellular immune “hubs” in colorectal cancer (Pelka et al., 2021). We used integrative NMF (iNMF) (Gao et al., 2021; Welch et al., 2019), which successfully corrected for batch differences while retaining sample-to-sample transcriptional variability (Figures S4A and S4B) and adapted a consensus approach (Kotliar et al., 2019) to identify activity programs that were consistent across replicates (Figure S4C; STAR Methods). The main user-supplied parameter in NMF is the number of programs identified (rank, K). None of the three commonly used heuristics for guiding the choice of K identified an obvious “elbow” in our dataset (Figures S4D and S4E). We therefore developed a metric based on the goal of identifying the greatest number of robust (i.e., consistent across values of K) and unique (i.e., distinct from other programs at the same K) activity programs (Figure S5; STAR Methods). This approach identified distinct “blocks” of activity programs in multiple cell types that co-varied across samples (Figure 2A). To build a tissue-level map of these cell-cell interactions, we constructed a weighted network of coordinated activity programs based on the pairwise Pearson correlations r (Figures S3B and S6). Based on this analysis, we identified eight major modules comprising highly correlated transcriptional states across cell types in the breast (Figures 2A and S6D).

To exclude non-cell-type-specific transcriptional responses—that are unlikely to be directly related to hormone signaling in the breast—we identified modules made up of activity programs with similar gene loadings. We found that modules 7 and 8 were highly enriched for activity programs with correlated gene loadings (Figures 2B and S7A). Programs in module 7 primarily consisted of ribosomal transcripts and genes involved in cellular respiration, whereas programs in module 8 consisted of stress response genes such as heat shock and chaperone proteins (Figures 2C and S7C). We speculate that module 8 represents an artifact of tissue processing rather than biologically meaningful transcriptional variation, since prior studies have identified a similar signature in dissociated solid tissues (O’Flanagan et al., 2019). However, one advantage of DECIPHER-seq is that it describes cells as a combination of activity programs rather than forcing cells into distinct clusters. Thus, samples with high expression of “dissociation-related” activity programs still contain biologically meaningful signals from other programs and can be retained in the analysis.

Next, we inferred modules enriched for putative direct cell-cell signaling interactions by identifying interactions between two nodes that depended on both the magnitude of activity program expression in a “sender” cell type and the proportion of that sender cell type in the tissue (Figures 1E and 2D). We reasoned that if one cell type was signaling to another, the activity program representing the transcriptional response in the “receiver” cell type should be sensitive to the proportion of sender cells in the tissue, particularly for direct interactions involving short-range

signaling molecules. While this simplified model does not consider the effects of signal amplification, cooperation between signaling pathways, or higher-order interactions between more than two cell types, it identifies a subset of “high-confidence” direct cell-cell interactions that meet a set of simple criteria. We annotated putative direct cell-cell signaling interactions as those where the combined effects of signaling from a sender cell type and its proportion in a tissue described over 50% of the variation in activity program expression across samples in a second “receiver” cell type, and the individual effects of signaling and cell proportions were not significant (Figure 2D; STAR Methods). As the proportion of epithelial versus stromal cells in our samples may be influenced by tissue dissociation, we restricted our analysis to links between epithelial cell types as “sender” cells (HR+ luminal, secretory luminal, or basal cells) and all cell types as “receivers.” We modeled each pairwise interaction as a linear response to three variables: signaling from a sender cell type (i.e., the mean expression score of an activity program in that cell type), the proportion of the sender cell type in the epithelium, and an interaction term representing the combined effects of signaling and cell proportions (Figure 2D). Consistent with our prediction about the nature of hormone signaling in the breast, four out of the five high-confidence direct cell-cell interactions (false discovery rate [FDR] < 0.01) were part of the same module (module 3), and consisted of a link between HR+ luminal cells as the “sender” cell type and a second “receiver” cell type (Figures 2D and 2E).

ER/PR signaling and the downstream transcriptional response

We next performed marker and gene set enrichment analysis to define potential functions for activity programs within each module and identify common pathways upregulated across multiple activity programs in a module (STAR Methods; Tables S3 and S4). We first focused on module 3 (Figures 3A and S8A), as our previous analysis demonstrated that this module was highly enriched for putative direct cell-cell signaling interactions. Since estrogen and progesterone are master regulators of breast development that act via paracrine signaling from HR+ luminal cells to other cell types, we predicted that ER/PR signaling and the downstream paracrine response would represent a major source of direct cell-cell signaling signatures present in our dataset.

Consistent with this hypothesis, activity programs in module 3—here annotated as the “ER/PR response” module—were highly enriched for genes previously found to be upregulated during the luteal phase of the menstrual cycle in a bulk RNA-seq analysis (module enrichment $p < 0.01$; Figure 3B; Table S5; Pardo et al., 2014). Activity program 1 in HR+ luminal cells (“ER/PR signaling”) was associated with high expression of the essential PR target genes WNT4 and TNFSF11 (RANKL) (Rajaram et al., 2015; Tanos et al., 2013), and enriched for transcripts in the Molecular Signatures Database Hallmark “early estrogen response” ($p < 0.001$) and “late estrogen response” ($p < 0.01$) gene sets (Figures 3C, S8B, and S8C; Liberzon et al., 2015). Additional canonical HR+

(E) Results from multiple linear regression analysis, depicting the four most significant (FDR < 0.01) inferred direct cell-cell interactions. For each pairwise combination, the response variable was modeled in response to three predictors: the expression score in a “sender” cell type (signaling), the proportion of the “sender” cell type, and an interaction term between both predictors. Points represent the regression coefficient for each predictor, and error bars depict the standard error.

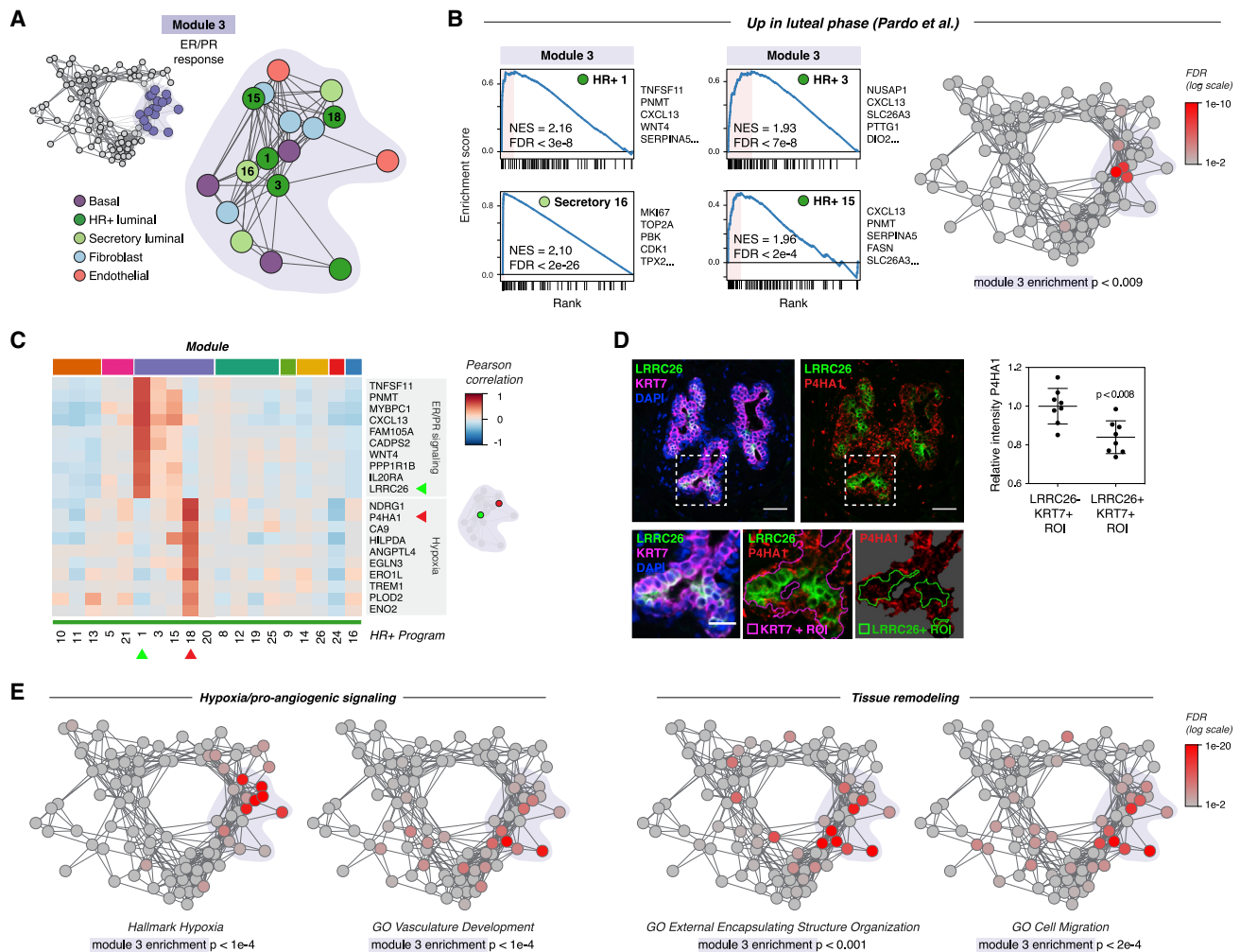


Figure 3. ER/PR signaling and the downstream response

(A) Diagram highlighting activity programs in the “ER/PR response” module.

(B) Left: gene set enrichment analysis of the indicated activity programs in the “ER/PR response” module, showing the enrichment of genes upregulated during the luteal phase of the menstrual cycle (Pardo et al., 2014). The top five leading edge genes for each activity program are listed. Right: network graph of activity programs, colored by the FDR for gene set enrichment of genes upregulated during the luteal phase of the menstrual cycle (log scale; Pardo et al., 2014). The FDR-corrected p value for enrichment of this gene set in the “ER/PR response” module was determined by permutation analysis.

(C) Heatmap of the top 10 marker genes for HR+ 1 and HR+ 8. Results depict the Pearson correlation between the expression score of the indicated activity programs and the normalized expression of the indicated genes across cells.

(D) Representative immunostaining for LRRC26, P4HA1, and KRT7 and quantification of the relative mean intensity of P4HA1 signal in LRRC26⁺/KRT7⁺ and LRRC26⁻/KRT7⁺ regions of interest ($p < 0.008$, unpaired t test). Data are represented as individual points, and error bars indicate mean \pm SEM of 8 regions from 3 samples with high ER/PR signaling. Scale bars, 20 μ m. Inset scale bars, 10 μ m.

(E) Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). FDR-corrected p values for overall enrichment of gene sets within module 1 were determined by permutation analysis.

genes including TFF1, AREG, PGR, and VEGFA were highly expressed across multiple activity programs in this module (Figure S8D; Aupperlee et al., 2013; Hyder et al., 2000; LaMarca and Rosen, 2007; Ribieras et al., 1998). Consistent with previous work demonstrating that STAT5 acts as a cofactor to mediate signaling downstream of PR activation in the breast, the ER/PR response module was also enriched for genes involved in IL-2/STAT5 signaling (module enrichment $p < 1e-4$; Figure S8E). Finally, gene set enrichment analysis identified a rare subpopulation of proliferative secretory luminal cells within the ER/PR response module (Figure 3B). This “proliferation” activity program

(Secretory program 16) was highly enriched for cell-cycle-related genes previously found to be upregulated during the luteal phase of the menstrual cycle (Figure 3B; Table S5; Pardo et al., 2014).

Our analysis also revealed that high levels of ER/PR signaling in HR+ cells (HR+ 1) coincided with the emergence of a second transcriptional state in a distinct subpopulation of HR+ luminal cells (HR+ 18) (Figures 3C and S8F). Marker and gene set enrichment analysis demonstrated that HR+ program 18 was characterized by upregulation of a hypoxia gene signature and pro-angiogenic factors such as VEGFA and ANGPTL4 (Figures S8D and S8G). The identification of this “hypoxia” gene signature is consistent

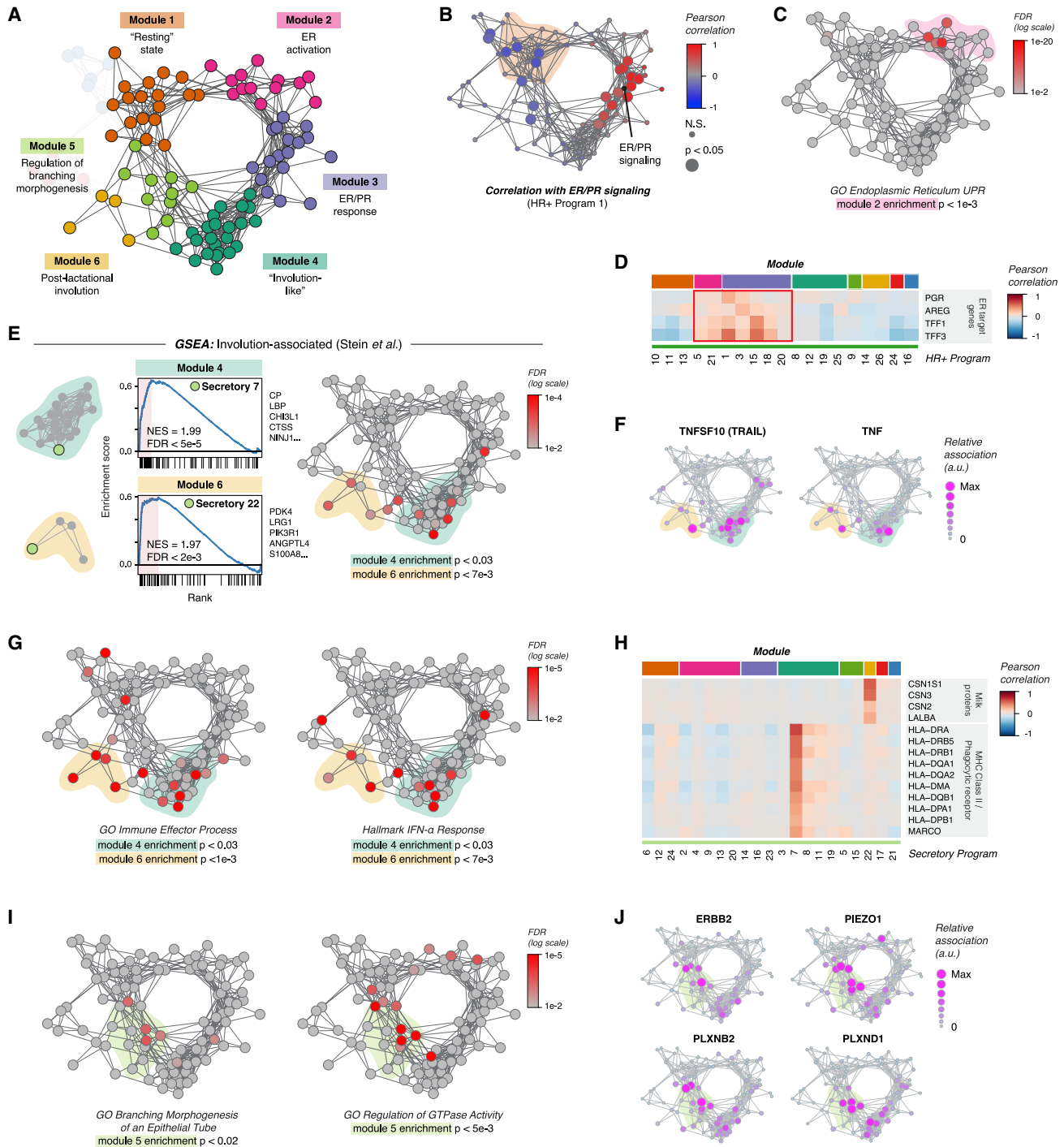


Figure 4. Coordinated changes in signaling states across cell types in the breast

(A) Network diagram highlighting modules 1–6.

(B) Network graph of activity programs in the human breast, colored by the Pearson correlation of each program's mean expression score across samples with ER/PR signaling (HR+ activity program 1). Significant positive and negative correlations ($p < 0.05$) as identified by bootstrap resampling are represented by larger nodes.

(C) Network graph of activity programs, colored by the FDR for enrichment of genes in the GO Biological Process set "endoplasmic reticulum unfolded protein response" (log scale). Overall enrichment of this gene set within module 2 was determined by permutation analysis.

(D) Heatmap of selected estrogen receptor (ER) target genes. Results depict the Pearson correlation between the expression score of the indicated activity programs and the normalized expression of ER target genes across cells.

with a previous study using microdialysis of healthy human breast tissue, which found that VEGF levels increased in the luteal phase of the menstrual cycle (Dabrosin, 2003). As estrogen response elements have been identified in the untranslated regions of VEGFA (Hyder et al., 2000), our results suggest that this increased expression may be, in part, a direct effect of hormone signaling to a subpopulation of HR+ cells.

To confirm these results *in vivo*, we performed marker analysis to identify genes specific to each cluster that could be used for immunostaining. We identified LRRC26 as a marker of the ER/PR signaling activity program HR+ 1 and P4HA1 as a marker of the hypoxia/pro-angiogenic activity program HR+ 18 (Figure 3C). In intact human tissue sections, we found that LRRC26 staining marked a distinct set of luminal cells from P4HA1 (Figure 3D). Moreover, these two subpopulations co-occurred within the same regions of the breast, demonstrating that they are unlikely to be an artifact of sample processing. Together, these results identify at least two diverging transcriptional states in HR+ cells in samples with high ER/PR signaling, one associated with signaling via RANK ligand and WNT4 to the surrounding epithelium and a second associated with a hypoxia-related/pro-angiogenic transcriptional signature.

We next expanded our analysis of gene activity programs to other epithelial lineages and stromal cell types in the “ER/PR response” module. Similar to program 18 in HR+ cells, multiple activity programs across other cell types in this module were enriched for transcripts involved in hypoxia and blood vessel remodeling including VEGFA and ANGPTL4 (Figures 3E, S8D, and S8G). The ER/PR response module was also enriched for genes involved in tissue remodeling, cell migration, and ECM organization (Figures 3E and S8H), consistent with previously reported morphological changes in the breast epithelium (Ramakrishnan et al., 2002) and alterations in stromal organization and ECM composition (Ferguson et al., 1992; Hallberg et al., 2010) across the menstrual cycle. Stromal cell types in this module were characterized by upregulation of ECM and matrix remodeling proteins including collagens (COL3A1, COL1A2), the cross-linking enzyme LOXL2, and the cytokine TGFB3 (Figure S8I). Together, these results identify distinct transcriptional signatures for ER/PR activation in HR+ luminal cells and the downstream paracrine response in other cell types.

Coordinated changes in signaling states across cell types in the breast

Next, we used a similar approach to analyze the remaining five major modules—annotated here as “resting state,” “ER activation,” “involution-like,” “post-lactational involution,” and

“regulation of branching morphogenesis”—each made up of highly interconnected transcriptional states across cell types in the breast (Figure 4A). The “resting state” module (module 1, Figure S9A) consisted of gene expression programs that were negatively correlated with ER/PR signaling (HR+ program 1) in HR+ luminal cells (Figure 4B). Activity programs in this module were enriched for pathways involved in RNA processing and transport (Figure S9B). The “ER activation” module (module 2, Figure S9C), consisted of activity programs linked to both the “resting state” and “ER/PR response” modules (Figure 4A). This module was enriched for genes involved in the unfolded protein response (UPR) and endoplasmic reticulum stress (Figures 4C, S9D, and S9E), as well as the response to estrogen (Figure S9D). Prior work has shown that ER activation leads to a rapid “anticipatory” activation of the UPR in the absence of accumulation of unfolded proteins (Andruska et al., 2015). In keeping with this, expression of canonical ER target genes such as PGR, AREG, TFF1, and TFF3 was most closely associated with HR+ cell activity programs in this “ER activation” module as well as the “ER/PR response” module (Figure 4D).

Gene set enrichment analysis of the “post-lactational involution” module (module 6, Figure S10A) and “involution-like” module (module 4, Figure S10D) uncovered transcriptional signatures in secretory luminal cells that were similar to those that have been described during post-lactational involution in the mouse (Figures 4E and S10E; Table S6; Stein et al., 2004). Activity programs in both modules were characterized by high expression of death receptor ligands such as TNFSF10 (TNF-related apoptosis-inducing ligand, TRAIL) and TNF (Figure 4F) and of genes involved in the immune response, including interferon-response genes (Figure 4G). We annotated module 6 as related to post-lactational involution, since activity program expression in secretory luminal cells within this module (secretory program 22) was highly associated with expression of milk proteins (Figure 4H) and genes involved in lactation (Figure S10B). Moreover, activity programs across all cell types in this module were more highly expressed in parous (P) versus nulliparous (NP) samples (Figure S10C). This “post-lactational involution” module was also enriched for genes involved in the acute phase response, complement proteins, and defense response, consistent with pathways that have been previously described as upregulated during post-lactational involution in the mouse (Figure S10F; Stein et al., 2004). Since prior studies in human tissue samples have shown that differences in lobular area and epithelial architecture between P and NP women persist for up to 18 months following weaning (Jindal et al., 2014), we speculate that activity programs in this module may be associated with the time since

- (E) Left: gene set enrichment analysis of the indicated activity programs in the “involution-associated” and “post-lactational involution” modules, showing the enrichment of genes previously shown to be upregulated during the post-lactational involution in the mouse (Stein et al., 2004). The top five leading edge genes for each activity program are listed. Right: network graph of activity programs, colored by the FDR for enrichment of genes upregulated in the Stein gene set (log scale). FDR-corrected p values for overall enrichment of this gene set in each of the indicated modules was determined by permutation analysis.
- (F) Network graph of activity programs, depicting the relative association of the indicated marker genes with each activity program (arbitrary units, linear scale).
- (G) Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). FDR-corrected p values for overall enrichment of gene sets within the indicated modules were determined by permutation analysis.
- (H) Heatmap of selected genes including milk proteins, MHC Class II molecules, and the phagocytic receptor MARCO. Results depict the Pearson correlation between the expression score of the indicated activity programs and the normalized expression of the indicated genes across cells.
- (I) Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). FDR-corrected p values for overall enrichment of gene sets within module 5 were determined by permutation analysis.
- (J) Network graph of activity programs, depicting the relative association of the indicated marker genes with each activity program (arbitrary units, linear scale).

weaning, although more complete patient data would be required to formally test this hypothesis.

The “involution-like” signature (module 4) in secretory luminal cells was characterized by expression of major histocompatibility complex class II (MHC-II) molecules and the phagocytic receptor MARCO (Figure 4H), suggesting that these cells play a role as non-professional phagocytes in the clearance of apoptotic cells, similar to what has been described during post-lactational involution (Monks et al., 2008). As previous data have demonstrated that the fraction of apoptotic cells in the mammary epithelium peaks between the late luteal and early follicular phases of the menstrual cycle, this module may represent the response to falling hormone levels at the end of the menstrual cycle (Anderson et al., 1982). TGFB3 is a major signaling molecule involved in post-lactational involution that enhances phagocytosis by mammary epithelial cells (Fornetti et al., 2016), suggesting that TGFB3 secreted by cells in response to ER/PR signaling (Figure S8I) activates a subset of secretory luminal cells that go on to express “involution-like” markers.

Finally, we annotated module 5 (Figure S10G) as associated with “regulation of branching morphogenesis” based on enrichment for the gene ontology (GO) term “branching morphogenesis of an epithelial tube” (Figures 4I and S10H). Consistent with the critical role of Rac and Rho GTPases in mammary branching (Ewald, 2008), the GO term “regulation of GTPase activity” was also highly enriched across this module (Figure 4I). Activity programs in this module were also associated with genes involved in cell motility, mechanotransduction, and invasion—including ERBB2, PIEZO1, PLXNB2, and PLXND1 (Figure 4J)—that have been previously described as important for epithelial remodeling (Gay et al., 2011; Stewart et al., 2021; Worzfeld et al., 2012).

Together, these results demonstrate how the underlying sample-to-sample variability in the breast can be used to infer functional connections between cell types in cell-cell interaction networks. Using DECIPHER-seq, we provide a comprehensive, systems-level view of the transcriptional changes that underlie normal breast morphogenesis.

The ER/PR signaling response of HR+ luminal cells is reduced in parous women

Previous epidemiologic analyses have demonstrated that prior pregnancy is highly protective against ER+/PR+ breast cancer (Fortner et al., 2019), and decreased hormone responsiveness following pregnancy is one proposed mechanism for this effect (Britt et al., 2007). Supporting this, previous studies demonstrated decreased expression of the PR effector WNT4 following pregnancy (Meier-Abt et al., 2014; Muenst et al., 2017). Moreover, in an explant culture model, estrogen consistently induced expression of the ER target gene AREG only in NP women (Dunphy et al., 2020). As our network analysis suggested that activity programs in the “ER/PR response” module were dependent on both the magnitude of signaling from HR+ luminal cells and their proportion in the tissue (Figures 2D and 2E), we hypothesized that decreased hormone responsiveness could be caused by either (1) a change in the magnitude of paracrine signals produced by each HR+ luminal cell and/or (2) a reduction in the overall proportion of HR+ luminal cells leading to a “dilution” of paracrine signals following ER/PR activation. It has been diffi-

cult to distinguish between these mechanisms using bulk tissue-level analyses. By individually probing the single-cell transcriptional landscape of the HR+ luminal cell population and downstream cell types, scRNA-seq provided a means to directly interrogate whether parity influences the per-cell hormone signaling response of HR+ luminal cells.

To quantify variation in ER/PR signaling in HR+ luminal cells, we first measured the similarity between each sample’s single-cell distribution across HR+ activity program 1 (ER/PR signaling). Hierarchical clustering identified two sets of samples, representing those with high or low ER/PR signaling (Figure 5A). Based on this, we found that while the levels of hormone signaling in HR+ luminal cells varied between NP women—likely reflecting differences in hormone levels across the menstrual cycle or due to hormonal contraceptive use—per-cell ER/PR signaling in HR+ luminal cells was significantly reduced in P women ($p < 0.02$, Mann-Whitney test; Figure 5B) and did not depend on other biological variables such as age and BMI (Figure S11A). Equal numbers of individuals from each cohort were using hormonal contraceptives ($n = 4$ out of 11 NP or P individuals, Table S1). For women not using hormonal contraceptives ($n = 7$ out of 11 NP or P individuals), we modeled the expected number of samples with high ER/PR signaling based on a binomial distribution using average menstrual cycle phase lengths (Bull et al., 2019). The number of NP samples with high ER/PR signaling was consistent with the expected number of samples in the luteal phase (2 of 7 samples, $p = 0.24$), whereas the number of P samples with high hormone signaling was significantly lower than expected based on the average length of the follicular and luteal phases of the menstrual cycle (0 of 7 samples, $p = 0.02$) (Figure 5C). These results remained consistent when we used a model accounting for previously reported differences in the relative lengths of the follicular versus luteal phases in P women (Figure S11B; Barrett et al., 2014). Thus, the decreased per-cell ER/PR signaling seen in HR+ luminal cells from P women cannot be explained by differences in hormonal contraceptive use or random sampling across the menstrual cycle.

To identify differentially expressed genes between NP and P women with high sensitivity, we generated a “pseudo-bulk” dataset of aggregated HR+ luminal cells from each sample (STAR Methods) and confirmed that P women had decreased expression of the canonical HR+ genes AREG, WNT4, PGR, TNFSF11 (RANKL), and TFF1 (Figure 5D; Table S7). The PR itself is an ER target gene (Kastner et al., 1990). Staining for the PR confirmed that PR expression was reduced in luminal cells in P samples in both our original sequenced cohort of samples (“discovery” set, $p < 0.005$) and a second independent cohort of samples (“validation” set, $p < 0.05$) (combined $p < 0.002$, Mann-Whitney test; Figure 5E). This reduction in PR expression was not due to broad changes in the lobular architecture of P women, as our results were consistent when we restricted our analysis to either lobular (TDLUs, terminal ductal lobular units) or ductal regions of the epithelium (Figure S11C).

Finally, we confirmed that paracrine signaling downstream of PR activation was reduced in P samples by assessing the effects of one of these genes, WNT4. As WNT4 from HR+ luminal cells has been shown to signal to basal cells (Rajaram et al., 2015), we performed co-immunostaining for the WNT effector TCF7 and basal/myoepithelial cell marker p63 and found that

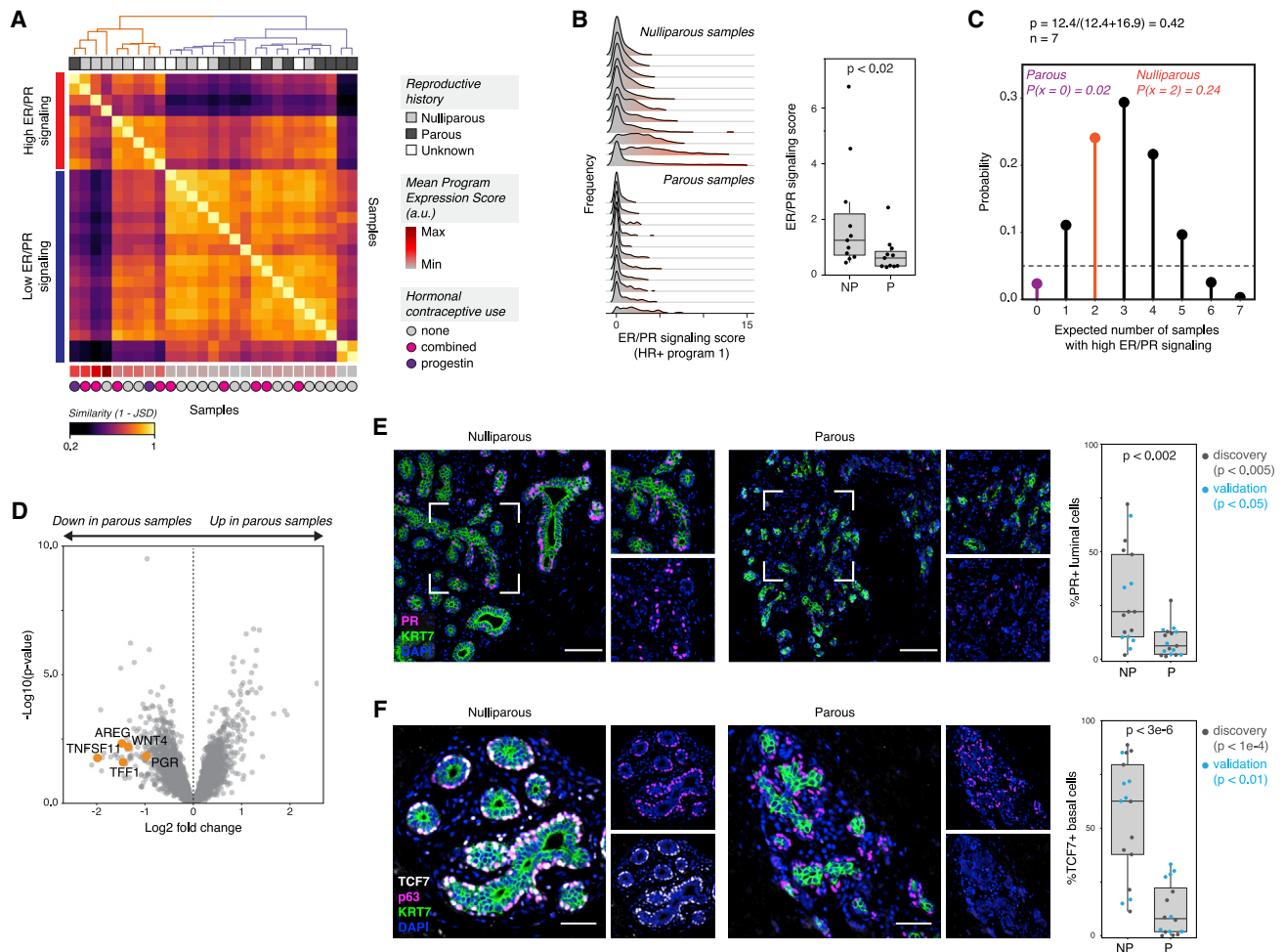


Figure 5. The ER/PR signaling response of HR+ luminal cells is reduced in parous women

(A) Heatmap showing the similarity between each sample's single-cell expression score distribution across HR+ activity program 1 (ER/PR signaling), measured as $(1 - \text{Jensen-Shannon distance})$. Hierarchical clustering (complete linkage) identifies two sets of samples representing high or low expression of the "ER/PR signaling" gene program. The mean expression score for HR+ activity program 1 is annotated at the bottom of the heatmap (arbitrary units, linear scale).

(B) Ridge plots depicting the distribution of HR+ program 1 (ER/PR signaling) expression in HR+ luminal cells across nulliparous (NP) versus parous (P) samples and quantification of the average expression score for HR+ program 1 ($n = 22$ samples, $p < 0.02$, Mann-Whitney test). Data are represented as individual points; box indicates the median and interquartile range (IQR) for 11 nulliparous and 11 parous samples; whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$.

(C) Binomial probability distribution for the expected number of samples with high ER/PR signaling. The binomial probability of high ER/PR signaling is modeled as the average length of the luteal phase of the menstrual cycle, in days, divided by the average total length of the menstrual cycle ($p = 0.42$) (Bull et al., 2019).

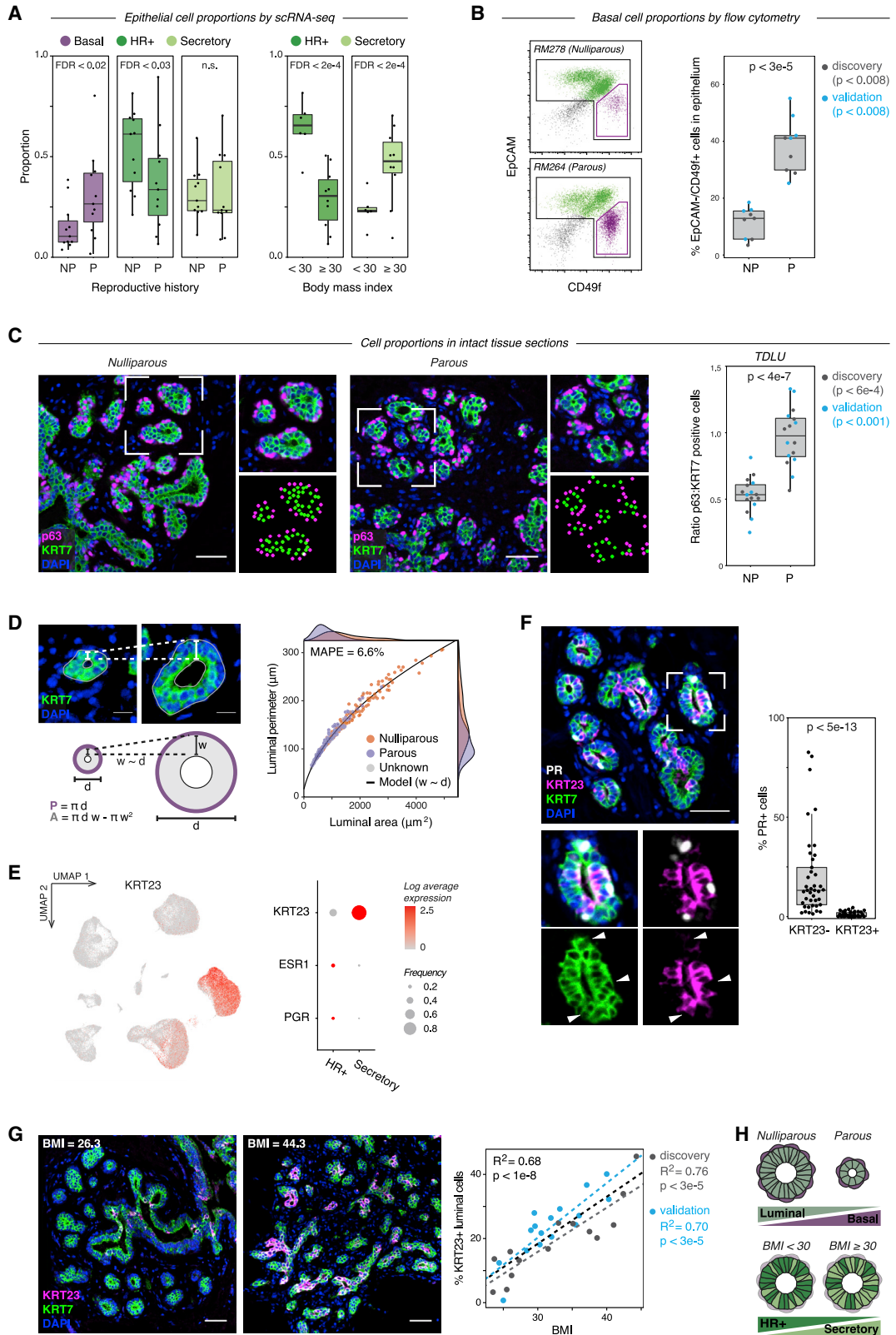
(D) Volcano plot highlighting the differential expression of canonical hormone-responsive genes between parous and nulliparous "pseudo-bulk" samples in HR+ luminal cells. Dots represent individual genes.

(E) Immunostaining for PR and KRT7 and quantification of the percentage of PR+ cells within the KRT7+ luminal compartment for nulliparous (NP) versus parous (P) samples ($n = 34$ samples, $p < 0.002$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples ("discovery set," $n = 19$ samples, $p < 0.005$) and a second independent cohort of samples ("validation" set, $n = 15$ samples, $p < 0.05$). Scale bars, 100 μm . Data are represented as individual points; box indicates the median and interquartile range (IQR) for the combined dataset ($n = 17$ nulliparous samples and 17 parous samples); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$.

(F) Immunostaining for TCF7, p63, and KRT7, and quantification of the percentage of TCF7+ cells within the p63+ basal/myoepithelial cell compartment for nulliparous (NP) versus parous (P) samples ($n = 33$ samples, $p < 3e-6$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples ("discovery set," $n = 18$ samples, $p < 1e-4$) and a second independent cohort of samples ("validation" set, $n = 15$ samples, $p < 0.01$). Scale bars, 50 μm . Data are represented as individual points; box indicates the median and interquartile range (IQR) for the combined dataset ($n = 17$ nulliparous samples and 16 parous samples); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$.

TCF7 expression was markedly decreased in P samples (overall $p < 3e-6$, "discovery" set $p < 1e-4$, "validation" set $p < 0.01$, Mann-Whitney test; Figure 5F). Again, this decrease was not due to differences in lobular architecture, as TCF7 staining was reduced in both ducts and TDLUs in P samples (Figure S11D).

Together, these data demonstrate that ER/PR signaling is a source of transcriptional variation among HR+ luminal cells, that transcription along this axis (HR+ activity program 1) is reduced in women with prior history of pregnancy, and that these transcriptional changes in HR+ cells coincide with a



(legend on next page)

reduction in downstream paracrine signaling to basal/myoepithelial cells.

Parity and BMI influence epithelial cell proportions

Based on our previous finding that paracrine signaling from HR+ luminal cells to other epithelial cell types is strongly influenced by the proportion of HR+ cells in the epithelium (Figures 2D and 2E), we next asked whether the architectural changes associated with parity would contribute to systematic changes in epithelial cell proportions, and thus influence hormone responsiveness across samples. The breast undergoes a major expansion of the mammary epithelium during pregnancy, followed by a regression back toward the pre-pregnant state after weaning in a process called involution. Following involution, the epithelial architecture remains distinct from that of women without prior pregnancy, consisting of larger TDLUs containing greater numbers of acini. At the same time, individual acini are reduced in size (Russo et al., 1992).

We focused our initial analysis on the 63,583 cells in the live/singlet and epithelial sort gates to get an unbiased view of how the epithelial composition of the breast changes with pregnancy. The proportion of basal/myoepithelial cells in the epithelium was approximately 2-fold higher in women with prior history of pregnancy (P) relative to women without prior pregnancy (NP) (Figures 6A and S12A; FDR < 0.02, Wald test with post hoc multiple-comparisons test). This effect remained significant when we controlled for menstrual cycle stage and/or exogenous hormones (i.e., hormonal contraceptive use) using our previously identified “ER/PR signaling” score (Table S8). We confirmed these results in an expanded cohort of samples using three additional methods. First, we measured basal cell proportions by flow cytometry analysis of EpCAM and CD49f. Consistent with scRNA-seq clustering results, parity was associated with an in-

crease in the average proportion of EpCAM[−]/CD49f⁺ basal cells from about 15% to about 40% of the epithelium (Figure 6B; overall $p < 3e-5$, “discovery” set $p < 0.008$, “validation” set $p < 0.008$, Mann-Whitney test). The proportion of basal cells did not vary with other discriminating factors such as BMI or hormonal contraceptive use but was weakly associated with age ($R^2 = 0.20$, $p < 0.04$, Wald test) (Figure S12C). To determine the relative effect of each factor, we performed multiple linear regression analysis and found that the basal cell fraction positively correlated with pregnancy history ($p < 2e-05$, Wald test), but not age ($p = 0.17$, Wald test) (Figure S12D; Table S9; $R^2 = 0.77$, $p < 8e-6$).

Dissociation of tissue for scRNA-seq or FACS may affect measurements of cell composition. We therefore performed two further analyses to confirm these findings in intact tissue. First, we reanalyzed two previously published microarray datasets of total RNA isolated from core needle biopsies from either premenopausal ($n = 71$ P/42 NP) or postmenopausal ($n = 79$ P/30 NP) women (Peri et al., 2012; Santucci-Pereira et al., 2019) and confirmed a significant increase in the basal/myoepithelial markers KRT5, KRT14, and TP63 relative to luminal markers in P samples (Figure S12E). Second, we performed immunostaining and confirmed an approximately 2-fold increase in the ratio of p63⁺ basal cells to KRT7⁺ luminal cells in intact tissue sections (Figure 6C; overall $p < 4e-7$, “discovery” set $p < 6e-4$, “validation” set $p < 0.001$, Mann-Whitney test). Immunostaining demonstrated that this change in epithelial proportions was specific to TDLUs rather than ducts (Figures 6C and S12F). We hypothesized that the increased frequency of basal/myoepithelial cells observed in P women could be explained, in part, by changes in TDLU architecture following pregnancy. To test this, we performed a morphometric comparison of TDLUs between P and NP samples in our dataset. Consistent with

Figure 6. Reproductive history and body mass index are associated with epithelial cell proportions

- (A) Quantification of the proportion of the indicated cell types by scRNA-seq for nulliparous versus parous samples ($n = 22$ samples, Wald test) and obese versus non-obese samples ($n = 16$ samples, Wald test). Data are represented as individual points; box indicates the median and interquartile range (IQR) (left: $n = 11$ nulliparous and 11 parous samples. Right: $n = 6$ samples with BMI < 30 and 10 samples with BMI ≥ 30); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$.
- (B) Representative flow cytometry analysis of the percentage of EpCAM[−]/CD49f⁺ basal cells within the Lin[−] epithelial population and quantification of the percentage of basal cells in nulliparous (NP) versus parous (P) women ($n = 18$ samples; $p < 3e-5$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples (“discovery set,” $n = 9$ samples, $p < 0.008$) and a second independent cohort of samples (“validation” set, $n = 9$ samples, $p < 0.008$). Data are represented as individual points; box indicates the median and interquartile range (IQR) for the combined dataset ($n = 9$ nulliparous and 9 parous samples); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$.
- (C) Immunostaining for the basal/myoepithelial marker p63 and pan-luminal marker KRT7 in terminal ductal lobular units (TDLUs), and quantification of the ratio of p63⁺ basal cells to KRT7⁺ luminal cells in nulliparous (NP) versus parous (P) women ($n = 32$ samples; $p < 4e-7$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples (“discovery set,” $n = 17$ samples, $p < 6e-4$) and a second independent cohort of samples (“validation” set, $n = 15$ samples, $p < 0.001$). Data are represented as individual points; box indicates the median and interquartile range (IQR) for the combined dataset ($n = 16$ nulliparous and 16 parous samples); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$. Scale bars, 50 μm .
- (D) Two-dimensional geometric model of the relative space available for basal cells (luminal perimeter, P) and luminal cells (luminal area, A) within individual acini. Acini were modeled as hollow circles with a shell thickness (w) proportional to their diameter (d). Dots represent measurements of individual acini from TDLUs in parous ($n = 158$ acini from 15 samples) or nulliparous ($n = 164$ acini from 16 samples) specimens as indicated. Line represents results from geometric model (mean absolute percentage error = 6.6%). Scale bars, 15 μm .
- (E) Left: UMAP depicting log normalized expression of KRT23 in reduction mammoplasty samples (GSE198732). Right: dot plot depicting the log normalized mean and frequency of KRT23, ESR1, and PGR expression across luminal cell types.
- (F) Co-immunostaining of PR, KRT23, and the pan-luminal marker KRT7, and quantification of the percentage of PR⁺ cells within the KRT23[−] and KRT23⁺ luminal cell populations ($n = 41$ samples; $p < 5e-13$, Mann-Whitney test). Data are represented as individual points; box indicates the median and interquartile range (IQR) for 41 samples; whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$. Scale bars, 50 μm .
- (G) Co-immunostaining of KRT23 and KRT7 and linear regression analysis of the percentage of KRT23⁺ luminal cells versus BMI ($n = 30$ samples; $R^2 = 0.68$, $p < 1e-8$, Wald test). Scale bars, 50 μm . Results are shown for a subset of the original cohort of sequenced samples (“discovery set,” $n = 14$ samples; $R^2 = 0.76$, $p < 3e-5$) and a second independent cohort of samples (“validation” set, $n = 16$ samples; $R^2 = 0.70$, $p < 3e-5$). Data are represented as individual points; dotted lines represent the best-fit lines for the discovery cohort (light gray), validation cohort (blue), and combined cohort (dark gray).
- (H) Summary of changes in epithelial cell proportions with prior pregnancy and obesity (BMI ≥ 30).

previous reports (Russo et al., 1992), we observed a marked decrease in the average diameter of individual acini in P women (Figure S12G; $p < 4e-5$, Mann-Whitney test). Additionally, we found that the average thickness of the luminal cell layer increased in proportion to acinus diameter (Figure S12H; $R^2 = 0.75$, $p < 3e-16$) and was thus higher in NP women (Figure S12I; $p < 7e-7$, Mann-Whitney test). These results were independent of ER/PR signaling and thus cannot be explained by differences in menstrual cycle stage (Table S10).

To determine how these parameters influence the relative proportions of each cell type, we implemented a simple geometric model (Figure 6D; STAR Methods). When normalized to cross-sectional area (for luminal cells) or perimeter (for basal cells), there was no change in mean luminal cell density or basal cell coverage between P versus NP samples (Figure S12J). Across all samples, the number of basal or luminal cells per acinus was proportional to the space available for each cell type (Figure S12K). Geometric modeling accurately predicted the relationship between the luminal area and outer perimeter for individual acini (mean absolute percentage error loss = 6.6%) and demonstrated that as individual acini increased in size, the space available for luminal cells (luminal area) increased at a faster rate than the space available for basal cells (luminal perimeter) (Figure 6D). Thus, the observed differences in epithelial cell proportions between P and NP samples are not due to a change in basal/myoepithelial coverage, but rather a change in the overall morphology of the luminal layer (e.g., thickness and/or diameter) and relative surface area of individual acini in P women.

While parity was associated with a decreased overall proportion of luminal cells in the epithelium, the proportions of individual HR+ and secretory subtypes within the luminal compartment were highly variable. Consistent with previous work (Meier-Abt et al., 2014; Muenst et al., 2017), we observed reduced frequencies of HR+ luminal cells in P women. However, the proportion of secretory luminal cells was not associated with parity (Figure 6A). Together, these data suggested that additional factors influence the relative proportion of HR+ versus secretory cells within the luminal compartment. Therefore, we performed linear regression analysis to test for the effects of parity, BMI, age, and hormonal contraceptive use on the proportions of HR+ versus secretory luminal cells. We found that the relative proportion of HR+ luminal cells versus secretory luminal cells was reduced in obese (BMI ≥ 30) women (Figures 6A and S12B; FDR < 0.0002 , Wald test with post hoc multiple-comparisons test) and did not vary significantly with other discriminating factors such as age, reproductive history, or hormonal contraceptive use (Figure S13A; Wald test with post hoc multiple-comparisons test). On a continuous scale, every 12 units of BMI was associated with a 2-fold reduction in the proportion of HR+ cells in the luminal compartment (Figure S13B; FDR < 0.001 , Wald test with post hoc multiple-comparisons test). Similar to our previous results, this effect remained significant when we controlled for ER/PR signaling (Table S11).

One limitation of this dataset derived from RM tissue was that all samples classified as non-obese were from NP women less than 24 years old, whereas obese samples were more likely to be from P and older age women (Table S1; Figure S13C). Therefore, we performed scRNA-seq analysis on an independent set of breast core biopsies from healthy premenopausal women

who donated tissue to the Komen Tissue Bank (KTB) (Figures S13D and S13E; Table S2). In contrast with the RM cohort, the KTB cohort consisted of older (37–47 years) P samples with BMI in the normal or overweight range (BMI 20.7–28.3) (Table S1; Figure S13C). Using the RM cohort as a training set, we accurately predicted the proportion of HR+ luminal cells in the KTB cohort as a function of BMI with a mean absolute percentage error of 14.8% (Figure S13F).

We next attempted to measure the relative proportion of the HR+ luminal lineage *in situ* by performing immunostaining for ER and PR. As in our previous analysis, we included samples from both our original sequenced cohort of samples (“discovery” set) and a second independent cohort of samples (“validation” set). The “validation” set was well-balanced across age and BMI, overcoming a limitation of the “discovery” set (Figure S12G). There was a weak trend toward decreased expression of ER and/or PR with increasing BMI, but the change was not statistically significant in the “validation” set or the combined cohort (Figure S13H). Consistent with the heterogeneous ESR1 and PGR transcript expression we observed in scRNA-seq data (Figure S1H), ER and PR protein expression by immunostaining was variable and partly non-overlapping, ranging from 11% to 71% overlap (Figure S13H, bottom panel). We hypothesized that the variability in hormone receptor staining was due to changes in ER/PR expression, stability, and nuclear localization that have all been previously observed based on hormone receptor activation status (Battersby et al., 1992; Métivier et al., 2003; Petz and Nardulli, 2000). Based on this, we predicted that ER transcript and protein levels would co-vary across samples due to the overall proportion of HR+ luminal cells and their hormonal microenvironment but would be stochastically expressed in individual cells at any one time due to fluctuations in mRNA and protein expression, localization, and stability. To test this, we performed co-immunostaining and RNA FISH and confirmed that although ER transcript and protein levels correlate across tissue sections ($R^2 = 0.60$, $p < 0.01$), they do not correlate on a per-cell basis ($\rho = 0.63$, Wilcoxon signed-rank test)—on average, only 31% of cells expressing ESR1 transcript also expressed ER protein (Figure S13I). Expression of ESR1 or PGR transcript was highly specific for cells in the HR+ luminal cluster, although the sensitivity of each transcript for the HR+ cluster was low and varied across individuals (Figure S13J). Thus, these data demonstrate that immunostaining or RNA FISH for nuclear hormone receptors underestimates the fraction of cells in the HR+ lineage and that lack of ER/PR expression cannot be used to reliably define a cell as part of the secretory versus HR+ luminal cell lineages.

On the basis of these results, we sought to identify another marker to distinguish between the luminal lineages, and identified keratin 23 (KRT23) as highly enriched in the secretory luminal cell cluster (Figure 6E), as was also reported by a previous scRNA-seq study (Nguyen et al., 2018). Immunohistochemistry for KRT23 and PR or ER confirmed that these proteins are expressed in mutually exclusive luminal populations (Figures 6F, S13K, and S13L). KRT23 thus represents a discriminatory marker between the two luminal populations. Staining in an expanded cohort of intact tissue sections confirmed that the proportion of KRT23⁺ secretory luminal cells increased by about 20% for every 10-unit increase in BMI (Figure 6G; overall

$R^2 = 0.68$, $p < 1e-8$; “discovery” set $R^2 = 0.76$, $p < 3e-5$; “validation” set $R^2 = 0.70$, $p < 3e-5$). Using multiple regression analysis, we confirmed that the proportion of KRT23-positive cells in the luminal compartment was significantly associated with BMI, but not parity or age (Figure S13M). Together, these data demonstrate that there are two independent effects of reproductive history and body weight on cell proportions in the mammary epithelium: parity affects the ratio of basal to luminal cells whereas BMI affects the ratio of HR+ versus secretory luminal cells (Figure 6H).

Biological variables impact coordinated changes in signaling states across cell types in the breast

Finally, we used the cell-cell interaction network identified by DECIPHER-seq to better understand how biological variables—such as BMI, parity, and hormonal contraceptive use—affect cell-cell interactions in the breast. Based on the above results, we propose that parity and BMI affect the hormone responsiveness of the breast through two distinct mechanisms: parity decreases the per-cell ER/PR signaling response in HR+ luminal cells, whereas BMI indirectly affects hormone signaling by reducing the proportion of HR+ luminal cells in the mammary epithelium (Figure 7A). Consistent with this, both prior pregnancy and increasing BMI were negatively associated with activity programs across the “ER/PR response” module and positively correlated with programs in the “resting state” module (Figures 7B and S14A). To confirm these results in intact tissue sections, we performed immunostaining for PR as a measure of ER activation in HR+ luminal cells, and for TCF7 as a measure of the downstream paracrine response (Wnt activation) in basal cells. As expected, we found that PR expression in the HR+ (KRT23⁻) luminal cell subpopulation was not significantly different between non-obese and obese women ($p = 0.17$, Mann-Whitney test; Figure 7C) but that Wnt signaling in basal cells was markedly reduced in obese samples ($p < 3e-5$, Mann-Whitney test; Figure 7D). We confirmed these results using multiple linear regression to simultaneously test the effects of prior pregnancy and obesity. Whereas PR expression in HR+ (KRT23⁻) cells was dependent on parity but not obesity, downstream Wnt signaling in basal cells was dependent on both variables (Figure 7E).

Second, we took advantage of the different dynamics of serum estrogen and progestin/progesterone in donors using combined hormonal contraceptives versus those undergoing natural menstrual cycles (Figure S14B) to ask whether activity program expression in the “involution-like” module (module 4) was influenced by the hormonal microenvironment. The natural menstrual cycle is characterized by an initial rise in estrogen levels during the follicular phase of the menstrual cycle followed by a combined surge of estrogen and progesterone during the luteal phase. In contrast, following oral contraceptive use, estrogen and progestin levels rise simultaneously, reach peak concentrations in the blood about 2 h following ingestion, and return fairly rapidly to a steady state level over the following 22 h (Figure S14B; Hampson, 2020). We found that activity programs in the “involution-like” module were highly correlated with the use of combined (estrogen/progestin) oral contraceptives (Figure S14C and S14D). These results suggest that the “involution-like” phenotype is influenced by hormone levels and

dynamics (since exogenous hormones are associated with increased expression across this module) but does not require the precise sequential estrogen/progesterone dynamics observed during natural menstrual cycles (since estrogen and progesterone rise simultaneously upon oral contraceptive ingestion). Overall, these results demonstrate how sample-to-sample variation in the breast can be used to predict how specific changes in transcriptional cell state and cell-type proportions influence cell-cell interactions in a tissue and to understand some of the sources of biological variation (e.g., metadata factors) that control the overall state of the tissue (Figure 7F).

DISCUSSION

In this study, we leverage inter-sample transcriptional variation in the breast to identify a set of highly correlated “activity programs” representing the *in situ* response to hormone receptor activation in HR+ cells and the effects of downstream paracrine signaling in other cell types. We uncover additional correlated programs representing the dynamic response of the breast to changing hormone levels (e.g., “involution-like”). Furthermore, we show that person-to-person heterogeneity in hormone responsiveness in the breast is directly linked to two factors known to be correlated with premenopausal breast cancer risk—reproductive history and BMI.

Cumulative lifetime hormone exposure is a major determinant of breast cancer risk (Collaborative Group on Hormonal Factors in Breast Cancer, 2012). Here, we mapped the coordinated changes in cell state that occur in response to paracrine signaling from HR+ luminal cells. Notably, many of these changes closely mimic those seen during the pregnancy/involution cycle that have been linked to a transient increased breast cancer risk following pregnancy (Lyons et al., 2011; O’Brien et al., 2010; Schedin et al., 2007). First, we identify a proliferative gene signature in secretory luminal cells that is highly correlated with hormone signaling in HR+ luminal cells, consistent with previous studies demonstrating that TNFSF11 (RANKL) and Wnt control progesterone-mediated epithelial proliferation (Joshi et al., 2015). Second, we identify previously uncharacterized subpopulations of HR+ and secretory luminal cells in the cycling premenopausal breast with transcriptional signatures closely matching that described for post-lactational involution (Clarkson et al., 2004; Stein et al., 2004), including upregulation of immune mediators, MHC class II molecules, and the phagocytic receptor MARCO. This idea that the menstrual cycle mimics a miniature pregnancy/involution cycle is consistent with studies showing that the fraction of apoptotic cells in the epithelium peaks between the late luteal and early follicular phases (Anderson et al., 1982). We also observe upregulation of hypoxic gene signatures in multiple epithelial and stromal cell types that are highly correlated with hormone signaling in HR+ cells. A previous study identified these same pathways as highly enriched during post-lactational involution in the mouse. More importantly from the perspective of breast cancer risk, this “hypoxia/pro-angiogenic” signature identified breast cancers with increased metastatic activity (Stein et al., 2009), suggesting that these pathways can be co-opted by cancer cells to support a permissive tumor microenvironment. Thus, we speculate that many of the same mechanisms underlie both the short-term increased breast cancer

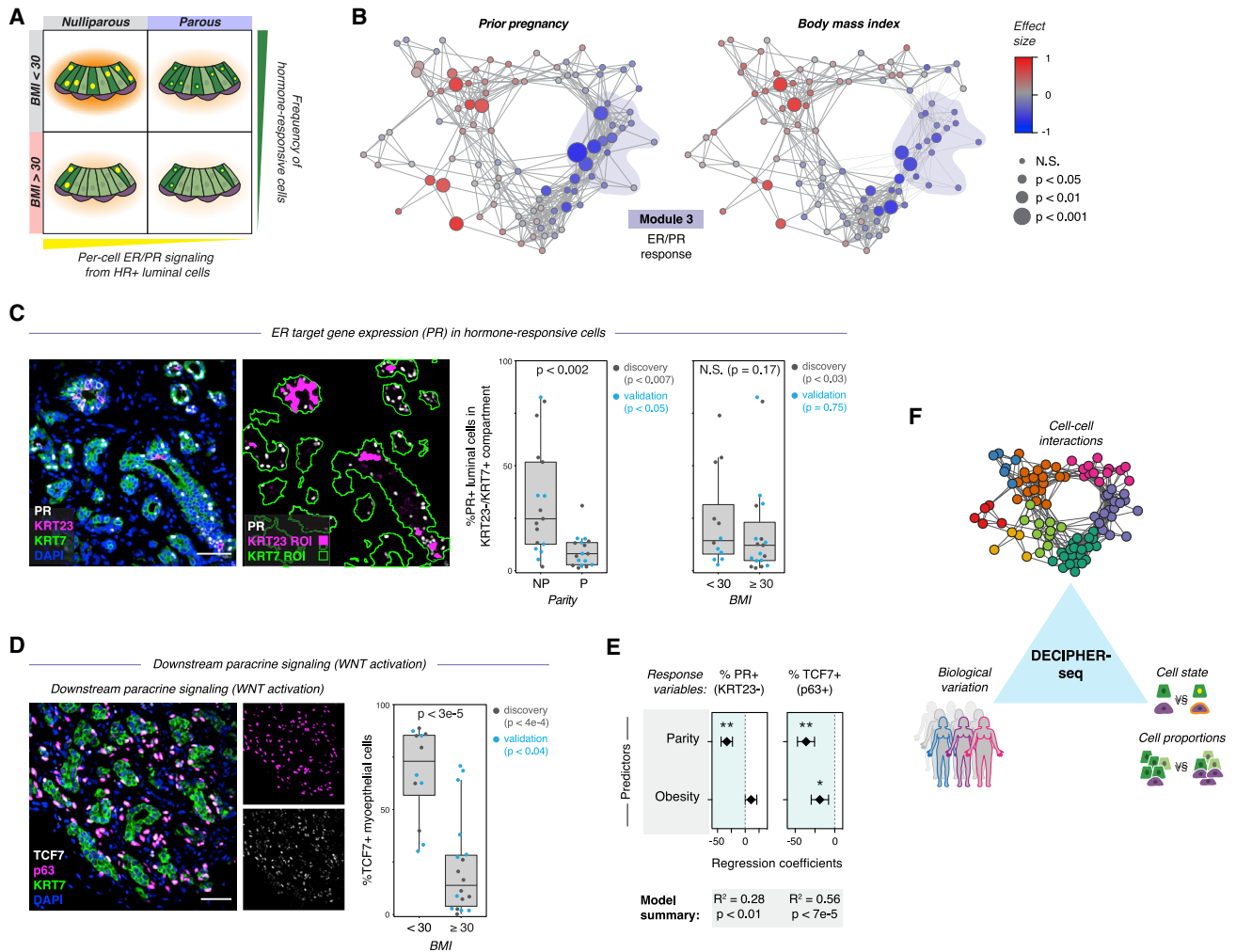


Figure 7. Biological variables are linked to predicted tissue states

(A) Schematic depicting the model for how parity and obesity impact hormone signaling in the breast through distinct mechanisms. Parity affects per-cell ER/PR signaling in HR+ luminal cells, and obesity (BMI ≥ 30) leads to a reduction in the proportion of hormone-responsive (HR+) luminal cells in the epithelium.

(B) Network graph of activity programs in the human breast, colored by the effect size of prior pregnancy (Wilcoxon effect size) or body mass index (Pearson correlation coefficient) on each activity program (linear scale). Significant positive and negative associations ($p < 0.05$) are represented by larger nodes (prior pregnancy: Mann-Whitney test; BMI: Wald test).

(C) Co-immunostaining of the estrogen receptor target gene PR, KRT23, and the pan-luminal marker KRT7 and quantification of the percentage of PR⁺ cells in the KRT23⁺/KRT7⁺ luminal cell population for nulliparous (NP) versus parous (P) samples ($n = 34$ samples, $p < 0.002$, Mann-Whitney test) or non-obese (BMI < 30) versus obese (BMI ≥ 30) samples ($n = 31$ samples, $p = 0.17$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples (“discovery” set) and a second independent cohort of samples (“validation” set). Data are represented as individual points; boxes indicate the median and interquartile range (IQR) for the combined datasets (Left: $N = 17$ nulliparous and 17 parous samples. Right: $n = 12$ samples with BMI < 30 and 19 samples with BMI ≥ 30); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$. Scale bars, 50 μm .

(D) Immunostaining for TCF7, p63, and KRT7, and quantification of the percentage of TCF7⁺ cells within the p63⁺ basal/myoepithelial cell compartment for non-obese (BMI < 30) versus obese (BMI ≥ 30) samples ($n = 30$ samples, $p < 3e-5$, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples (“discovery set,” $n = 14$ samples, $p < 4e-4$) and a second independent cohort of samples (“validation” set, $n = 16$ samples, $p < 0.04$). Data are represented as individual points; box indicates the median and interquartile range (IQR) for the combined dataset ($n = 12$ samples with BMI < 30 and 18 samples with BMI ≥ 30); whiskers extend from $Q1 - 1.5IQR$ to $Q3 + 1.5IQR$. Scale bars, 50 μm .

(E) Results from multiple linear regression analysis, with prior pregnancy (parity) and obesity (BMI ≥ 30) as predictors and the percentage of PR⁺ cells in the KRT23⁺/KRT7⁺ luminal cell population or of TCF7⁺ cells in the p63⁺ basal cell compartment as response variables. Points represent the regression coefficient for each predictor, and error bars depict the standard error.

(F) Summary of results. DECIPHER-seq predicts how specific changes in transcriptional cell state and cell-type proportions influence cell-cell interactions in the human breast and links specific sources of biological variation (e.g., Parity and BMI) to the overall signaling state of the tissue.

risk following pregnancy and the lifetime increased risk due to total number of menstrual cycles.

Pregnancy has two opposing effects on breast cancer risk: although breast cancer risk is increased for the first 5–10 years following pregnancy, it also has a pronounced long-term protective effect, with up to a 50% reduction in ER+/PR+ breast cancer risk for women with multiple full-term pregnancies at a young age (Britt et al., 2007). The cellular basis for this long-term protective effect remains an area of active research. Our analysis revealed that parity is associated with a stark increase in the proportion of basal and/or myoepithelial cells within the breast epithelium, as well as decreased hormone signaling in HR+ luminal cells. While the precise role of myoepithelial cells during cancer progression remains an active area of research (Risom et al., 2022), previous work has described two properties of myoepithelial cells consistent with a tumor-protective effect: they are resistant to malignant transformation (Koren et al., 2015; Proia et al., 2011) and may also act as a dynamic barrier to prevent tumor cell invasion (Sirka et al., 2018; Sternlicht et al., 1997). Thus, our data support the notion that pregnancy protects against breast cancer risk through multiple mechanisms: by decreasing the relative frequency of luminal cells—the tumor cell-of-origin for most breast cancer subtypes (Keller et al., 2012; Melchor et al., 2014; Molyneux et al., 2010), by reducing the overall hormone responsiveness of HR+ cells and subsequent pro-tumorigenic microenvironmental changes, and by suppressing progression to invasive carcinoma (Sirka et al., 2018; Sternlicht et al., 1997).

Finally, we found that paracrine signaling from HR+ cells to other cell types depends on both the magnitude of signaling from HR+ cells and the overall proportion of HR+ cells in the epithelium. Prior pregnancy and obesity are specifically associated with a reduced risk of ER+/PR+ breast cancer in premenopausal women (Fortner et al., 2019; Premenopausal Breast Cancer Collaborative Group et al., 2018), and our data are consistent with a model that these biological variables lead to reduced paracrine signaling downstream of estrogen and progesterone via two distinct mechanisms. First, parity leads to a reduced per-cell hormone signaling response in HR+ luminal cells. Second, we identify a marked decrease in the ratio of HR+ cells relative to secretory luminal cells with increasing BMI. Both changes are associated with reduced paracrine signaling across the ER/PR response module.

Several potential mechanisms could account for the decreased hormone signaling response observed in HR+ luminal cells in P women. Previous studies have identified small reductions in the levels of estrogen metabolites in the urine of P women, which may be indicative of lower serum levels of estradiol (Barrett et al., 2014). Since PR expression is induced downstream of ER activation, lower levels of serum estradiol could lead to reduced signaling through both ER and PR. A second possibility is that structural differences in the mammary epithelium of P women, such as increased lobular density (Russo et al., 1992) or alterations in vascularization could lead to decreased access of hormones to HR+ luminal cells. Finally, changes in the differentiation state or epigenetic remodeling of HR+ luminal cells following pregnancy could lead to a direct change in the ability of these cells to respond to hormone. Interestingly, recent work has shown that matrix stiffness and/or compressive stress is required for maintenance of ER expres-

sion in explant cultures, via H3K27me3-dependent epigenetic regulation (Munne et al., 2021), and previous work has shown that parity is associated with decreased mammographic density (Vachon et al., 2000). Further studies are required to determine whether the decreased hormone response of HR+ luminal cells in P women is a result of these or other processes.

A key insight of our computational approach is that a subset of “high-confidence” direct cell-cell interactions can be inferred based on their dependence on the proportion of one cell type in the tissue. Because the DECIPHER-seq workflow corrects for batch effects while maintaining meaningful biological variation and optimizes both the granularity and robustness of identified activity programs, it has the potential to be flexibly adapted to a broad range of preexisting single-cell datasets, or across datasets from multiple sources. Further, we find that the coordinated activity programs in our dataset naturally self-organize into a cycle, precisely as we would expect based on hormone fluctuations across menstrual cycles. This raises the intriguing possibility that a similar computational approach could reveal cyclical cellular programs in other tissue types in the body, such as circadian rhythms, feeding cycles, or the response to wounding. While we focus on single-cell transcriptional data in this study, iNMF has also been applied to multi-omic datasets containing spatial or epigenetic data together with transcriptional information (Welch et al., 2019; Gao et al., 2021).

In summary, using scRNA-seq of a unique cohort of 28 healthy premenopausal women, we provide a comprehensive, systems-level view of the cellular and transcriptional variation within the human breast, which profoundly affects the response to hormones and may impact breast cancer risk. As the human breast is one of the only human organs that undergoes repeated cycles of morphogenesis and involution, this study serves as a roadmap for deeper interrogation of the cell state changes associated with hormone dynamics. Finally, it provides a foundation for future systems-level studies dissecting how the paracrine communication networks downstream of hormone signaling are altered during ER+/PR+ breast cancer progression.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [RESOURCE AVAILABILITY](#)
 - Lead contact
 - Materials availability
 - Data and code availability
- [EXPERIMENTAL MODEL AND SUBJECT DETAILS](#)
 - Human tissue samples
- [METHOD DETAILS](#)
 - Tissue processing
 - Dissociation to single cells
 - MULTI-seq sample barcoding (Batches 3, 4, and KTB)
 - Sorting for scRNA-seq
 - scRNA-seq library preparation
 - Expression library pre-processing
 - Cell calling

- MULTI-seq barcode library pre-processing
- Sample demultiplexing
- Dataset integration and cell type identification
- PC analysis of individual cell types
- Activity program identification in each cell type
- Network clustering of correlated activity programs
- Fluorescent immunohistochemistry
- Morphometric analysis and geometric modeling
- RNA FISH analysis of ESR1 transcripts
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
 - Quantification of sample-to-sample heterogeneity
 - Testing for changes in cell type proportions
 - Identification of non-cell-type specific programs
 - Inferring direct cell-cell interactions
 - Gene set enrichment analysis
 - Enrichment of gene sets within modules
 - Sample-to-sample variability in ER/PR signaling
 - Pseudo-bulk differential gene expression analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cels.2022.06.005>.

ACKNOWLEDGMENTS

We thank Drs. Tom Norman and Jonathan Weissman for technical support and for generously providing access to equipment and computing resources. Sequencing was performed in the Center for Advanced Technology at UCSF. Tissue samples were provided by the Cooperative Human Tissue Network (CHTN), which is funded by the National Cancer Institute. Other investigators may have received specimens from the same subjects. Samples from the Susan G. KTB at the IU Simon Cancer Center were used in this study. We thank contributors, including Indiana University who collected samples used in this study, as well as donors and their families, whose help and participation made this work possible. This research was supported by grants from the Department of Defense Breast Cancer Research Program (W81XWH-10-1-1023 and W81XWH-13-1-0221), NIH (U01CA199315 and DP2 HD080351-01), the NSF (MCB-1330864), and the UCSF Center for Cellular Construction (DBI-1548297), an NSF Science and Technology Center, to Z.J.G. Z.J.G. is a Chan Zuckerberg BioHub Investigator. L.M.M. is a former Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-2239-15).

AUTHOR CONTRIBUTIONS

L.M.M., R.J.W., and Z.J.G. conceived the project. L.M.M., J.A.C., R.J.W., C.S.M., and K.P. performed the sequencing experiments. C.S.M. generated aligned reads and barcode matrices. C.S.M. and L.M.M. performed sample demultiplexing. P.G. coordinated sample acquisition and provided critical guidance for sample selection. L.M.M. and J.A.C. performed fluorescent immunohistochemistry and RNA-FISH experiments. L.M.M. and J.A.C. performed flow cytometry experiments. L.M.M. and J.A.C. performed histopathology on tissue sections. A.D.B. performed histopathological tissue analysis. L.M.M. analyzed and visualized the data. L.M.M., C.S.M., and G.R. wrote and tested the code used in data analysis. M.T. provided guidance in data analyses and computational approaches. T.T. and A.D.B. provided guidance in human breast biology. T.T., M.T., and Z.J.G. provided critical resources. T.A.D., M.T., T.T., and Z.J.G. supervised the project. L.M.M. and Z.J.G. wrote the manuscript. All authors reviewed and edited the manuscript.

DECLARATION OF INTERESTS

Z.J.G. and C.S.M. hold patents related to the MULTI-seq barcoding method. Z.J.G. is an equity holder in Scribe Biosciences and Provenance bio and a member of the scientific advisory board of Serotiny Bio. C.S.M. is a consultant

for ImYoo. Since January 10, 2022, L.M.M. is an employee of Genentech, a member of the Roche group.

INCLUSION AND DIVERSITY

We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as living with a disability.

Received: February 2, 2021

Revised: March 2, 2022

Accepted: June 22, 2022

Published: July 20, 2022

REFERENCES

- Anderson, T.J., Ferguson, D.J., and Raab, G.M. (1982). Cell turnover in the “resting” human breast: influence of parity, contraceptive pill, age and laterality. *Br. J. Cancer* 46, 376–382. <https://doi.org/10.1038/bjc.1982.213>.
- Andruska, N., Zheng, X., Yang, X., Helferich, W.G., and Shapiro, D.J. (2015). Anticipatory estrogen activation of the unfolded protein response is linked to cell proliferation and poor survival in estrogen receptor α -positive breast cancer. *Oncogene* 34, 3760–3769. <https://doi.org/10.1038/onc.2014.292>.
- Aupperlee, M.D., Leipprandt, J.R., Bennett, J.M., Schwartz, R.C., and Haslam, S.Z. (2013). Amphiregulin mediates progesterone-induced mammary ductal development during puberty. *Breast Cancer Res.* 15, R44. <https://doi.org/10.1186/bcr3431>.
- Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P.V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Meth.* 16, 695–698. <https://doi.org/10.1038/s41592-019-0466-z>.
- Barrett, E.S., Parlett, L.E., Windham, G.C., and Swan, S.H. (2014). Differences in ovarian hormones in relation to parity and time since last birth. *Fertil. Steril.* 101, 1773–1780.e1. <https://doi.org/10.1016/j.fertnstert.2014.02.047>.
- Battersby, S., Robertson, B.J., Anderson, T.J., King, R.J., and McPherson, K. (1992). Influence of menstrual cycle, parity and oral contraceptive use on steroid hormone receptors in normal breast. *Br. J. Cancer* 65, 601–607. <https://doi.org/10.1038/bjc.1992.122>.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* 57, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Bhat-Nakshatri, P., Gao, H., Sheng, L., McGuire, P.C., Xuei, X., Wan, J., Liu, Y., Althouse, S.K., Colter, A., Sandusky, G., et al. (2021). A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep. Med.* 2, 100219. <https://doi.org/10.1016/j.xcrm.2021.100219>.
- Britt, K., Ashworth, A., and Smalley, M. (2007). Pregnancy and the risk of breast cancer. *Endocr. Relat. Cancer* 14, 907–933. <https://doi.org/10.1677/ERC-07-0137>.
- Bull, J.R., Rowland, S.P., Scherwitzl, E.B., Scherwitzl, R., Danielsson, K.G., and Harper, J. (2019). Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *NPJ Digit. Med.* 2, 83. <https://doi.org/10.1038/s41746-019-0152-7>.
- Butler, A., Hoffman, P., Smibert, P., Papalex, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. <https://doi.org/10.1038/nbt.4096>.
- Cao, Y., Lin, Y., Ormerod, J.T., Yang, P., Yang, J.Y.H., and Lo, K.K. (2019). scDC: single cell differential composition analysis. *BMC Bioinformatics* 20, 721. <https://doi.org/10.1186/s12859-019-3211-9>.
- Clarke, R.B., Howell, A., Potten, C.S., and Anderson, E. (1997). Dissociation between steroid receptor expression and cell proliferation in the human breast. *Cancer Res.* 57, 4987–4991.

- Clarkson, R.W.E., Wayland, M.T., Lee, J., Freeman, T., and Watson, C.J. (2004). Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. *Breast Cancer Res.* 6, R92–R109. <https://doi.org/10.1186/bcr754>.
- Collaborative Group on Hormonal Factors in Breast Cancer (2012). Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* 13, 1141–1151. [https://doi.org/10.1016/S1470-2045\(12\)70425-4](https://doi.org/10.1016/S1470-2045(12)70425-4).
- Dabrosin, C. (2003). Variability of vascular endothelial growth factor in normal human breast tissue *in vivo* during the menstrual cycle. *J. Clin. Endocrinol. Metab.* 88, 2695–2698. <https://doi.org/10.1210/jc.2002-021584>.
- Desper, R., and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* 9, 687–705. <https://doi.org/10.1089/106652702761034136>.
- Dontu, G., and Ince, T.A. (2015). Of mice and women: a comparative tissue biology perspective of breast stem cells and differentiation. *J. Mammary Gland Biol. Neoplasia* 20, 51–62. <https://doi.org/10.1007/s10911-015-9341-4>.
- Drost, H.G. (2018). Philentropy: information theory and distance quantification with R. *Journal Software. JOSS* 3, 26. <https://doi.org/10.21105/joss.00765>.
- Dunphy, K.A., Black, A.L., Roberts, A.L., Sharma, A., Li, Z., Suresh, S., Browne, E.P., Arcaro, K.F., Ser-Dolansky, J., Bigelow, C., et al. (2020). Inter-individual variation in response to estrogen in human breast explants. *J. Mammary Gland Biol. Neoplasia* 25, 51–68. <https://doi.org/10.1007/s10911-020-09446-3>.
- Ewald, A.J., Brenot, A., Duong, M., Chan, B.S., and Werb, Z. (2008). Collective epithelial migration and cell rearrangements drive mammary branching morphogenesis. *Dev. Cell* 14, 570–581. <https://doi.org/10.1016/j.devcel.2008.03.003>.
- Ferguson, J.E., Schor, A.M., Howell, A., and Ferguson, M.W. (1992). Changes in the extracellular matrix of the normal human breast during the menstrual cycle. *Cell Tissue Res.* 268, 167–177. <https://doi.org/10.1007/BF00338066>.
- Fornetti, J., Flanders, K.C., Henson, P.M., Tan, A.-C., Borges, V.F., and Schedin, P. (2016). Mammary epithelial cell phagocytosis downstream of TGF- β 3 is characterized by adherens junction reorganization. *Cell Death Differ.* 23, 185–196. <https://doi.org/10.1038/cdd.2015.82>.
- Fortner, R.T., Sisti, J., Chai, B., Collins, L.C., Rosner, B., Hankinson, S.E., Tamimi, R.M., and Eliassen, A.H. (2019). Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses' Health Studies. *Breast Cancer Res.* 21, 40. <https://doi.org/10.1186/s13058-019-1119-y>.
- Gao, C., Liu, J., Kriebel, A.R., Preissl, S., Luo, C., Castanon, R., Sandoval, J., Rivkin, A., Nery, J.R., Behrens, M.M., et al. (2021). Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.* 39, 1000–1007. <https://doi.org/10.1038/s41587-021-00867-x>.
- Gay, C.M., Zygmunt, T., and Torres-Vázquez, J. (2011). Diverse functions for the semaphorin receptor PlexinD1 in development and disease. *Dev. Biol.* 349, 1–19. <https://doi.org/10.1016/j.ydbio.2010.09.008>.
- Gysi, D.M., Voigt, A., Fragoso, T.M., Almaas, E., and Nowick, K. (2018). wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics* 19, 392. <https://doi.org/10.1186/s12859-018-2351-7>.
- Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al. (2017). A single-cell survey of the small intestinal epithelium. *Nature* 551, 333–339. <https://doi.org/10.1038/nature24489>.
- Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20, 296. <https://doi.org/10.1186/s13059-019-1874-1>.
- Hallberg, G., Andersson, E., Naessén, T., and Ordeberg, G.E. (2010). The expression of syndecan-1, syndecan-4 and decorin in healthy human breast tissue during the menstrual cycle. *Reprod. Biol. Endocrinol.* 8, 35. <https://doi.org/10.1186/1477-7827-8-35>.
- Hampson, E. (2020). A brief guide to the menstrual cycle and oral contraceptive use for researchers in behavioral endocrinology. *Horm. Behav.* 119, 104655. <https://doi.org/10.1016/j.yhbeh.2019.104655>.
- Heaton, H., Talman, A.M., Knights, A., Imaz, M., Gaffney, D.J., Durbin, R., Hemberg, M., and Lawnczak, M.K.N. (2020). Soupcorell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Meth.* 17, 615–620. <https://doi.org/10.1038/s41592-020-0820-1>.
- Hyder, S.M., Nawaz, Z., Chiappetta, C., and Stancel, G.M. (2000). Identification of functional estrogen response elements in the gene coding for the potent angiogenic factor vascular endothelial growth factor. *Cancer Res.* 60, 3183–3190.
- Jindal, S., Gao, D., Bell, P., Albrektsen, G., Edgerton, S.M., Ambrosone, C.B., Thor, A.D., Borges, V.F., and Schedin, P. (2014). Postpartum breast involution reveals regression of secretory lobules mediated by tissue-remodeling. *Breast Cancer Res.* 16, R31. <https://doi.org/10.1186/bcr3633>.
- Joshi, P.A., Waterhouse, P.D., Kannan, N., Narala, S., Fang, H., Di Grappa, M.A., Jackson, H.W., Penninger, J.M., Eaves, C., and Khokha, R. (2015). RANK signaling amplifies WNT-responsive mammary progenitors through R-SPONDIN1. *STEMCR* 5, 31–44. <https://doi.org/10.1016/j.stemcr.2015.05.012>.
- Kastner, P., Krust, A., Turcotte, B., Stropp, U., Tora, L., Gronemeyer, H., and Chambon, P. (1990). Two distinct estrogen-regulated promoters generate transcripts encoding the two functionally different human progesterone receptor forms A and B. *EMBO J.* 9, 1603–1614. <https://doi.org/10.1002/j.1460-2075.1990.tb08280.x>.
- Keller, P.J., Arendt, L.M., Skibinski, A., Logvinenko, T., Klebba, I., Dong, S., Smith, A.E., Prat, A., Perou, C.M., Gilmore, H., et al. (2012). Defining the cellular precursors to human breast cancer. *Proc. Natl. Acad. Sci. USA* 109, 2772–2777. <https://doi.org/10.1073/pnas.1017626108>.
- Koren, S., Reavie, L., Couto, J.P., De Silva, D., Stadler, M.B., Roloff, T., Britschgi, A., Eichlisberger, T., Kohler, H., Aina, O., et al. (2015). PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* 525, 114–118. <https://doi.org/10.1038/nature14669>.
- Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. Preprint at bioRxiv. <https://doi.org/10.1101/060012>.
- Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* 8, 507. <https://doi.org/10.7554/eLife.43803>.
- LaMarca, H.L., and Rosen, J.M. (2007). Estrogen regulation of mammary gland development and breast cancer: amphiregulin takes center stage. *Breast Cancer Res.* 9, 304. <https://doi.org/10.1186/bcr1740>.
- Legland, D., Arganda-Carreras, I., and Andrey, P. (2016). MorphoLibJ: integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics* 32, 3532–3534. <https://doi.org/10.1093/bioinformatics/btw413>.
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. <https://doi.org/10.1016/j.cels.2015.12.004>.
- Lim, E., Wu, D., Pal, B., Bouras, T., Asselin-Labat, M.-L., Vaillant, F., Yagita, H., Lindeman, G.J., Smyth, G.K., and Visvader, J.E. (2010). Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* 12, R21. <https://doi.org/10.1186/bcr2560>.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. <https://doi.org/10.1186/s13059-014-0550-8>.
- Lyons, T.R., O'Brien, J., Borges, V.F., Conklin, M.W., Keely, P.J., Eliceiri, K.W., Marusyk, A., Tan, A.-C., and Schedin, P. (2011). Postpartum mammary gland involution drives progression of ductal carcinoma *in situ* through collagen and COX-2. *Nat. Med.* 17, 1109–1115. <https://doi.org/10.1038/nm.2416>.
- McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019a). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest

- neighbors. *Cell Syst.* 8, 329–337.e4. <https://doi.org/10.1016/j.cels.2019.03.003>.
- McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019b). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Meth.* 30, 1. <https://doi.org/10.1038/s41592-019-0433-8>.
- Meier-Abt, F., Brinkhaus, H., and Bentires-Alj, M. (2014). Early but not late pregnancy induces lifelong reductions in the proportion of mammary progesterone sensing cells and epithelial Wnt signaling. *Breast Cancer Res.* 16, 402. <https://doi.org/10.1186/bcr3626>.
- Melchor, L., Molyneux, G., Mackay, A., Magnay, F.A., Atienza, M., Kendrick, H., Nava-Rodrigues, D., López-García, M.Á., Milanezi, F., Greenow, K., et al. (2014). Identification of cellular and genetic drivers of breast cancer heterogeneity in genetically engineered mouse tumour models. *J. Pathol.* 233, 124–137. <https://doi.org/10.1002/path.4345>.
- Métivier, R., Penot, G., Hübner, M.R., Reid, G., Brand, H., Kos, M., and Gannon, F. (2003). Estrogen receptor- α directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* 115, 751–763. [https://doi.org/10.1016/s0092-8674\(03\)00934-6](https://doi.org/10.1016/s0092-8674(03)00934-6).
- Molyneux, G., Geyer, F.C., Magnay, F.-A., McCarthy, A., Kendrick, H., Natrajan, R., Mackay, A., Grigoriadis, A., Tutt, A., Ashworth, A., et al. (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell* 7, 403–417. <https://doi.org/10.1016/j.stem.2010.07.010>.
- Monks, J., Smith-Steinhart, C., Kruk, E.R., Fadok, V.A., and Henson, P.M. (2008). Epithelial cells remove apoptotic epithelial cells during post-lactation involution of the mouse mammary gland. *Biol. Reprod.* 78, 586–594. <https://doi.org/10.1095/biolreprod.107.065045>.
- Mueller, S.O., Clark, J.A., Myers, P.H., and Korach, K.S. (2002). Mammary gland development in adult mice requires epithelial and stromal estrogen receptor α . *Endocrinology* 143, 2357–2365. <https://doi.org/10.1210/endo.143.6.8836>.
- Muenst, S., Mechera, R., Däster, S., Piscuoglio, S., Ng, C.K.Y., Meier-Abt, F., Weber, W.P., and Soysal, S.D. (2017). Pregnancy at early age is associated with a reduction of progesterone-responsive cells and epithelial Wnt signaling in human breast tissue. *Oncotarget* 8, 22353–22360. <https://doi.org/10.18632/oncotarget.16023>.
- Munne, P.M., Martikainen, L., Rätty, I., Bertula, K., Nonappa, R., Ruuska, J., Ala-Hongisto, H., Peura, A., Hollmann, B., Euro, L., et al. (2021). Compressive stress-mediated p38 activation required for ER α + phenotype in breast cancer. *Nat. Commun.* 12, 6967. <https://doi.org/10.1038/s41467-021-27220-9>.
- Nakshatri, H., Anjanappa, M., and Bhat-Nakshatri, P. (2015). Ethnicity-dependent and -independent heterogeneity in healthy normal breast hierarchy impacts tumor characterization. *Sci. Rep.* 5, 13526. <https://doi.org/10.1038/srep13526>.
- Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* 9, 2028. <https://doi.org/10.1038/s41467-018-04334-1>.
- O'Brien, J., Lyons, T., Monks, J., Lucia, M.S., Wilson, R.S., Hines, L., Man, Y.G., Borges, V., and Schedin, P. (2010). Alternatively activated macrophages and collagen remodeling characterize the postpartum involuting mammary gland across species. *Am. J. Pathol.* 176, 1241–1255. <https://doi.org/10.2353/ajpath.2010.090735>.
- O'Flanagan, C.H., Campbell, K.R., Zhang, A.W., Kabeer, F., Lim, J.L.P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., et al. (2019). Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* 20, 210. <https://doi.org/10.1186/s13059-019-1830-0>.
- Palmieri, C., Saij, S., Sakaguchi, H., Cheng, G., Sunters, A., O'Hare, M.J., Warner, M., Gustafsson, J.-A., Coombes, R.C., and Lam, E.W.-F. (2004). The expression of oestrogen receptor (ER)- β and its variants, but not ER α , in adult human mammary fibroblasts. *J. Mol. Endocrinol.* 33, 35–50. <https://doi.org/10.1677/jme.0.0330035>.
- Pardo, I., Lillemoe, H.A., Blosser, R.J., Choi, M., Sauder, C.A.M., Doxey, D.K., Mathieson, T., Hancock, B.A., Baptiste, D., Atale, R., et al. (2014). Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank. *Breast Cancer Res.* 16, R26. <https://doi.org/10.1186/bcr3627>.
- Parmar, H., and Cunha, G.R. (2004). Epithelial-stromal interactions in the mouse and human mammary gland *in vivo*. *Endocr. Relat. Cancer* 11, 437–458. <https://doi.org/10.1677/erc.1.00659>.
- Pelka, K., Hofree, M., Chen, J.H., Sarkizova, S., Pirl, J.D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W.H., Xu, K.H., et al. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* 184, 4734–4752.e20. <https://doi.org/10.1016/j.cell.2021.08.003>.
- Peri, S., de Cicco, R.L., Santucci-Pereira, J., Slifker, M., Ross, E.A., Russo, I.H., Russo, P.A., Arslan, A.A., Belitskaya-Lévy, I., Zeleniuch-Jacquotte, A., et al. (2012). Defining the genomic signature of the parous breast. *BMC Med. Genomics* 5, 46. <https://doi.org/10.1186/1755-8794-5-46>.
- Petz, L.N., and Nardulli, A.M. (2000). Sp1 binding sites and an estrogen response element half-site are involved in regulation of the human progesterone receptor A promoter. *Mol. Endocrinol.* 14, 972–985. <https://doi.org/10.1210/mend.14.7.0493>.
- Premenopausal Breast Cancer Collaborative Group, Schoemaker, M.J., Nichols, H.B., Wright, L.B., Brook, M.N., Jones, M.E., O'Brien, K.M., Adami, H.-O., Baglietto, L., Bernstein, L., et al. (2018). Association of body mass index and age with subsequent breast cancer risk in premenopausal women. *JAMA Oncol.* 4, e181771. <https://doi.org/10.1001/jamaoncol.2018.1771>.
- Proia, T.A., Keller, P.J., Gupta, P.B., Klebba, I., Jones, A.D., Sedic, M., Gilmore, H., Tung, N., Naber, S.P., Schnitt, S., et al. (2011). Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. *Cell Stem Cell* 8, 149–163. <https://doi.org/10.1016/j.stem.2010.12.007>.
- Prosperi, M.C.F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., et al. (2011). ARCA collaborative group. *Nat. Commun.* 2, 321. <https://doi.org/10.1038/ncomms1325>.
- Rajaram, R.D., Buric, D., Caikovski, M., Ayyanan, A., Rougemont, J., Shan, J., Vainio, S.J., Yalcin-Ozuysal, O., and Brisken, C. (2015). Progesterone and Wnt4 control mammary stem cells via myoepithelial crosstalk. *EMBO J.* 34, 641–652. <https://doi.org/10.15252/emboj.201490434>.
- Ramakrishnan, R., Khan, S.A., and Badve, S. (2002). Morphological changes in breast tissue with menstrual cycle. *Mod. Pathol.* 15, 1348–1356. <https://doi.org/10.1097/01.MP.0000039566.20817.46>.
- Ribieras, S., Tomasetto, C., and Rio, M.C. (1998). The pS2/TFF1 trefoil factor, from basic research to clinical applications. *Biochim. Biophys. Acta* 1378, F61–F77. [https://doi.org/10.1016/s0304-419x\(98\)00016-x](https://doi.org/10.1016/s0304-419x(98)00016-x).
- Risom, T., Glass, D.R., Averbukh, I., Liu, C.C., Baranski, A., Kagel, A., McCaffrey, E.F., Greenwald, N.F., Rivero-Gutiérrez, B., Strand, S.H., et al. (2022). Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* 185, 299–310.e18. <https://doi.org/10.1016/j.cell.2021.12.023>.
- Rosenbluth, J.M., Schackmann, R.C.J., Gray, G.K., Selfors, L.M., Li, C.M.-C., Boedicker, M., Kuiken, H.J., Richardson, A., Brock, J., Garber, J., et al. (2020). Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. *Nat. Commun.* 11, 1711. <https://doi.org/10.1038/s41467-020-15548-7>.
- Russo, J., Rivera, R., and Russo, I.H. (1992). Influence of age and parity on the development of the human breast. *Breast Cancer Res. Treat.* 23, 211–218. <https://doi.org/10.1007/BF01833517>.
- Santucci-Pereira, J., Zeleniuch-Jacquotte, A., Afanasyeva, Y., Zhong, H., Slifker, M., Peri, S., Ross, E.A., de Cicco, R.L., Zhai, Y., Nguyen, T., et al. (2019). Genomic signature of parity in the breast of premenopausal women. *Breast Cancer Res.* 21, 1–19. <https://doi.org/10.1186/s13058-019-1128-x>.
- Schedin, P., O'Brien, J., Rudolph, M., Stein, T., and Borges, V. (2007). Microenvironment of the involuting mammary gland mediates mammary

cancer progression. *J. Mammary Gland Biol. Neoplasia* 12, 71–82. <https://doi.org/10.1007/s10911-007-9039-3>.

Sirka, O.K., Shamir, E.R., and Ewald, A.J. (2018). Myoepithelial cells are a dynamic barrier to epithelial dissemination. *J. Cell Biol.* 217, 3368–3381. <https://doi.org/10.1083/jcb.201802144>.

Söderqvist, G., Isaksson, E., Schoultz, von, B., Carlström, K., Tani, E., and Skoog, L. (1997). Proliferation of breast epithelial cells in healthy women during the menstrual cycle. *Am. J. Obstet. Gynecol.* 176, 123–128. [https://doi.org/10.1016/s0002-9378\(97\)80024-5](https://doi.org/10.1016/s0002-9378(97)80024-5).

Stein, T., Morris, J.S., Davies, C.R., Weber-Hall, S.J., Duffy, M.-A., Heath, V.J., Bell, A.K., Ferrier, R.K., Sandilands, G.P., and Gusterson, B.A. (2004). Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3. *Breast Cancer Res.* 6, R75–R91. <https://doi.org/10.1186/bcr753>.

Stein, T., Salomonis, N., Nuyten, D.S.A., van de Vijver, M.J., and Gusterson, B.A. (2009). A mouse mammary gland involution mRNA signature identifies biological pathways potentially associated with breast cancer metastasis. *J. Mammary Gland Biol. Neoplasia* 14, 99–116. <https://doi.org/10.1007/s10911-009-9120-1>.

Sternlicht, M.D., Kedeshian, P., Shao, Z.M., Safarians, S., and Barsky, S.H. (1997). The human myoepithelial cell is a natural tumor suppressor. *Clin. Cancer Res.* 3, 1949–1958.

Stewart, T.A., Hughes, K., Stevenson, A.J., Marino, N., Ju, A.L., Morehead, M., and Davis, F.M. (2021). Mammary mechanobiology - investigating roles for mechanically activated ion channels in lactation and involution. *J. Cell Sci.* 134. <https://doi.org/10.1242/jcs.248849>.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>.

Tanos, T., Siflomos, G., Echeverria, P.C., Ayyanan, A., Gutierrez, M., Delaloye, J.F., Raffoul, W., Fiche, M., Dougall, W., Schneider, P., et al. (2013). Progesterone/RANKL is a major regulatory axis in the human breast. *Sci. Transl. Med.* 5, 182ra55. <https://doi.org/10.1126/scitranslmed.3005654>.

Traag, V.A., Van Dooren, P., and Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 84, 016114. <https://doi.org/10.1103/PhysRevE.84.016114>.

Vachon, C.M., Kuni, C.C., Anderson, K., Anderson, V.E., and Sellers, T.A. (2000). Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States). *Cancer Causes Control* 11, 653–662. <https://doi.org/10.1023/a:1008926607428>.

Vogel, P.M., Georgiade, N.G., Fetter, B.F., Vogel, F.S., and McCarty, K.S. (1981). The correlation of histologic changes in the human breast with the menstrual cycle. *Am. J. Pathol.* 104, 23–34.

Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873–1887.e17. <https://doi.org/10.1016/j.cell.2019.05.006>.

Worzfeld, T., Swiercz, J.M., Looso, M., Straub, B.K., Sivaraj, K.K., and Offermanns, S. (2012). ErbB-2 signals through Plexin-B1 to promote breast cancer metastasis. *J. Clin. Invest.* 122, 1296–1305. <https://doi.org/10.1172/JCI60568>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
FITC-EpCAM	Stem Cell Technologies	60136FI
APC-CD49f	BioLegend	313616; RRID: AB_1575047
Biotin-CD2	BD Biosciences	555325; RRID: AB_395732
Biotin-CD3	BD Biosciences	555338; RRID: AB_395744
Biotin-CD16	BD Biosciences	555405; RRID: AB_395805
Biotin-CD64	BD Biosciences	555526; RRID: AB_395912
Biotin-CD31	Invitrogen	MHCD31154; RRID: AB_2539742
Biotin-CD45	BioLegend	304004; RRID: AB_314392
p63	Cell Signaling Technology	13109; RRID: AB_2637091
KRT7	Abcam	AB68459; RRID: AB_1139824
KRT23	Abcam	AB156569
ER	Fisher Scientific	RM9101S; RRID: AB_149901
PR	Cell Signaling Technology	8757; RRID: AB_2797144
TCF7	Cell Signaling Technology	2203; RRID: AB_2199302
P4HA1	Thermo Fisher	PA5-55353; RRID: AB_2645147
LRRC26	Thermo Fisher	PA5-63285; RRID: AB_2643502
Biological samples		
Human breast specimens from reduction mammoplasty surgeries	CHTN and Kaiser Permanente Northern California	Table S1
Human breast specimens from core biopsies	Komen Tissue Bank	Table S1
Chemicals, peptides, and recombinant proteins		
BV785-Streptavidin	BioLegend	405249
Collagenase Type 3	Worthington	CLS-3
Collagenase Type 2	Worthington	CLS-2
Hyaluronidase	Sigma Aldrich	H3506
RPMI with HEPES	Corning	10-041-CV
Amphotericin B	Lonza	17-836E
Gentamicin	Lonza	17-518
Dispase	Stem Cell Technologies	07913
DNase I	Stem Cell Technologies	07900
MEGM	Lonza	CC-3150
MEBM	Lonza	CC-3151
Lab Vision Ultra-V Block	Thermo Fisher	TA-125-UB
UltraVision LP Detection System	Thermo Fisher	TL-060-HL
Vectashield HardSet Mounting Media with DAPI	Vector Labs	H-1400
FITC-TSA	Akoya Biosciences	NEL701A001KT
Cy3-TSA	Akoya Biosciences	NEL744001KT
Cy5-TSA	Akoya Biosciences	NEL745E001KT
Probe Hs-ESR1	ACD Bio	310301
Critical commercial assays		
Chromium Single Cell 3' Library & Gel Bead Kit v2	10X Genomics	PN-120237
Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3	10X Genomics	PN-1000075
Chromium Single Cell A Chip Kit	10X Genomics	PN-120236

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chromium Single Cell B Chip Kit	10X Genomics	PN-1000153
Chromium i7 Multiplex Kit	10X Genomics	PN-120262
MULTI-seq Lipid-Modified Oligos	Millipore Sigma	LMO001-100RXN
Bioanalyzer High Sensitivity DNA Kit	Agilent	5067-4626
Qubit dsDNA HS Assay Kit	Thermo Fisher	Q32851
RNAscope Multiplex Fluorescent Reagent Kit V2	ACD Bio	323100

Deposited data

Raw data and processed scRNA-seq UMI counts and barcode matrices of reduction mammoplasty breast specimens	This study	GEO: GSE198732
Raw data and processed scRNA-seq UMI counts and barcode matrices of Komen Tissue Bank breast specimens	This study	GEO: GSE198732

Software and algorithms

CellRanger v3.0.2	10x Genomics	GitHub: https://github.com/10XGenomics/cellranger
MULTI-seq	McGinnis et al., 2019b	GitHub: https://github.com/chris-mcginnis-ucsf/MULTI-seq
SoupOrCell	Heaton et al., 2020	GitHub: https://github.com/wheaton5/souporcell
Seurat v3.1.5	Stuart et al., 2019; Hafemeister and Satija, 2019	GitHub: https://github.com/satijalab/seurat
DoubletFinder	McGinnis et al., 2019a	GitHub: https://github.com/chris-mcginnis-ucsf/DoubletFinder
LIGER	Gao et al., 2021; Welch et al., 2019	GitHub: https://github.com/welch-lab/liger
DECIPHER-seq computational workflow	This study	GitHub: https://github.com/lmurrow/DECIPHER-seq and https://doi.org/10.5281/zenodo.6596414
ape	Desper and Gascuel, 2002	GitHub: https://github.com/emmanuelparadis/ape
Leidenalg	Traag et al., 2011	GitHub: https://github.com/vtraag/leidenalg
wTO	Gysi et al., 2018	GitHub: https://github.com/cran/wTO
fgsea	Korotkevich et al., 2019	GitHub: https://github.com/ctlab/fgsea
DESeq2	Love et al., 2014	GitHub: https://github.com/mikelove/DESeq2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zev Gartner (zev.gartner@ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Single-cell RNA-seq data (raw FASTQ files, processed gene expression and barcode count matrices, and de-identified patient metadata) have been deposited at the Gene Expression Omnibus (GEO: GSE198732) and are publicly available as of the date of publication. Accession numbers are listed in the [key resources table](#). All original code has been deposited at Zenodo and Github and is publicly available as of the date of publication. DOIs are listed in the [key resources table](#). Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human tissue samples

Reduction mammoplasty tissue samples were obtained from the Cooperative Human Tissue Network (CHTN, Vanderbilt University Medical Center, Nashville, TN) and Kaiser Permanente Northern California (KPNC, Oakland, CA). Core biopsy samples were provided

by the Susan G. Komen Tissue Bank (KTB). Tissues were obtained as de-identified samples and all subjects provided written informed consent. When possible, medical reports or other patient data were obtained with personally identifiable information redacted. Use of breast tissue specimens to conduct the studies described above were approved by the UCSF Committee on Human Research under Institutional Review Board protocols 16-18865 and 10-01532. Donor information for all tissue specimens and their use in scRNA-seq, flow cytometry, and immunostaining experiments is detailed in [Table S1](#).

METHOD DETAILS

Tissue processing

A portion of each sample was fixed in formalin and paraffin-embedded using standard procedures. The remainder was dissociated mechanically and enzymatically to obtain epithelial-enriched tissue fragments. Tissue was minced, followed by enzymatic dissociation with 200 U/mL collagenase type III (Worthington CLS-3, samples RM108 - RM203) or collagenase type II (Worthington CLS-2, samples RM216 - RM314) and 100 U/mL hyaluronidase (Sigma H3506) in RPMI 1640 with HEPES (Corning 10-041-CV) plus 10% (v/v) dialyzed FBS, penicillin, streptomycin, amphotericin B (Lonza 17-836E), and gentamicin (Lonza 17-518) at 37 °C for 16 h. For KTB samples, the resulting cell suspension containing single cells and stroma was frozen and maintained at -180 °C until use. For reduction mamoplasty samples, the cell suspension was centrifuged at 400 x g for 10 min and resuspended in RPMI 1640 plus 10% FBS. Digested tissue fragments enriched for epithelial cells and closely-associated stroma were collected after serial filtration through 150 μ m and 40 μ m nylon mesh strainers. Following centrifugation, tissue fragments and filtrate were frozen and maintained at -180 °C until use.

Dissociation to single cells

The day of sorting, epithelial-enriched tissue fragments from the 150 μ m fraction, or total banked material for the KTB samples, were thawed and digested to single cells by trituration in 0.05% trypsin for 2 min, followed by trituration in 5 U/mL dispase (Stem Cell Technologies 07913) plus 1 mg/mL DNase I (Stem Cell Technologies 07900) for 2 min. Single-cell suspensions were resuspended in HBSS supplemented with 2% FBS, filtered through a 40 μ m cell strainer, and pelleted at 400 x g for 5 min. The pellets were resuspended in 10 mL of complete mammary epithelial growth medium with 2% v/v FBS without GA-1000 (MEGM; Lonza CC-3150). Cells were incubated at 37 °C for 1 h, rotating on a hula mixer, to regenerate surface antigens.

MULTI-seq sample barcoding (Batches 3, 4, and KTB)

Single-cell suspensions were pelleted at 400 x g for 5 min and washed once with 10 mL mammary epithelial basal medium (MEBM; Lonza CC-3151). For each sample, one million cells were aliquoted, washed a second time with 200 μ L MEBM, and resuspended in 90 μ L of a 200 nM solution containing equimolar amounts of anchor lipid-modified oligonucleotides (LMOs) and sample barcode oligonucleotides in phosphate buffered saline (PBS). Following a 5-minute incubation on ice with anchor-LMO/barcode, 10 μ L of 2 μ M co-anchor LMO in PBS was added to each sample (for a final concentration of 200 nM), and wells were mixed by gentle pipetting and incubated for an additional 5 min on ice. Following incubation, cells were washed twice in 200 μ L PBS with 1% BSA and pooled together into a single 15 mL conical tube containing 10 mL PBS/1% BSA. All subsequent steps were performed on ice.

Sorting for scRNA-seq

Cells were pelleted at 400 x g for 5 min and resuspended in PBS/1% BSA at a concentration of 1 million cells per 100 μ L, and incubated with primary antibodies. Cells were stained with Alexa 488-conjugated anti-CD49f to isolate basal/myoepithelial cells, PE-conjugated anti-EpCAM to isolate luminal epithelial cells, and biotinylated antibodies for lineage markers CD2, CD3, CD16, CD64, CD31, and CD45 to remove hematopoietic (CD16/CD64-positive), endothelial (CD31-positive), and leukocytic (CD2/CD3/CD45-positive) lineages by negative selection (Lin⁻). Sequential incubation with primary antibodies was performed for 30 min on ice in PBS/1% BSA, and cells were washed with cold PBS/1% BSA. Biotinylated primary antibodies were detected with a streptavidin-Brilliant Violet 785 conjugate. After incubation, cells were washed once and resuspended in PBS/1% BSA plus 1 μ g/mL DAPI for live/dead discrimination. Cell sorting was performed on a FACSAria II cell sorter. Live/singlet (DAPI⁻), luminal (DAPI⁻/Lin⁻/CD49f⁺/EpCAM⁺), basal/myoepithelial (DAPI⁻/Lin⁻/CD49f⁺/EpCAM⁻), or total epithelial (pooled luminal and basal/myoepithelial) cells were collected for each sample as specified in [Table S2](#) and resuspended in PBS/1% BSA at a concentration of 1000 cells/ μ L. For Batch 4, an aliquot of MULTI-seq barcoded cells were separately stained with biotinylated-CD45/streptavidin-Brilliant Violet 785 to enrich for immune cells, and sorted CD45⁺ cells were pooled with the Live/singlet fraction as specified in [Table S2](#).

Antibodies and dilutions used (μ L/million cells) were as follows: FITC-EpCAM (1.5 μ L, Stem Cell Technologies 60136FI, clone VU1D9), APC-CD49f (4 μ L, BioLegend 313616, clone GoH3), Biotin-CD2 (8 μ L; BD 555325, clone RPA-2.10), Biotin-CD3 (8 μ L; BD 55338, clone HIT3a), Biotin-CD16 (8 μ L; BD 555405, clone 3G8), Biotin-CD64 (8 μ L; BD 555526, clone 10.1), Biotin-CD31 (4 μ L; Invitrogen MHCD31154, clone MBC78.2), Biotin-CD45 (1 μ L; BioLegend 304004, clone HI30), BV785-Streptavidin (1 μ L; BioLegend 405249).

scRNA-seq library preparation

cDNA libraries were prepared using the 10X Genomics Single Cell V2 (CG00052 Single Cell 3' Reagent Kit v2: User Guide Rev B) or Single Cell V3 (CG000183 Single Cell 3' Reagent Kit v3: User Guide Rev B) standard workflows as specified in [Table S2](#). Library

concentrations were quantified using high sensitivity DNA Bioanalyzer chips (Agilent, 5067-4626) and Qubit dsDNA HS Assay Kit (Thermo Fisher Q32851). Individual libraries were sequenced on a lane of a HiSeq4500 or NovaSeq, as specified in [Table S2](#), for an average of ~150,000 reads/cell.

Expression library pre-processing

Cell Ranger (10x Genomics) was used to align sequences, filter data and count unique molecular identifiers (UMIs). Data were mapped to the human reference genome GRCh37 (hg19). The resulting sequencing statistics are summarized in [Table S2](#). For samples run across multiple 10X lanes, the cellranger aggr pipeline (10X Genomics) was used to normalize read depth across droplet microfluidic lanes (see “sort gate” information in [Table S2](#)).

Cell calling

For V2 experiments, cell-associated barcodes were defined using Cell Ranger. For V3/MULTI-seq experiments, cells were defined as barcodes associated with ≥ 600 total RNA UMIs and $\leq 20\%$ of reads mapping to mitochondrial genes. We manually selected 600 RNA UMIs and 20% mitochondrial genes to exclude low-quality cell barcodes.

MULTI-seq barcode library pre-processing

Raw barcode FASTQs were converted to barcode UMI count matrices as described previously ([McGinnis et al., 2019b](#)). Briefly, FASTQs were parsed to discard reads where: 1) the first 16 bases of read 1 did not match a list of cell barcodes generated as described above, and 2) the first 8 bases of read 2 did not align with any reference barcode with less than 1 mismatch. Duplicated UMIs, defined as reads with the same cell barcode where bases 17-28 (V3 chemistry) of read 2 exactly matched, were removed to produce a final barcode UMI count matrix.

Sample demultiplexing

Barcode UMI count matrices were used to classify cells using the MULTI-seq classification suite ([McGinnis et al., 2019b](#)). In Batch 3, sample RM192 was poorly labeled for the lane of cells from the epithelial cell sort gate. Therefore, to reduce spurious doublet calls in this dataset, we manually set UMI counts which were < 10 for this barcode to zero. For all experiments, raw barcode reads were \log_2 -transformed and mean-centered, the top and bottom 0.1% of values for each barcode were excluded, and a probability density function (PDF) was constructed for each barcode. Next, all local maxima were computed for each PDF, and the negative and positive maxima were selected. To define a threshold between these two maxima, we iterated across 0.02-quantile increments and chose the quantile maximizing the number of singlet classifications, defined as cells surpassing the threshold for a single barcode. Multiplets were defined as cells surpassing two or more thresholds, and unlabeled cells were defined as cells surpassing zero thresholds. Unclassified cells were removed and the procedure was repeated until all remaining cells were classified.

To classify cells that were identified as unlabeled by MULTI-seq, we used the SoupOrCell pipeline ([Heaton et al., 2020](#)) to assign cells to different individuals based on single nucleotide polymorphisms (SNPs). For each dataset, we set the number of clusters (k) to the total number of samples in that experiment. To avoid local minima, SoupOrCell restarts clustering multiple times and takes the solution that minimizes the loss function. For Batch 3, we chose the number of restarts that produced less than a 1.5% misclassification rate between MULTI-seq and SoupOrCell singlet sample classifications (Live/singlet: 30 restarts/1.2% mismatch rate; Epithelial: 75 restarts/1.5% mismatch rate). SoupOrCell classification performed more poorly across parameters for Batch 4 (Live/singlet plus CD45+: 50 restarts/8.1% mismatch rate, 75 restarts/4.8% mismatch rate; Epithelial: 50 restarts/8.6% mismatch rate, 75 restarts/14.9% mismatch rate, 100 restarts/4.1% mismatch rate). Therefore, for these datasets we used sample classifications that were consistent across two restarts (Live/singlet plus CD45+: consistent calls across 50 and 75 restarts/0.4% overall mismatch rate; Epithelial: consistent calls across 50 and 100 restarts/1% overall mismatch rate) to identify high-confidence singlets.

Dataset integration and cell type identification

Cell type identification was performed using the Seurat package (version 3.1.5) in R ([Stuart et al., 2019](#)). To identify and remove doublets formed from cells from the same sample that would not be identified by MULTI-seq or SoupOrCell, we filtered each lane to remove cells with greater than 20% of reads mapping to mitochondrial genes and ran DoubletFinder (version 2.0) on each data subset ([McGinnis et al., 2019a](#)), using parameters identified by the ‘paramSweep_v3’ function. Aggregated data for singlet cells for each batch was filtered to remove cells that had fewer than 200 genes and genes that appeared in fewer than 3 cells. Cells with a Z score of 4 or greater for the total number of genes expressed were presumed to be doublets and removed from analysis. The remaining cells were log transformed and scaled to a total of $1e4$ molecules per cell, and the top 2000 most variable genes based on variance stabilizing transformation were identified for each batch ([Hafemeister and Satija, 2019](#)). Data from all four batches were integrated using the standard workflow and default parameters from Seurat v3 ([Stuart et al., 2019](#)). This data integration workflow identifies pairwise correspondences between cells across datasets and uses these anchors to transform datasets into a shared expression space. Following dataset integration, the resulting batch-corrected expression matrix was scaled, and principal component (PC) analysis was performed using the identified integration genes. The top 28 statistically significant PCs as determined by visual inspection of elbow plots were used as an input for UMAP visualization and k-nearest neighbor (KNN) modularity optimization-based clustering using Seurat’s ‘FindNeighbors’ and ‘FindClusters’ functions.

PC analysis of individual cell types

To perform principal component analysis on individual cell types, we subset out each cluster from the integrated dataset and repeated the standard workflow from Seurat v3 to identify integration genes specific to this cell type. The resulting batch-corrected expression matrices were scaled, and PC analysis was performed using the identified integration genes.

Activity program identification in each cell type

To identify gene expression signatures, or “activity programs”, within individual cell types, we subset raw counts data from each of the five most abundant cell type clusters (HR+ luminal cells, secretory luminal cells, basal/myoepithelial cells, fibroblasts, and endothelial cells) and performed matrix factorization. We chose to perform matrix factorization independently on each cell type rather than on the combined dataset, as preliminary analyses demonstrated that the number of gene programs identified for each cell type was highly dependent on the relative sizes of each cluster in the combined dataset. To correct for batch differences between samples run on different days, we used the LIGER package in R to perform integrative NMF (iNMF) (Gao et al., 2021; Welch et al., 2019), and performed subsequent gene set enrichment analyses on shared, rather than batch-specific, gene loadings for each activity program. Activity program expression in cells from the same sample run across different batches was more similar than program expression in cells from different samples processed in the same batch, demonstrating that this approach successfully corrected for batch differences while retaining sample-to-sample transcriptional variability (Figures S4A and S4B). To avoid identification of gene signatures dominated by highly-expressed transcripts, we normalized the raw counts matrix for each cell based on its total expression, multiplied by a scale factor of $1e4$, and log-transformed and scaled the result without centering. The resulting datasets (one for each cell type) were decomposed using the ‘online_iNMF’ function from LIGER (Gao et al., 2021). Online iNMF uses an online learning algorithm to iteratively cycle through the data in small mini-batches, greatly increasing convergence times for large datasets. We performed 10 complete passes (‘max.epochs’ parameter) through each dataset, and chose the mini-batch size (‘miniBatch_size’) by rounding down to the nearest 500 from the smallest batch size in that cell type (HR+ luminal cells: 1000, Secretory luminal cells: 2000, Basal cells: 500, Fibroblasts: 500, Endothelial cells: 500).

Since solutions to NMF are non-unique, we adapted a consensus matrix factorization approach from (Kotliar et al., 2019) to identify activity programs that were consistent across multiple replicates. For each cell type, we ran 20 replicates of iNMF on the same normalized dataset with the same choice of rank K , starting from different random seeds. We row normalized the resulting 20 shared gene loading matrices (W , each of dimension $K_{programs} \times N_{genes}$) to have an L2 norm of one. Following normalization, we combined the shared gene loading matrices from each matrix into a $20K_{programs} \times N_{genes}$ dimensional matrix, where each row represents the gene loading from one activity program in one replicate. Next, we filtered out programs with a high mean Euclidean distance from their 6 nearest neighbors (30% of replicates), using the third quartile plus 1.5 times the interquartile range ($q_{0.75} + 1.5 \cdot IQR$) as an outlier threshold. After filtering outlier programs, we grouped the rows of the resulting matrix using k-means clustering, with the number of clusters set to the chosen iNMF rank K . Next, we collapsed each group of shared gene loadings to a single consensus vector by taking the median value for each gene across activity programs in that cluster, to produce a final $K_{programs} \times N_{cells}$ consensus program matrix, W . We performed the same row normalization on the batch-specific gene loading matrices, filtered programs identified as outliers in the shared gene loading matrix, and collapsed groups of batch gene loadings into a consensus vector by taking the median value for each gene across programs in that cluster to produce consensus batch matrices V_{batch} , each of dimension $K_{programs} \times N_{genes}$. Finally, we solve for the consensus cell expression score matrix H ($X_{cells} \times K_{programs}$), by using non-negative least squares initialized with the consensus shared (W) and batch-specific (V_{batch}) gene loading matrices.

A key parameter in matrix factorization is the choice of rank K . This parameter determines the granularity of identified activity programs. Three commonly used heuristics for guiding the optimum choice of K are: 1) minimizing the Frobenius reconstruction error of the final solution (Kotliar et al., 2019), 2) maximizing the median Kullback-Leibler (KL) divergence of activity program loadings across cells relative to a uniform distribution (Welch et al., 2019), and 3) estimating the “dimensionality” of the dataset via elbow plot of the proportion of variance explained across principal components (Kotliar et al., 2019). We propose a metric for choosing an optimum K , based on the goal of identifying the greatest number of activity programs that are robust (i.e. consistent across multiple choices of K) and unique (i.e. distinct from other programs at a particular choice of K). First, we perform consensus iNMF as described above over a range of ranks, with the sweep range guided by the heuristics described above. Here, we chose a range of 2 to 40 for all cell types. Next, we use the ‘fastme.bal’ function in the ‘ape’ R package to build a balanced minimum evolution phylogenetic tree based on the correlation matrix of the gene loadings for activity programs across all ranks (Desper and Gascuel, 2002). For each cell type, we partitioned the resulting phylogenetic tree into clusters using an empirical distance threshold to define distinct groups of activity programs (Prosperi et al., 2011; Figures S5A and S5B). To identify partitions, we first artificially rooted each tree by taking the median of the activity programs at $K = 2$. Next, we identified clusters by performing a depth-first search starting from this artificial root, stopping at sub-trees where the median value of the pairwise patristic distance between all programs in that sub-tree was below an empirically determined threshold of 0.3 (see Figure S5B). To filter out “outlier” activity programs that are expressed in only rare contaminating cells (e.g. a “fibroblast-like” gene signature in HR+ luminal cells), we calculated the maximum expression score for each activity program divided by the mean expression score for the next 50 highest-scoring cells, and removed programs where this ratio was greater than 5 (Figure S5C). We also removed subtrees with fewer than 5 total activity programs. Finally, we plotted the number of subtrees identified at each K (excluding outlier programs), weighted by the total number of programs in each subtree. We choose the optimum K (K_{opt}) as the saturation point in this curve, representing the point at which increasing the granularity of matrix factorization does not identify activity programs that comprise major new subtrees (Figure S5D).

Network clustering of correlated activity programs

To identify sets of activity programs that co-varied across samples, we first decomposed each cell type into a set of distinct gene expression signatures, or “activity programs”, using consensus iNMF with K_{opt} chosen for each cell type as described above. We then quantified the average expression of each gene program in each sample and constructed a weighted network of coordinated gene expression programs based on the pair-wise Pearson correlations between gene programs. To account for correlations driven by outlier samples, we used bias-corrected and accelerated bootstrap resampling to estimate confidence intervals associated with each correlation coefficient. The resulting Pearson correlation matrix was transformed into a weighted adjacency matrix by setting all Pearson correlation coefficients with p-values greater than 0.05 (based on the null hypothesis $r = 0$) to zero (Figures S6A and S6B). We identified modules of highly correlated gene expression programs using a Constant Potts Model for community detection in signed graphs in the ‘leidenalg’ package in python (Figure S6B; Traag et al., 2011). We ran this algorithm at a range of resolutions from 0.001 to 0.4 and chose the resolution that maximized overall modularity. To filter out isolated links and modules, we calculated the signed weighted topological overlap (wTO) between activity programs in each module (Gysi et al., 2018) and filtered nodes with low wTO and modules containing fewer than four nodes (Figure S6C). In contrast to Pearson correlation values which consider each pair of nodes in isolation, wTO is based on the similarity of two activity programs’ correlation values with all other programs in the network. We calculated the mean wTO between each node and all other nodes in the same module, and compared this to the value calculated for nodes in randomly selected modules of equal size. We determined p-values for each node’s mean wTO by determining the fraction of permutation trials where the mean wTO of nodes from “random” modules was greater than the mean wTO of nodes from tested modules, and removed nodes where $p > 0.01$. Community detection results remained unchanged after this filtering step (Figures S6B and S6D). For visualization, we use positive edges to create a force-directed layout. Consistent with our goal of choosing the rank K that captured the greatest number of unique activity programs (see above), the overall organization of modules into cell-cell interaction networks remained highly robust to the choice of rank at values of $K \geq K_{opt}$, whereas the network structure at $K \leq K_{opt}$ had much sparser connections between modules (Figure S6E).

Fluorescent immunohistochemistry

For immunofluorescent staining, formalin-fixed paraffin-embedded tissue sections were deparaffinized and rehydrated using standard methods. Endogenous peroxidases were blocked using 3% hydrogen peroxide in PBS, and antigen retrieval was performed in 0.1 M citrate buffer pH 6.0. Sections were blocked for 5 min at room temperature using Lab Vision Ultra-V block (Thermo TA-125-UB) and rinsed with TNT wash buffer (1X Tris-buffered saline with 5 mM Tris-HCl and 0.5% TWEEN-20). Primary antibody incubations were performed for 1 hour at room temperature or overnight at 4°C. Sections were washed three times for 5 min each with TNT wash buffer, incubated with Lab Vision UltraVision LP Detection System HRP Polymer (Thermo Fisher TL-060-HL) for 15 min at room temperature, washed, and incubated with one of three colors of tyramide signal amplification (TSA) reagent at a 1:50 dilution. After TSA, antibody complexes were removed by boiling in citrate buffer, followed by blocking and incubation with additional primary antibodies as above. Finally, sections were rinsed with deionized water and mounted using Vectashield HardSet Mounting Media with DAPI (Vector H-1400). Immunofluorescence was analyzed by spinning disk confocal microscopy using a Zeiss Cell Observer Z1 equipped with a Yokagawa spinning disk and running Zeiss Zen Software.

Antibodies, TSA reagents, and dilutions used are as follows: p63 (1:2000; CST 13109, clone D2K8X), KRT7 (1:4000; Abcam AB68459, clone EPR1619Y), KRT23 (1:2000; Abcam AB156569, clone EPR10943), ER (1:4000; Thermo Scientific RMM-9101-S, clone SP1), PR (1:3000; CST 8757, clone D8Q2J), TCF7 (1:2000; CST 2203, clone C63D9), P4HA1 (1:9000; Thermo PA5-55353), LRRc26 (1:2000; Thermo PA5-63285), FITC-TSA (2 min; Akoya Biosciences NEL701A001KT), Cy3-TSA (3 min; Akoya Biosciences NEL744001KT), Cy5-TSA (7 min; Akoya Biosciences NEL745E001KT).

Morphometric analysis and geometric modeling

Formalin-fixed paraffin-embedded tissue sections were immunostained for the pan-luminal marker KRT7, counterstained with DAPI and imaged as described above. Images containing lobular tissue were acquired randomly, and the area and perimeter of the KRT7-positive luminal layer of each acinus was analyzed in ImageJ. To reduce noise and remove small gaps in KRT7 fluorescence, we applied a closing filter from the MorphoLibJ plugin with a 2-pixel (1.33 μm) radius disk (Legland et al., 2016). The resulting image was smoothed by applying a Gaussian filter with sigma 5 pixels (3.33 μm), and binarized using the default thresholding algorithm in ImageJ. Finally, individual acini with visible lumens were manually selected and the area (A), perimeter (P), and circularity of the KRT7-positive region was measured for each structure. To estimate the average diameter (d) and luminal thickness (w) of each acinus, we used area and perimeter measurements to fit a circle containing a hollow lumen to each structure. Based on these results, we implemented a geometric model in which each acinus was represented as a hollow circle with shell thickness that was linearly related to diameter (d). Since basal cells form a monolayer along the luminal surface, we represented the space available for basal cells as the outer perimeter of the luminal layer, and the space available for luminal cells as the area of the luminal layer. To estimate the linear relationship between w and d , we performed linear regression analysis using measurements from all structures.

RNA FISH analysis of ESR1 transcripts

Combined RNA FISH and immunofluorescence analysis of estrogen receptor transcript (RNAscope Probe Hs-ESR1; ACD 310301) and protein (anti-ER; Thermo RMM-9101-S, clone SP1) was performed using the RNAscope in situ hybridization kit (RNAscope Multiplex Fluorescent Reagent Kit V2, ACD 323100) according to the manufacturer’s instructions and fluorescent immunohistochemistry

protocol outlined above with the following modifications. Immunostaining for ER was performed prior to *in situ* hybridization, using the hydrogen peroxide and antigen retrieval solutions supplied with the RNAscope kit and the mildest recommended conditions. After ER immunostaining and tyramide signal amplification, *in situ* hybridization for ESR1 was performed according to the manufacturer's instructions, followed by immunostaining for KRT7 as described above. For all RNA FISH experiments, we used positive (PPIB) and negative controls (DAPB) to verify staining conditions and probe specificity.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of sample-to-sample heterogeneity

Cluster entropy: To measure how well-mixed cells from different samples were across cell type clusters, we quantified the normalized relative cluster entropy for our dataset, weighted by cluster size (Barkas et al., 2019). A cluster entropy value of 1 represents complete intermixing of samples across clusters.

Similarity scores/alignment: To measure transcriptional variation in cell state within cell types between cells from the same versus different batches and/or samples, we measured the pairwise alignment between each sample/batch (Butler et al., 2018), where batches consisted of sets of samples processed on the same day (Table S2). This "similarity score" examines the local neighborhood of each cell in a particular sample/batch, asks how many of its k nearest neighbors (in PC or iNMF space) belong to a second sample/batch, and averages this over all cells. We chose k to be 1% of the total number of cells within a cluster. The result was normalized by the expected number of cells from each sample/batch. For repeat measurements, samples run across multiple batches were highly similar. For Figure S2E, we calculated the pairwise similarity score between each sample/batch using the first 14 principal components for each cell type (See also Figure S4E depicting the standard deviation of each principal component). For Figures S4A and S4B, we calculated the pairwise similarity score between each sample/batch using all iNMF components for each cell type (at K_{opt} , see text below for optimization of K).

Testing for changes in cell type proportions

We modeled the detected number of each cell type in each sample as a random count variable using a quasi-Poisson process to allow for overdispersion, with the condition being tested (e.g. parity, BMI, obesity) as a predictor and the total number of detected epithelial or luminal cells in each sample as an offset variable (Haber et al., 2017). To account for uncertainty due to variable numbers of profiled cells in each sample, we used bootstrap resampling to estimate confidence intervals associated with detection of each cell type (Cao et al., 2019). Results from 1000 bootstrap replicates were pooled using the 'mice::pool' function in R, and the model was fit using a quasi-Poisson generalized linear model from the 'stats' R package. Tests for statistical significance were performed using a Wald test on the regression coefficient. Multiple hypothesis correction was controlled using the false discovery rate. For the Komen Tissue Bank (KTB) data set, a quasi-Poisson model was trained on the reduction mammoplasty cohort as described above, and the 'predict' function in the 'stats' R package was used to predict the proportion of HR+ luminal cells in the KTB samples based on BMI.

Identification of non-cell-type specific programs

To identify transcriptionally similar activity programs representing non-cell-type specific responses, we calculated the Pearson correlation of gene loadings between activity programs using pairwise complete observations (i.e. excluding genes that are not expressed in either cell type). We defined each node's "mean gene loading similarity" as the mean correlation between the tested node and all other nodes in the same module. To determine p -values for each node's gene loading similarity, we compared this value to that calculated for nodes in randomly selected modules of equal size. The reported p -values represent the fraction of permutation trials where the mean gene loading similarity for nodes from "random" modules was greater than the mean gene loading similarity for nodes in tested modules.

Inferring direct cell-cell interactions

To infer modules enriched for putative direct cell-cell signaling interactions, we identified links between nodes that depended on both the magnitude of activity program expression in a "sender" cell type and the proportion of that "sender" cell type in the tissue. Since the proportion of epithelial versus stromal cells in our samples was highly dependent on tissue dissociation conditions, we restricted this analysis to links between epithelial cell types as "sender" cells (HR+ luminal, secretory luminal, or basal cells) and other cell types as "receivers". We modeled activity program expression in the "receiver" cell type as a linear response to three predictors: activity program expression \mathbf{Y} in the "sender" cell type (i.e. "signaling" from that cell type), the proportion $\mathbf{P}_{\text{sender}}$ of the "sender" cell type in the epithelium, and an interaction term representing the combined effects of signaling and cell proportions (Signaling \times Proportions). For links between two epithelial cell types, we tested both directions as "sender" versus "receiver" nodes. To infer high-confidence direct cell-cell signaling interactions, we identified pairwise combinations of activity programs where a) the individual effects of \mathbf{Y} and $\mathbf{P}_{\text{sender}}$ were not significant ($p > 0.05$), b) there was a positive interaction effect between \mathbf{Y} and $\mathbf{P}_{\text{sender}}$ (Signaling \times Proportions; $p < 0.01$ and $\beta > 0$), c) the adjusted R-squared for the overall model was at least 0.5, and d) the false discovery rate-corrected p -value for the overall model was less than 0.05.

Gene set enrichment analysis

To identify marker genes statistically associated with each gene program, we used ordinary least squares regression of each gene's normalized (z -scored) expression against the activity program expression score for each program in each cell type, after filtering

genes not expressed in that cell type (Kotliar et al., 2019). This results in a vector of regression coefficients representing the strength of the relationship between a cell's expression score for a particular activity program and its scaled expression of each gene (e.g. see Figure 4F). The resulting ranked gene lists (Table S3) were analyzed by gene set enrichment analysis, using the 'fgsea' package in R (Korotkevich et al., 2019).

Enrichment of gene sets within modules

To identify gene sets enriched across activity programs in a module, we first calculated the false discovery rate (FDR) for each gene set in each node. We performed false discovery rate correction for Hallmark and GO Biological Process gene sets separately, as many of the pathways in each database are highly related. For all gene sets enriched across at least 5 activity programs in our network, we calculated the number of activity programs in each module that were significantly enriched for each gene set (FDR < 0.01), and compared this value to randomly selected modules of equal size. We determined p-values for enrichment of gene sets in each module by determining the fraction of permutation trials where the number of significantly enriched nodes from "random" modules was greater than number of significantly enriched nodes from tested modules.

Sample-to-sample variability in ER/PR signaling

To quantify variation in expression of the "ER/PR signaling" gene program in HR+ luminal cells (HR+ gene program 1), we performed the following workflow. First, we used the cell loadings across HR+ gene program 1 for each sample to compute kernel density estimations using the 'density' function in the 'stats' R package. We excluded sample RM172 from this analysis as it had fewer than 50 HR+ luminal cells; thus, the resulting kernel density estimation was highly sensitive to individual outliers. Second, we used the 'JSD' function in the 'philentropy' R package (Drost, 2018) to measure the pairwise Jensen-Shannon divergence between samples. Third, we converted this to a distance metric (Jensen-Shannon Distance, JSD) by taking the square root and performed hierarchical clustering using the 'hclust' function in the 'stats' R package, using 'ward.D2' linkage. The similarity between samples was plotted on a heatmap as (1-JSD).

Pseudo-bulk differential gene expression analysis

To identify genes differentially expressed between samples from parous and nulliparous individuals in specific cell types, we constructed pseudo-bulk datasets consisting of the summed raw read counts across all single HR+ luminal cells for each batch and sample. We restricted our analysis to samples/batches that had at least 100 HR+ luminal cells. Each dataset was then randomly down-sampled to the lowest library size, and differential expression analysis was performed using DESeq2 (version 1.18.1) to test for genes differentially expressed between samples from parous and nulliparous individuals, using batch as a covariate (Love et al., 2014). As certain samples were sequenced across more than one batch (Table S2), replicates of the same sample from different batches were combined using the 'collapseReplicates' function. False discovery rate corrected p-values were calculated using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Cell Systems, Volume 13

Supplemental information

**Mapping hormone-regulated
cell-cell interaction networks
in the human breast at single-cell resolution**

Lyndsay M. Murrow, Robert J. Weber, Joseph A. Caruso, Christopher S. McGinnis, Kiet Phong, Philippe Gascard, Gabrielle Rabadam, Alexander D. Borowsky, Tejal A. Desai, Matthew Thomson, Thea Tlsty, and Zev J. Gartner

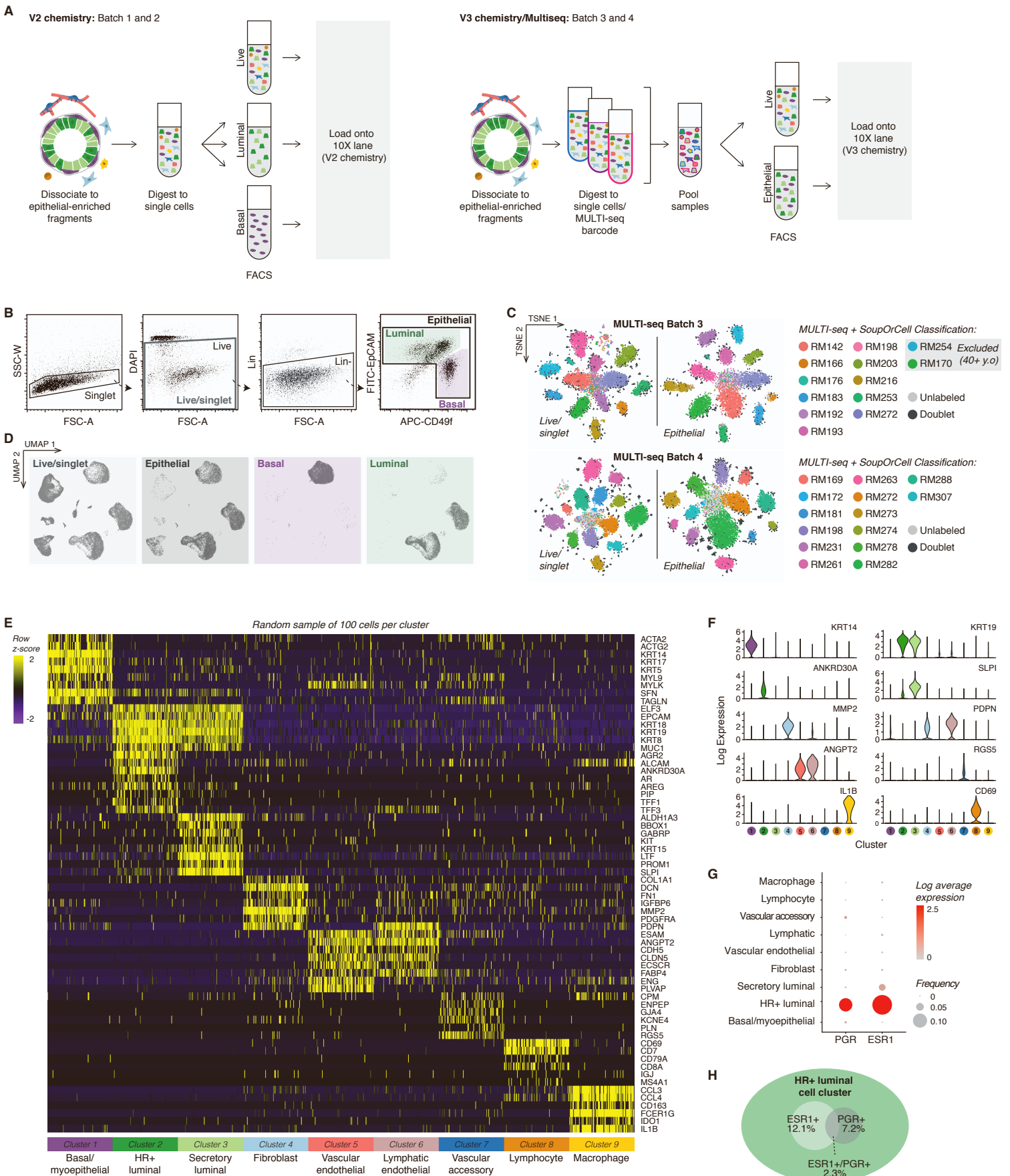


Figure S1 - Overview of scRNAseq workflow and basic clustering results

Figure S1. Overview of scRNA-seq workflow and basic clustering results; Related to Figure 1

(A) Overview of sample multiplexing and sorting strategies across batches (also see Table S2).

(B) FACS plots depicting sort gates used for sequencing.

(C) TSNE dimensionality reduction of the normalized barcode count matrices and final sample classification for MULTI-seq experiments (Batches 3 and 4, also see *STAR Methods*).

(D) UMAP dimensionality reduction of the combined data from twenty-eight reduction mammoplasty samples (GSE198732) for each sort population.

(E) Heatmap highlighting marker genes used to identify each cell type, colored by row z-score (linear scale). For visualization purposes, we randomly selected 100 cells from each cluster.

(F) Violin plot highlighting the log expression of selected marker genes in each cluster.

(G) Dot plot depicting the log normalized average and frequency of ESR1 and PGR expression across cell type clusters.

(H) Euler diagram highlighting the frequency of ESR1 and PGR expression and percent overlap in the HR+ luminal cell cluster.

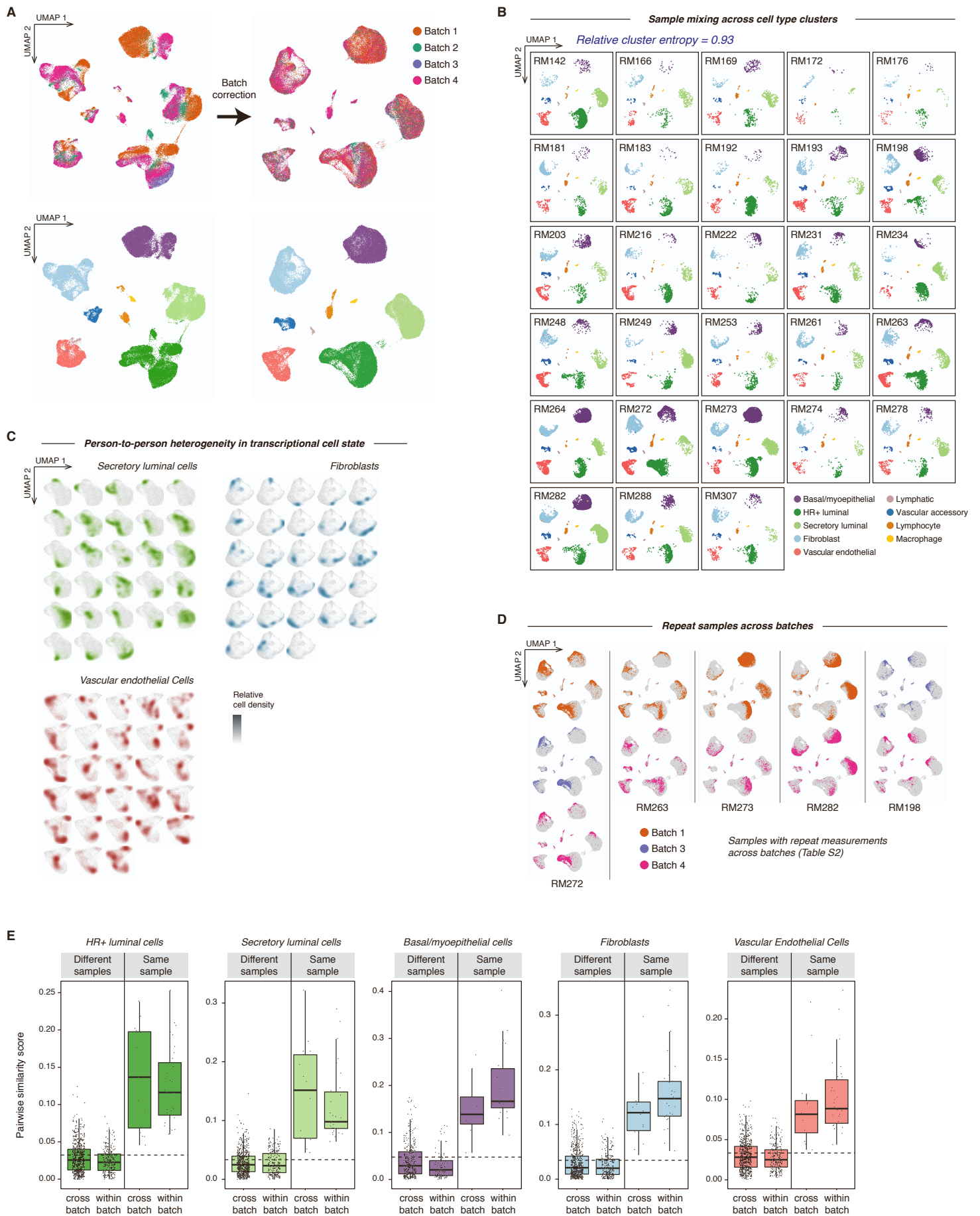


Figure S2 - Transcriptional variation between samples is independent of batch effects

Figure S2. Transcriptional variation between samples is independent of batch effect; Related to Figure 1

(A) UMAP dimensionality reduction of the combined data from twenty-eight reduction mammoplasty samples (GSE198732) for each sort population before (left) and after (right) batch correction.

(B) UMAP for each reduction mammoplasty sample highlighting cell types identified by unsupervised clustering. Cells from different individuals are well-mixed across all clusters (cluster entropy = 0.93, *STAR Methods*).

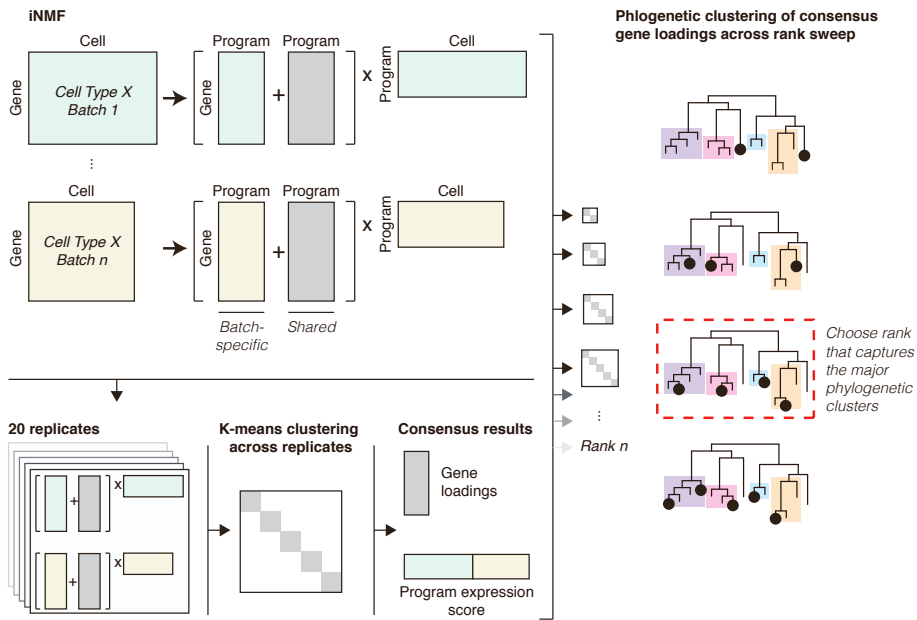
(C) Density plots (arbitrary units, linear scale) highlighting the transcriptional cell state of the indicated cell types from each sample.

(D) UMAP of reduction mammoplasty samples that were run as repeat measurements across multiple batches, highlighting cells from each batch. See Table S2 for sample and batch information.

(E) Quantification of the pairwise alignment—or “similarity score”—between cells from the same or different sample and within versus across batches for the indicated cell types. See Table S2 for sample and batch information. The dashed line represents the expected similarity score for random mixing.

A

Activity program identification in each cell type (NMF)



B

Network clustering of correlated activity programs

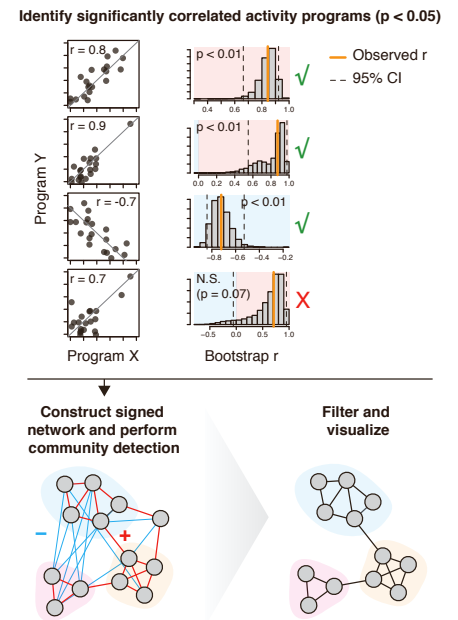


Figure S3 - Overview of DECIPHER-seq workflow

Figure S3. Overview of DECIPHER-seq workflow; Related to Figure 2

(A) To identify gene activity programs in the breast, we performed non-negative matrix factorization (NMF) on each of the major cell types in our dataset, using integrative NMF (iNMF) to account for batch differences. As NMF solutions are non-unique, we adapted a consensus matrix factorization approach to identify activity programs that were consistent across multiple replicates. To optimize rank K , we used phylogenetic clustering of consensus gene loadings across a range of K values and chose the point at which increasing granularity did not identify major new subtrees.

(B) To build a network map of cell-cell interactions, we quantified the average activity program expression for each sample and constructed a weighted network based on the pair-wise Pearson correlations r . The resulting correlation matrix was transformed into a signed weighted adjacency matrix by using bootstrap resampling to estimate confidence intervals associated with each correlation, and setting all correlations with p-values > 0.05 to zero. Finally, we identified modules of highly correlated gene expression programs using a community detection algorithm, and filtered out isolated links and modules using weighted topological overlap.

A Pairwise alignment of NMF results for samples run across multiple batches with at least 100 cells per condition

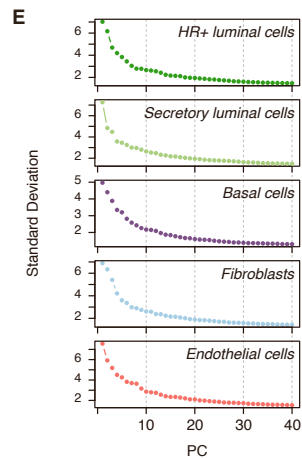
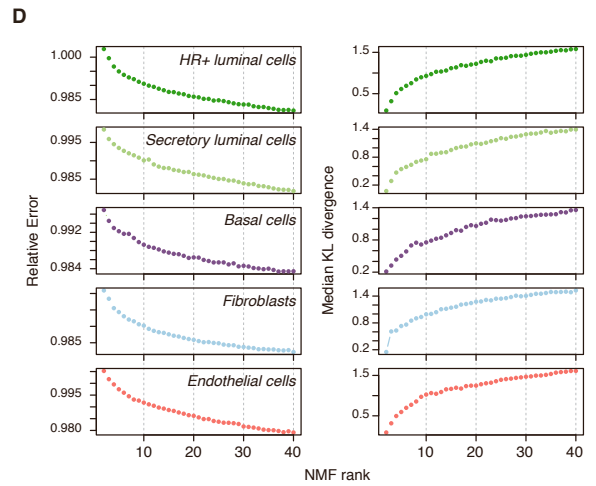
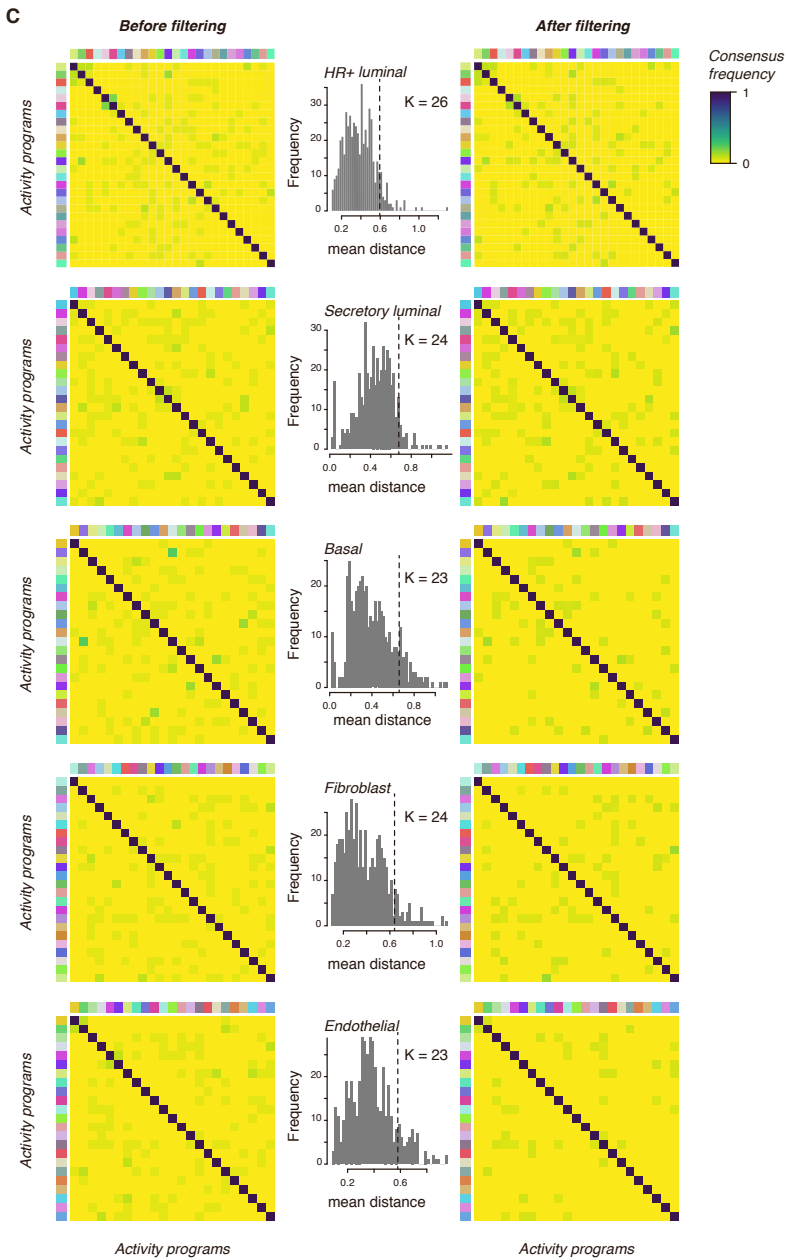
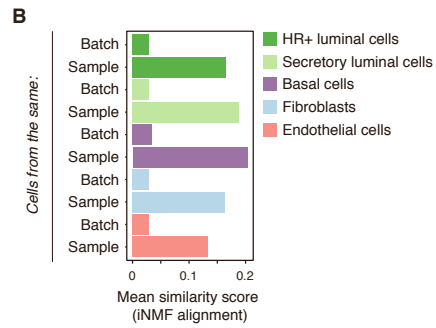
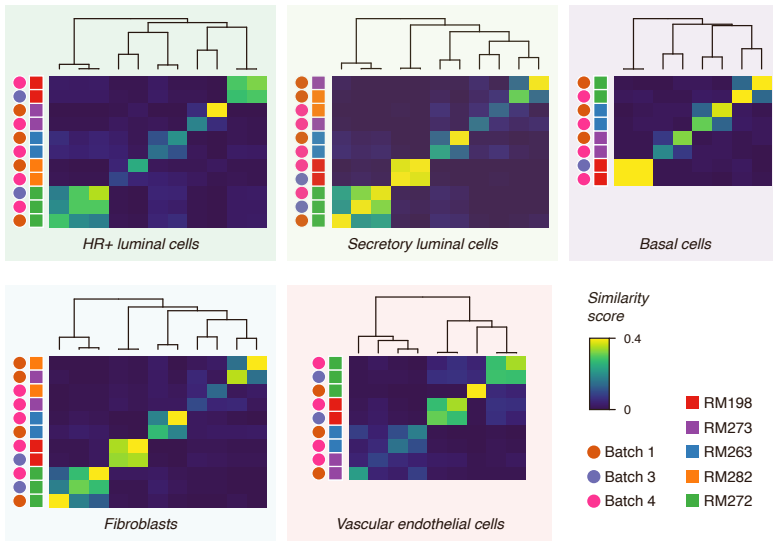


Figure S4 - NMF integration, consensus clustering of replicates, and metrics for choice of rank sweep

Figure S4. NMF integration, consensus clustering of replicates, and metrics for choice of rank sweep; Related to Figure 2

(A) Heatmaps showing the pairwise alignment (“similarity score”; fraction of nearest neighbors from the corresponding sample/batch, linear units) of NMF results for samples run across multiple batches. Hierarchical clustering (Ward D2 linkage) groups cells by sample rather than batch.

(B) Quantification of the mean pairwise alignment—or “similarity score”—of iNMF results for cells from the same sample but different batch (“Sample”) or different sample but same batch (“Batch”) for the indicated cell types. See Table S2 for sample and batch information.

(C) Consensus matrix (frequency, linear units) showing the clustered NMF components for the indicated values of K , combined across 20 replicates, before (left) and after filtering (right). The histogram shows the mean distance of each component to its 6 nearest neighbors with a dashed line showing the threshold for filtering outliers (see *STAR Methods*).

(D) Parameter selection for K sweep range. Plots depict the relative Frobenius reconstruction error and median KL divergence for consensus results at each value of K .

(E) Parameter selection for K sweep range. Plot depicting the standard deviation of the first forty principal components for each cell type.

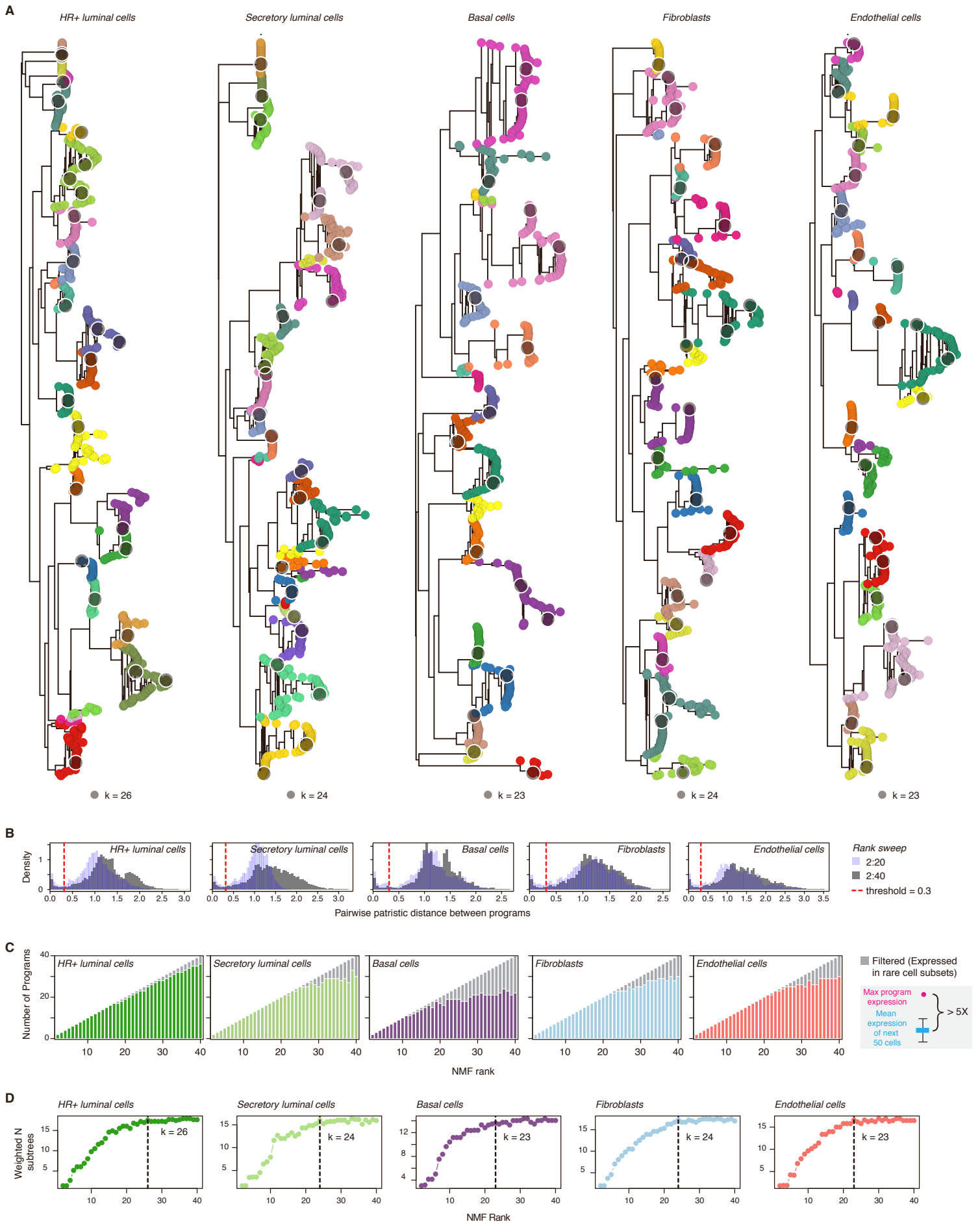


Figure S5 - Phylogenetic clustering of gene programs across rank sweep and final choice of rank

Figure S5. Phylogenetic clustering of gene programs across rank sweep and final choice of rank; Related to Figure 2

(A) Phylogenetic trees of consensus iNMF components (“activity programs”) for K values ranging from 2 to 40, based on a balanced minimum evolution algorithm. Each tree was partitioned into distinct clusters using a depth-first search to identify sub-trees with median patristic distance below a threshold of 0.3 (see below).

(B) Histogram of each phylogenetic tree’s patristic distance distribution, for phylogenetic trees constructed from K sweep values ranging from 2 to 40 (grey) or 2 to 20 (blue). The vertical red line shows the distance threshold used in this study of 0.3. We chose an absolute rather than relative (e.g. percentile) distance threshold as it was robust to different sweep ranges of K (e.g. the blue versus grey histograms).

(C) Plot depicting the number of “outlier” activity programs filtered at each value of K . Outliers were defined as activity programs where the maximum expression score of an activity program across all cells was more than 5 times greater than the mean expression score of that activity program in the next 50 highest-scoring cells.

(D) Plots depicting the number of subtrees identified at each K , weighted by the total number of programs in each subtree (see *STAR Methods*). Dashed lines depict the chosen values for K (K_{opt}). We chose K_{opt} as the saturation point representing the point at which increasing the granularity of matrix factorization did not identify activity programs that comprise major new subtrees.

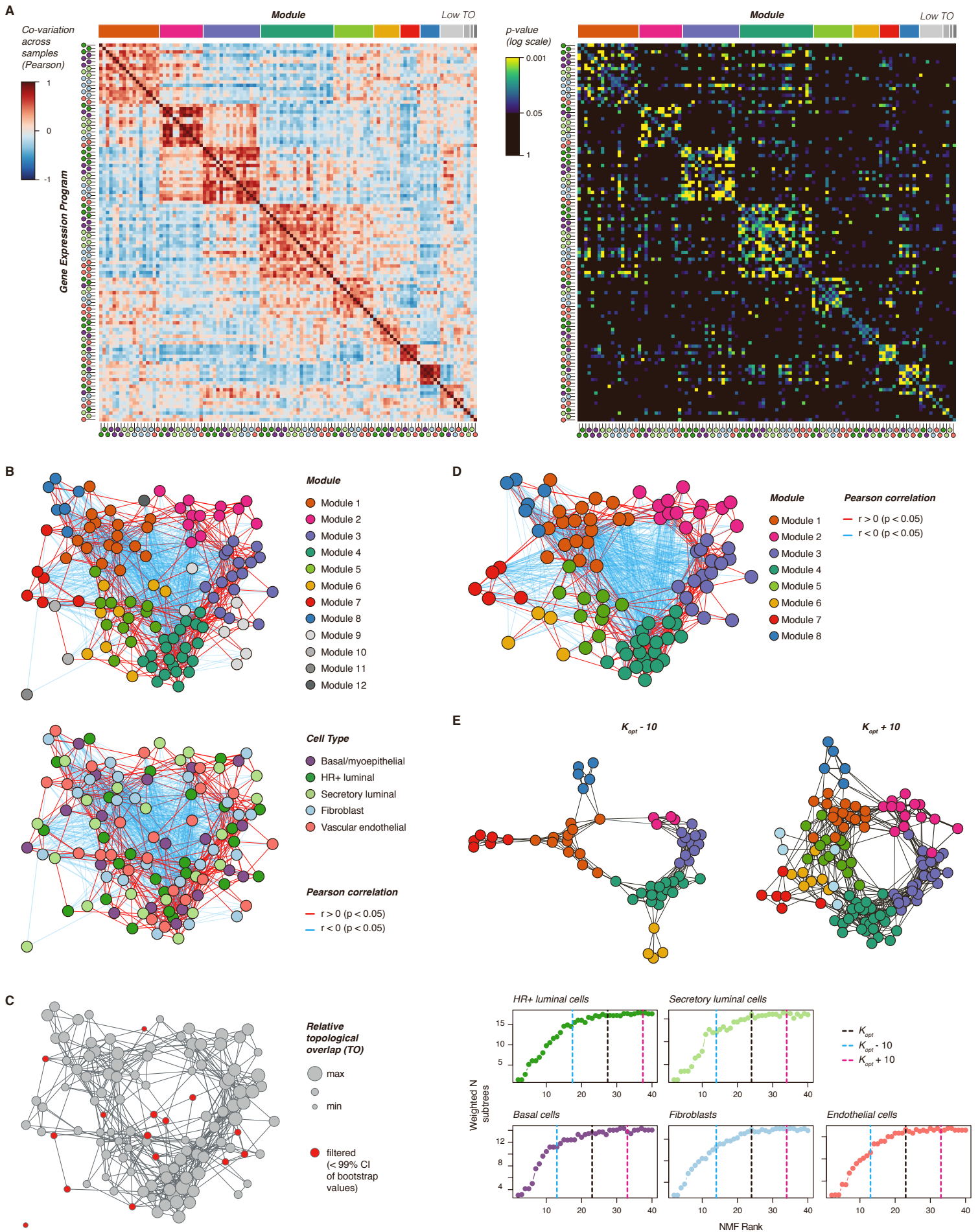


Figure S6 - Network clustering of correlated activity programs

Figure S6. Network clustering of correlated activity programs; Related to Figure 2

(A) *Left*: Heatmap depicting Pearson correlation coefficients (r) between activity programs in all identified modules, including filtered modules with low topological overlap. Modules were identified using a Constant Potts Model for community detection. *Right*: Heatmap depicting p-values for the Pearson correlation between activity programs in all identified modules (log scale). To estimate p-values, we used bootstrap resampling to determine confidence intervals for each correlation coefficient, and calculated p-values based on the null hypothesis that $r = 0$.

(B) Network graph of correlated activity programs in the human breast, including modules with low topological overlap, and depicting both positive and negative edges. Nodes represent distinct activity programs in the indicated cell types, and edges connect significantly correlated programs (red: Pearson correlation coefficient > 0 , p-value < 0.05 ; blue: Pearson correlation coefficient < 0 , p-value < 0.05). Modules were identified using a Constant Potts Model for community detection.

(C) Network graph of correlated activity programs in the human breast, with node sizes proportional to the relative weighted topological overlap (wTO) of each node with other nodes in the same module (arbitrary units, linear scale). Filtered activity programs with a wTO within the 99% confidence interval for nodes samples from “random” modules of the same size are highlighted in red.

(D) Network graph of correlated activity programs in the human breast, excluding modules with low topological overlap, and depicting both positive and negative edges. Nodes represent distinct activity programs in the indicated cell types, and edges connect significantly correlated programs (red: Pearson correlation coefficient > 0 , p-value < 0.05 ; blue: Pearson correlation coefficient < 0 , p-value < 0.05).

(E) *Top*: Network graph of correlated activity programs in the human breast, excluding modules with low topological overlap, for values of K at $(K_{opt} - 10)$ and $(K_{opt} + 10)$. Nodes represent distinct activity programs in the indicated cell types, and edges connect significantly correlated programs (Pearson correlation coefficient > 0 , p-value < 0.05). Modules of correlated programs were identified using a Constant Potts Model for community detection. *Bottom*: Plots depicting the number of subtrees identified at each K , weighted by the total number of programs in each subtree (see *STAR Methods*). Dashed lines depict the chosen values for K (K_{opt}), $(K_{opt} - 10)$, and $(K_{opt} + 10)$ for each cell type.

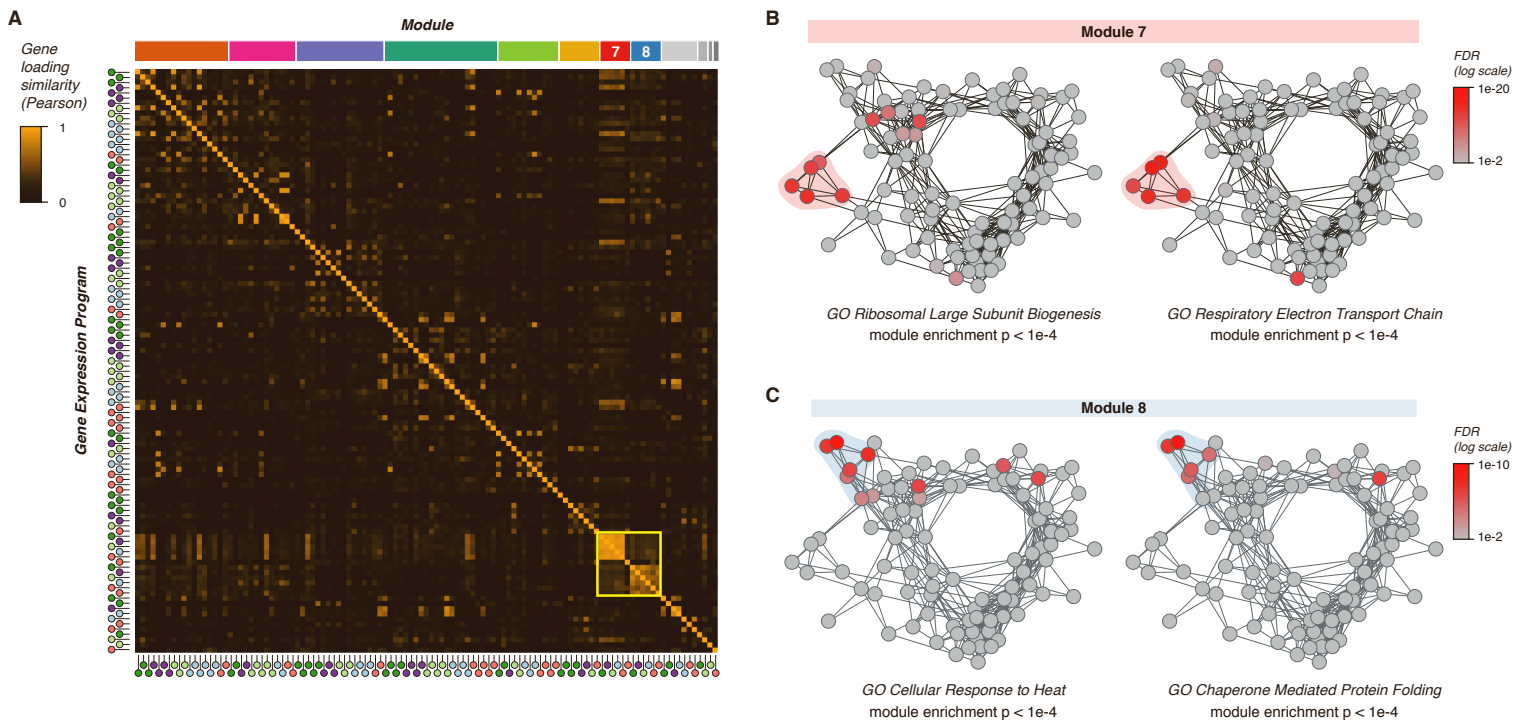


Figure S7 - Non-cell-type-specific shared transcriptional responses across cell types in the human breast

Figure S7. Non-cell-type-specific shared transcriptional responses across cell types in the human breast; Related to Figure 2

(A) Heatmap depicting Pearson correlation coefficients between gene loadings for all activity programs. Activity programs in modules 7 and 8 are highlighted by a yellow box.

(B) Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). Overall enrichment of gene sets within module 7 was determined by permutation analysis.

(C) Network graph of activity programs, colored by the FDR for enrichment of the indicated gene sets in each activity program (log scale). Overall enrichment of gene sets within module 8 was determined by permutation analysis.

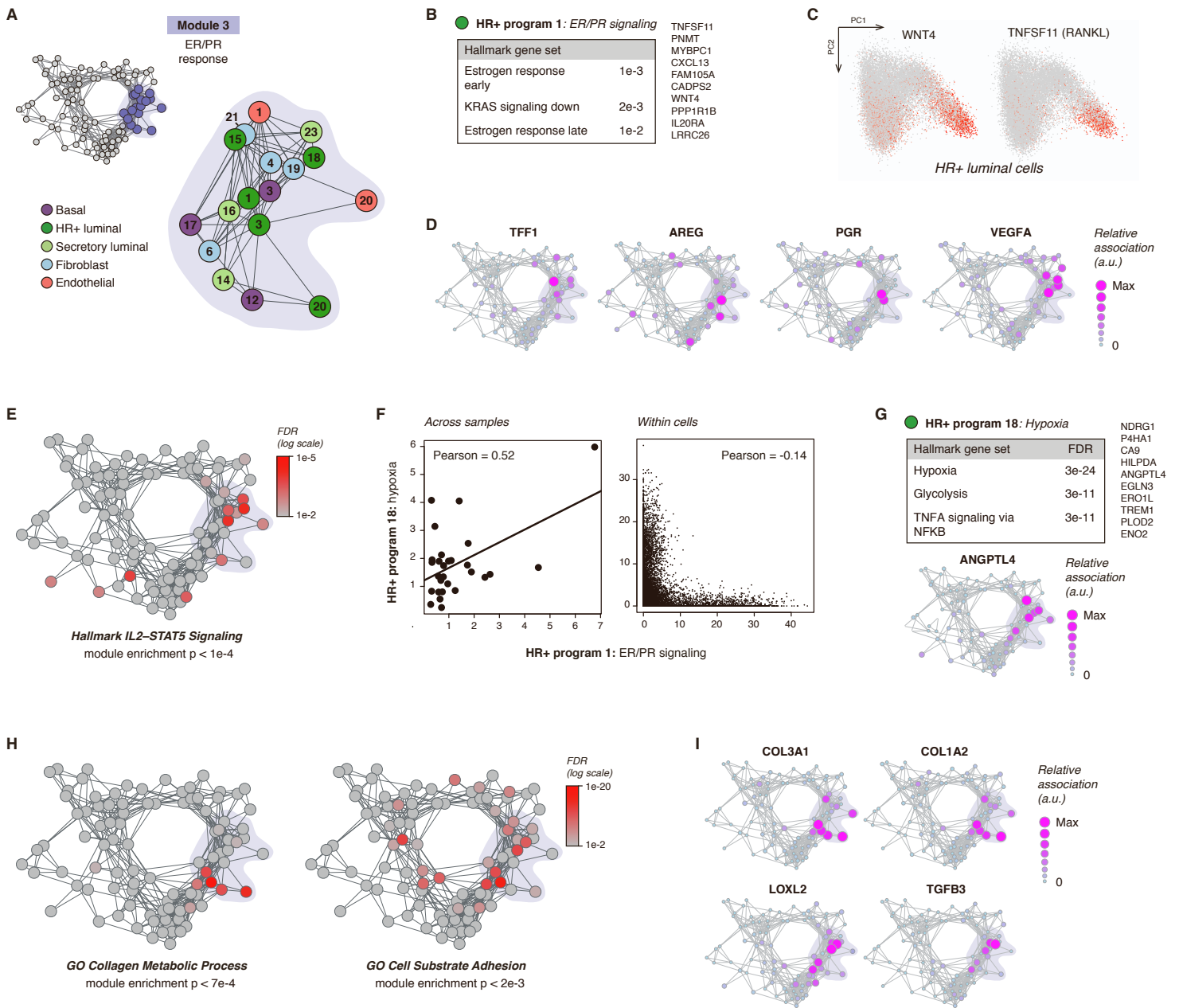


Figure S8 - The “ER/PR response” module

Figure S8. The “ER/PR response” module; Related to Figure 3

(A) Diagram highlighting activity programs in the “ER/PR response” module.

(B) Gene set enrichment analysis for HR+ activity program 1 (“ER/PR signaling”), showing the top pathways identified from the Molecular Signatures Database Hallmark gene sets, and the top 10 gene loadings in HR+ program 1.

(C) PCA plot of HR+ luminal cells depicting expression of WNT4 and TNFSF11 (RANKL) in log normalized counts.

(D) Network graph of activity programs, depicting the relative association of the indicated hormone-responsive genes with each activity program (arbitrary units, linear scale).

(E) Network graph of activity programs, colored by the FDR for enrichment of the “IL2-STAT5 signaling” Hallmark gene set in each activity program (log scale). Overall enrichment within module 3 was determined by permutation analysis.

(F) *Left:* Plot of the average expression score for HR+ activity program 1 (“ER/PR signaling”) versus activity program 18 (“hypoxia”) (Pearson $r = 0.52$). Dots represent the average expression score of each gene program within a sample. *Right:* Scatter plot of the per-cell average expression score for HR+ activity program 1 (“ER/PR signaling”) versus activity program 18 (“hypoxia”) (Pearson $r = -0.14$). Dots represent the expression score of each activity program within individual HR+ luminal cells.

(G) *Top:* Gene set enrichment analysis for HR+ activity program 18 (“hypoxia”), showing the top pathways identified from the Molecular Signatures Database Hallmark gene sets, and the top 10 gene loadings in HR+ program 18. *Bottom:* Network graph of activity programs, depicting the relative association of the hypoxia-related gene ANGPTL4 with each activity program (arbitrary units, linear scale).

(H) Network graph of activity programs, colored by the FDR for enrichment of the indicated Molecular Signatures Database gene sets in each activity program (log scale). Overall enrichment of gene sets within module 3 was determined by permutation analysis.

(I) Network graph of activity programs, depicting the relative association of the indicated marker genes with each activity program (arbitrary units, linear scale).

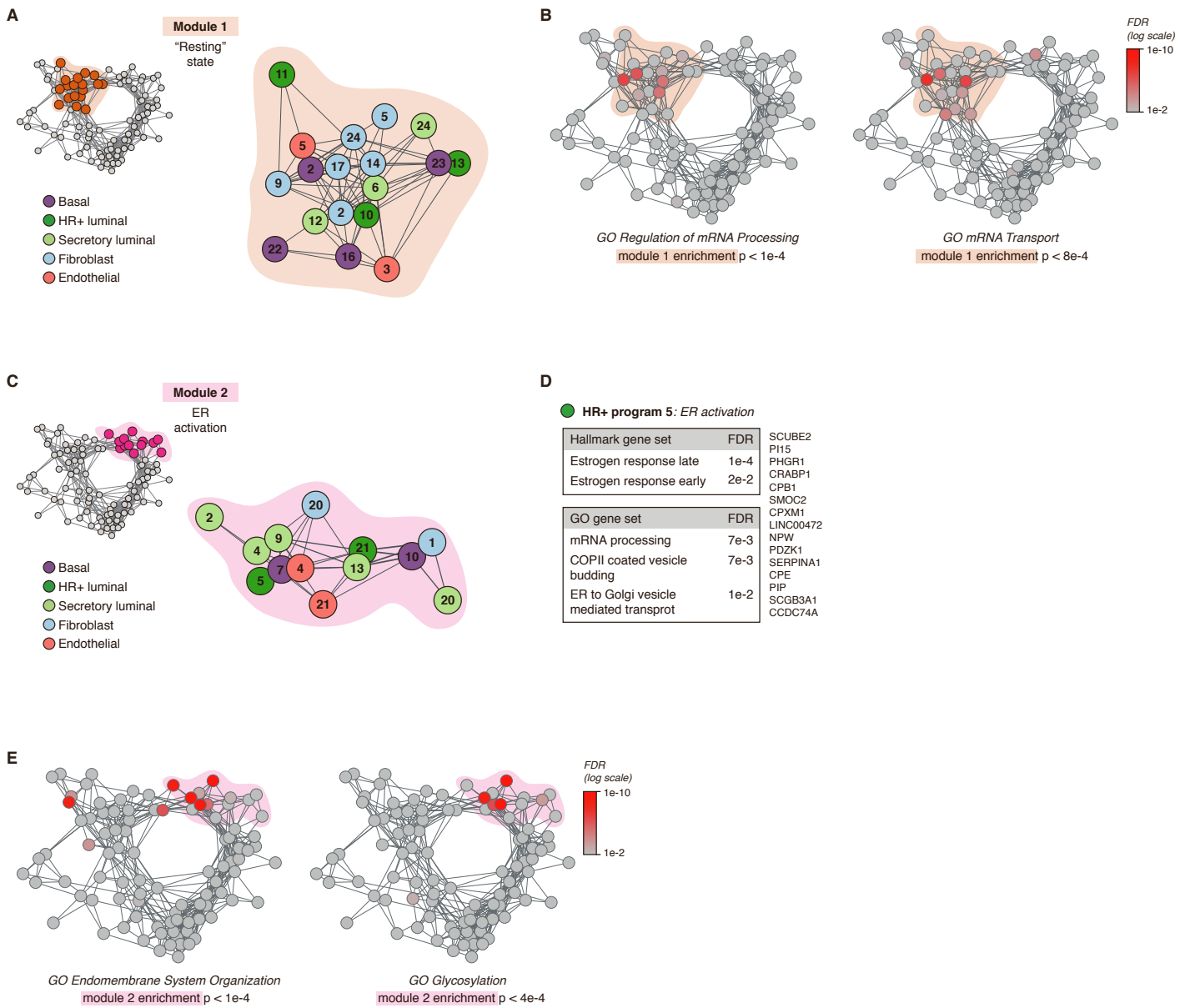


Figure S9 - The "Resting state" and "Estrogen receptor activation" modules

Figure S9. The “Resting state” and “Estrogen receptor (ER) activation” modules; Related to Figure 4

(A) Diagram highlighting activity programs in the “Resting state” module.

(B) Network graph of activity programs, colored by the FDR for enrichment of the indicated Molecular Signatures Database gene sets in each activity program (log scale). Overall enrichment of gene sets within module 1 was determined by permutation analysis.

(C) Diagram highlighting activity programs in the “ER activation” module.

(D) Gene set enrichment analysis for HR+ activity program 5 (“ER activation”), showing the top pathways identified from the Molecular Signatures Database Hallmark and GO Biological Process gene sets, and the top 15 gene loadings in HR+ program 5.

(E) Network graph of activity programs, colored by the FDR for enrichment of the indicated Molecular Signatures Database gene sets in each activity program (log scale). Overall enrichment of gene sets within module 2 was determined by permutation analysis.

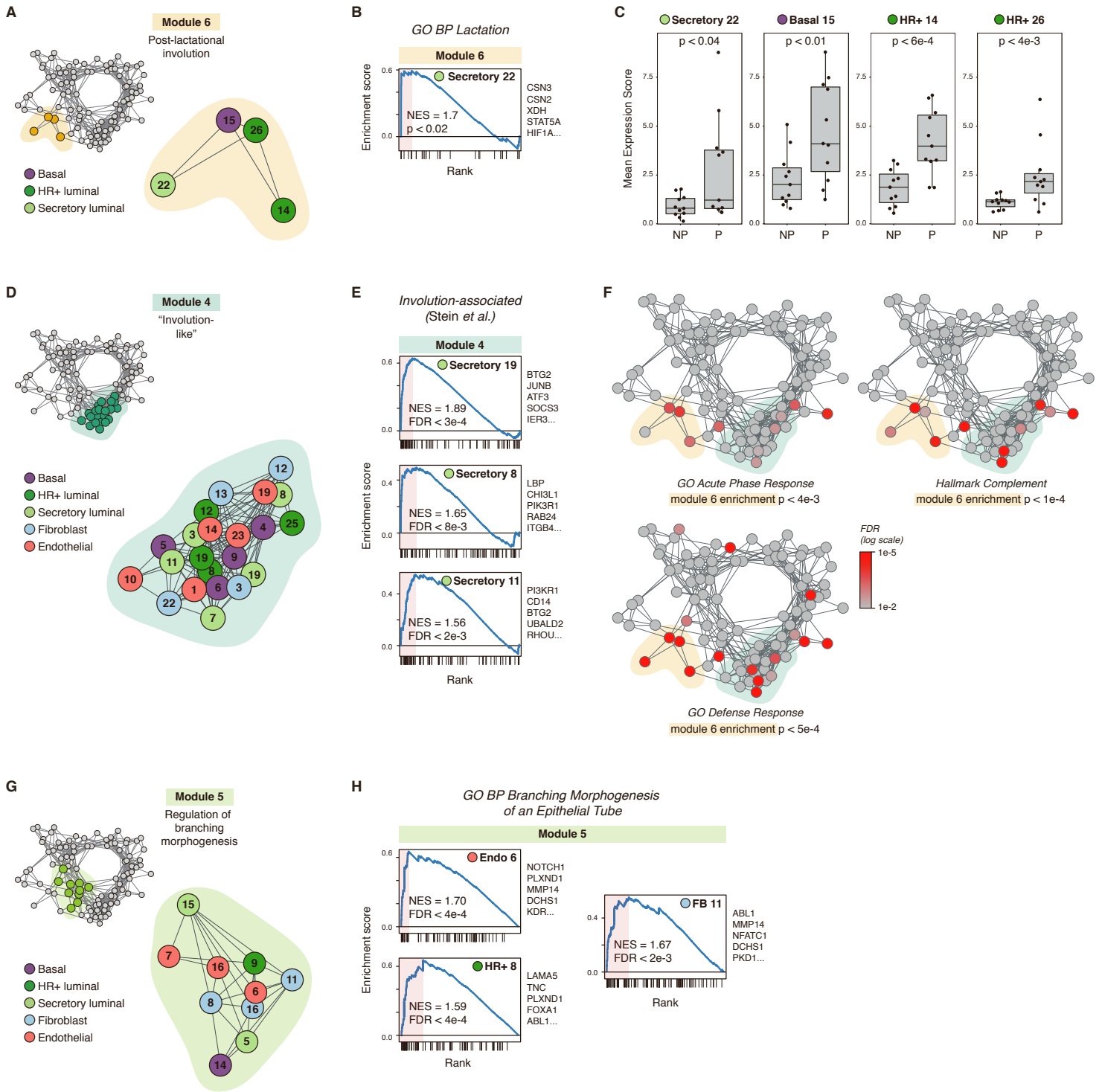


Figure S10 - The "Post-lactational involution", "Involution-like", and "Regulation of branching morphogenesis" modules

Figure S10. The “Post-lactational involution”, “Involution-like”, and “Regulation of branching morphogenesis” modules; Related to Figure 4

(A) Diagram highlighting activity programs in the “Post-lactational involution” module.

(B) Gene set enrichment analysis of Secretory cell activity program 22 in the “Post-lactational involution” module, showing enrichment of genes in the Molecular Signatures Database GO “Lactation” gene set. The top five leading edge genes are listed.

(C) Quantification of the mean expression score for the indicated activity programs for nulliparous (NP) versus parous (P) samples (n = 22 samples, $p < 0.05$, Mann-Whitney test).

(D) Diagram highlighting activity programs in the “Involution-like” module.

(E) Gene set enrichment analysis of the indicated activity programs in the “Involution-like” module, showing enrichment of genes previously shown to be upregulated during post-lactational involution in the mouse (Stein et al., 2004). The top five leading edge genes for each activity program are listed.

(F) Network graph of activity programs, colored by the FDR for enrichment of the indicated Molecular Signatures Database gene sets in each activity program (log scale). Overall enrichment of gene sets within module 6 was determined by permutation analysis.

(G) Diagram highlighting activity programs in the “Regulation of branching morphogenesis” module.

(H) Gene set enrichment analysis of the indicated activity programs in the “Regulation of branching morphogenesis” module, showing enrichment of the Molecular Signatures Database gene set “Branching Morphogenesis of an Epithelial Tube”. The top five leading edge genes for each activity program are listed.

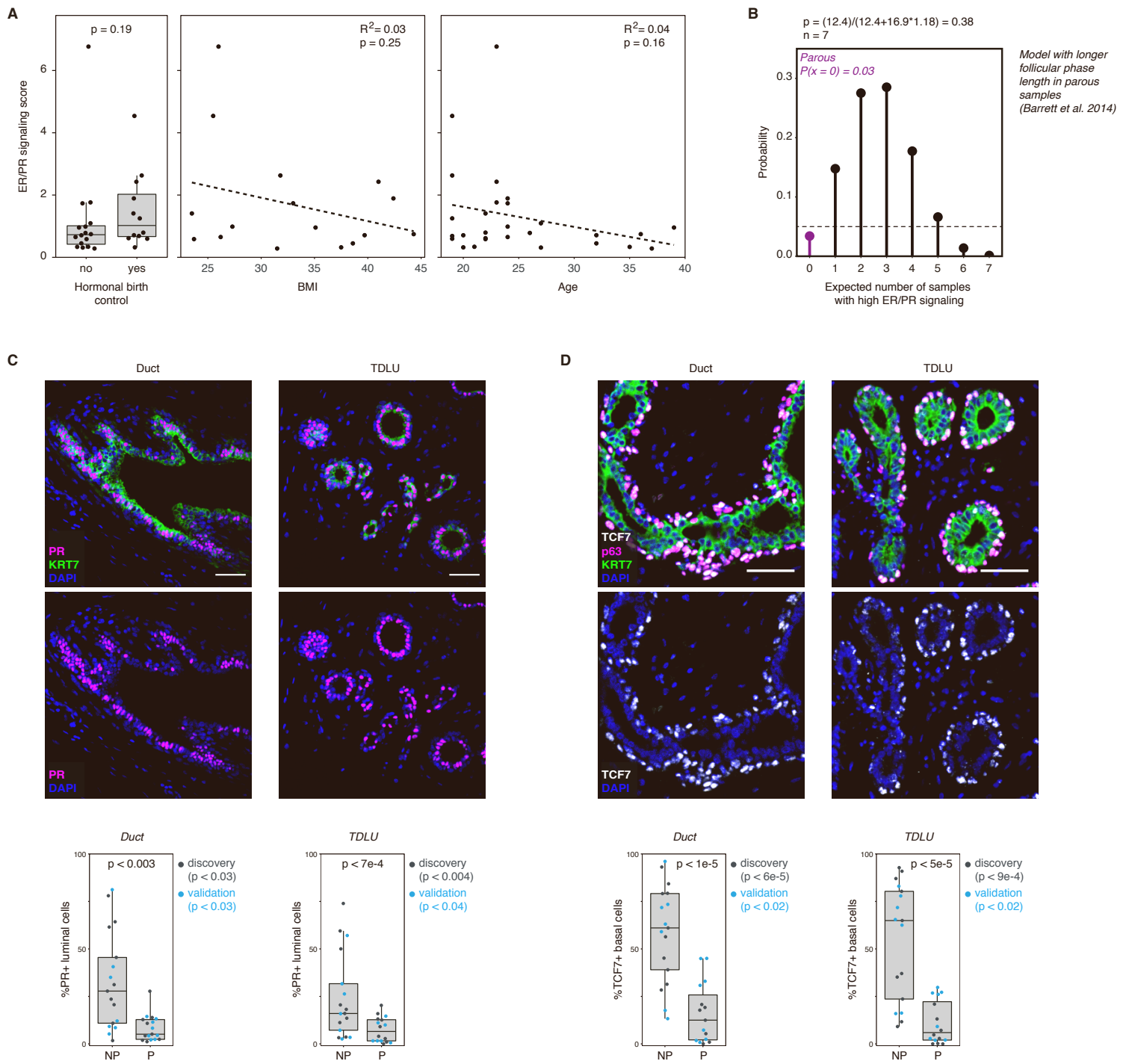


Figure S11 - Reduced ER/PR signaling in parous samples

Figure S11. Reduced ER/PR signaling in parous samples; Related to Figure 5

(A) Quantification of the average expression score of HR+ program 1 (ER/PR signaling) across the indicated biological variables (Hormonal contraceptive use: $n = 28$ samples, $p = 0.19$, Mann-Whitney test; BMI: $n = 16$, $R^2 = 0.01$, $p = 0.25$, Wald test; Age: $n = 28$ samples, $R^2 = 0.04$, $p = 0.16$, Wald test).

(B) Binomial probability distribution for the expected number of parous samples with high ER/PR signaling. The binomial probability of high ER/PR signaling was modeled as the average length of the luteal phase of the menstrual cycle in parous women, in days, divided by the average total length of the menstrual cycle in parous women ($P = 0.03$) (Barrett et al., 2014).

(C) Immunostaining for PR and KRT7 in ducts and terminal ductal lobular units (TDLUs), and quantification of the percentage of PR+ cells within the KRT7+ luminal compartment for nulliparous (NP) versus parous (P) samples ($n = 34$ samples, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples ("discovery set", $n=19$ samples) and a second independent cohort of samples ("validation" set, $n = 15$ samples). Scale bars 100 μm .

(D) Immunostaining for TCF7, p63, and KRT7 in ducts and terminal ductal lobular units (TDLUs), and quantification of the percentage of TCF7+ cells within the p63+ basal/myoepithelial cell compartment for nulliparous (NP) versus parous (P) samples ($n = 33$ samples, Mann-Whitney test). Results are shown for both the original cohort of sequenced samples ("discovery set", $n=18$ samples) and a second independent cohort of samples ("validation" set, $n = 15$ samples). Scale bars 50 μm .

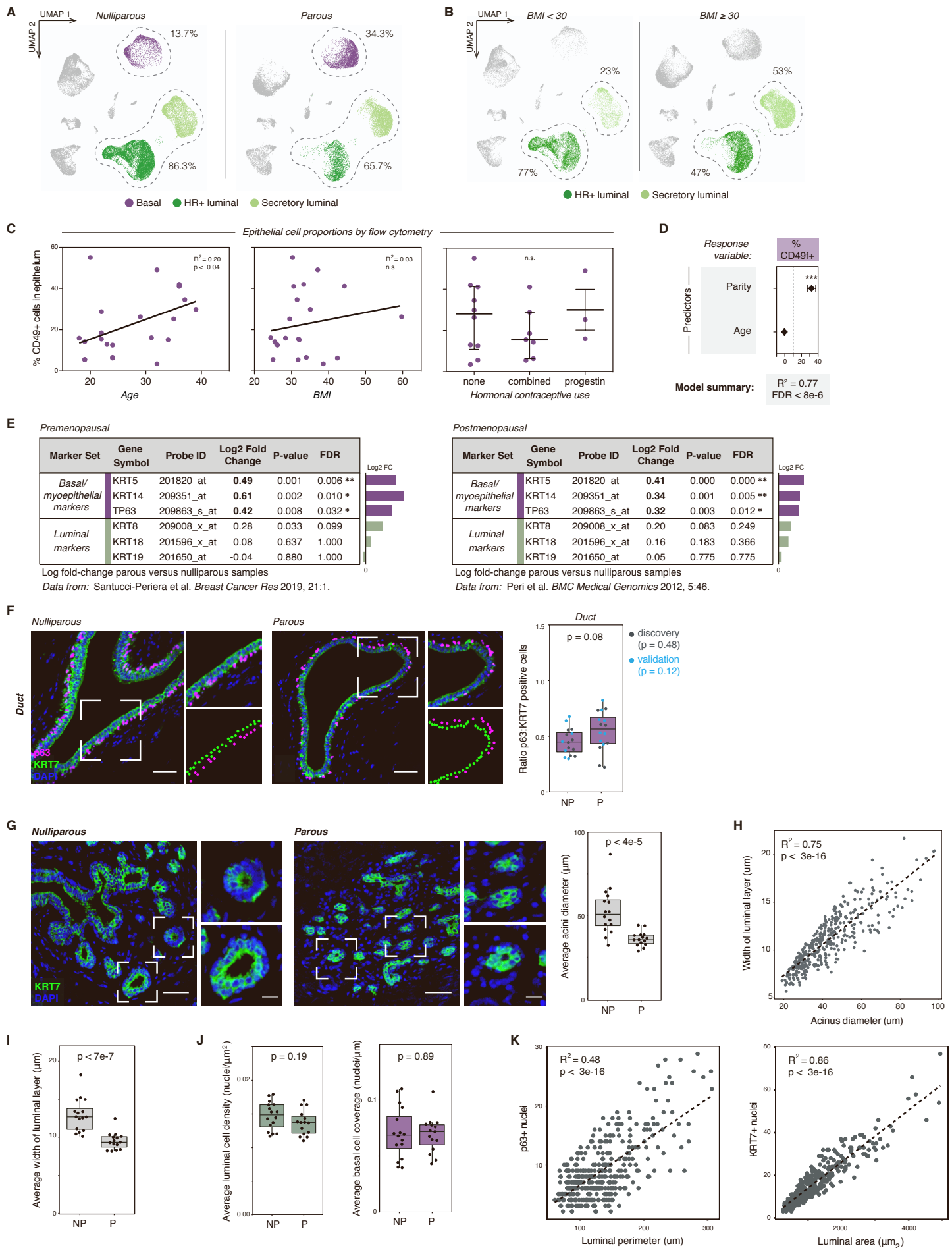


Figure S12 - Prior history of pregnancy is associated with an increased proportion of basal cells in the mammary epithelium

Figure S12. Prior history of pregnancy is associated with an increased proportion of basal cells in the mammary epithelium; Related to Figure 6

(A) UMAP plot of sorted live/singlet and epithelial cells from nulliparous and parous samples from the reduction mammoplasty dataset (GSE198732), with the percent of luminal and basal/myoepithelial cells highlighted.

(B) UMAP plot of sorted live/singlet and epithelial cells from non-obese (BMI < 30) and obese (BMI ≥ 30) samples from the reduction mammoplasty dataset (GSE198732), highlighting hormone-responsive (HR+) and secretory luminal cells.

(C) Quantification of the percentage of EpCAM⁻/CD49f⁺ basal cells identified by flow cytometry analysis versus age (n = 23; R² = 0.20; p < 0.04, Wald test), body mass index (n = 21; R² = 0.03; p = 0.44, Wald test), or hormonal contraceptive use (n = 23; p = 0.50, Kruskal-Wallis test).

(D) Results from multiple linear regression analysis, with prior pregnancy (parity) and age as predictors and the percentage of EpCAM⁻/CD49f⁺ basal cells in the epithelium as the response variable (Parity p < 2e-5; Age p = 0.17; overall R² = 0.77, p < 8e-6, Wald test).

(E) Microarray differential expression analysis for selected genes from Santucci-Periera *et al.* and Peri *et al.* (Peri *et al.*, 2012; Santucci-Pereira *et al.*, 2019).

(F) Immunostaining for the basal/myoepithelial marker p63 and pan-luminal marker KRT7 in ductal regions, and quantification of the ratio of p63+ basal cells to KRT7+ luminal cells in nulliparous (NP) versus parous (P) women (n = 32 samples; p = 0.08, Mann-Whitney test). Results are shown for a subset of the original cohort of sequenced samples (“discovery set”, n=17 samples, p = 0.48) and a second independent cohort of samples (“validation” set, n = 15 samples, p = 0.12). Scale bars 50 μm.

(G) Immunostaining for the pan-luminal marker KRT7, and quantification of the average acinar diameter in TDLUs from nulliparous (NP) versus parous (P) samples (n = 31 samples, p < 4e-5, Mann-Whitney test). Scale bars 50 μm. Inset scale bars 15 μm.

(H) Linear regression analysis of the width of the luminal layer versus acinus diameter for individual acini (n = 391 acini from 37 samples; R² = 0.75, p < 3e-16, Wald test).

(I) Quantification of the average width of the luminal layer in acini from TDLUs in nulliparous (NP) versus parous (P) samples (n = 37 samples; p < 7e-7, Mann-Whitney test).

(J) Quantification of the average luminal cell density (nuclei per μm² of luminal area) and basal cell coverage (nuclei per μm of luminal perimeter) in acini from TDLUs in nulliparous (NP) versus parous (P) samples (n = 37 samples; luminal cell density: p = 0.19, Mann-Whitney test; basal cell coverage: p = 0.89, Mann-Whitney test).

(K) *Left*: Linear regression analysis of the perimeter of the luminal layer versus the number of p63+ basal cells for individual acini (n = 391 acini from 37 samples; R² = 0.48, p < 3e-16, Wald test). *Right*: Linear regression analysis of the area of the luminal layer versus the number of KRT7+ luminal cells for individual acini (n = 391 acini from 37 samples; R² = 0.86, p < 3e-16, Wald test).

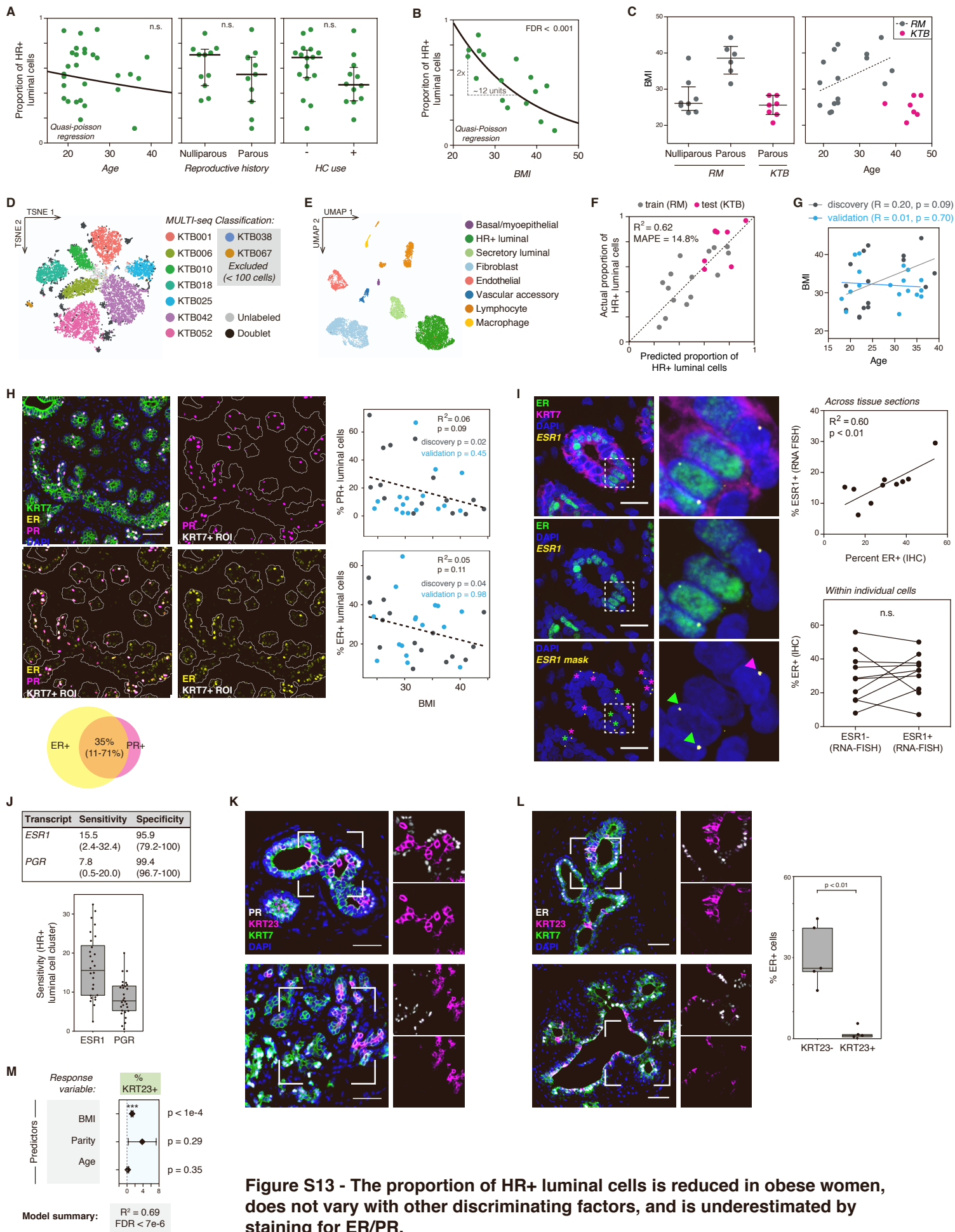


Figure S13 - The proportion of HR+ luminal cells is reduced in obese women, does not vary with other discriminating factors, and is underestimated by staining for ER/PR.

Figure S13. The proportion of HR+ luminal cells is reduced in obese women, does not vary with other discriminating factors, and is underestimated by staining for ER/PR; Related to Figure 6

(A) Proportion of HR+ luminal cells in each sample identified by scRNA-seq clustering, as a function of age, reproductive history, or hormonal contraceptive (HC) use (Wald test).

(B) Quasi-Poisson regression model of the proportion of HR+ cells in the luminal compartment as a function of BMI (FDR < 0.001, Wald test).

(C) Scatter plots highlighting differences in body mass index (BMI), reproductive history, and age between the Komen Tissue Bank (KTB) and reduction mammoplasty cohorts (see also Table S1). The trend line depicts the positive association of BMI with age in the reduction mammoplasty cohort, which is absent in the KTB cohort.

(D) TSNE dimensionality reduction of the normalized barcode count matrices and final sample classification for MULTI-seq barcoding of KTB samples.

(E) UMAP dimensionality reduction and unsupervised clustering of the combined data from seven KTB samples (GSE198732, Table S1) identifies the major epithelial and stromal cell types in the breast.

(F) A quasi-Poisson regression model accurately predicts the proportion of HR+ cells in the luminal compartment as a function of BMI in an independent cohort of core biopsy samples from the Komen Tissue Bank (KTB; mean absolute percentage error = 14.8%).

(G) Linear regression analysis of BMI versus age in the “discovery” and “validation” cohorts. The “validation” cohort is well-balanced across age and BMI.

(H) *Left:* Representative image of co-immunostaining of ER, PR, and KRT7. *Top right:* Linear regression analysis of the percentage of PR+ luminal cells versus BMI (n = 31 samples, $R^2 = 0.06$, $p = 0.09$, Wald test). *Bottom right:* Linear regression analysis of the percentage of ER+ luminal cells versus BMI (n = 29 samples $R^2 = 0.05$, $p = 0.11$, Wald test). Scale bars 50 μm . Venn diagram highlights the average percent overlap between ER and PR as measured by immunostaining, for samples where both ER and PR were co-immunostained (n = 5 samples, range = 11-71%).

(I) Multiplexed *in situ* hybridization of estrogen receptor transcript (ESR1) and immunostaining for estrogen receptor protein (ER) and KRT7. Scale bars 25 μm . *Right:* Quantification of the expression of ESR1 and ER across multiple tissue sections ($R^2 = 0.6$, $p < 0.01$, Wald test) or within individual cells ($p = 0.63$, Wilcoxon matched pairs signed-rank test).

(J) Table and bar plot depicting the sensitivity and specificity for ESR1 or PGR transcript expression in the HR+ luminal cell versus secretory luminal cell cluster based on scRNA-seq analysis.

(K) Representative images of co-immunostaining of PR, KRT23, and the pan-luminal marker KRT7. (See quantification of n = 41 samples in Figure 6F).

(L) Co-immunostaining of ER, KRT23, and the pan-luminal marker KRT7 and quantification of the percentage of ER+ cells within the KRT7+/KRT23- and KRT7+/KRT23+ luminal cell populations (n = 5 samples; $p < 0.01$ Mann-Whitney test). Scale bars = 50 μm .

(M) Results from multiple linear regression analysis, with body mass index (BMI), prior pregnancy (parity), and age as predictors and the percentage KRT23+ cells in the KRT7+ luminal compartment as the response variable (BMI $p < 1e-4$; Parity $p = 0.29$; Age $p = 0.35$; overall $R^2 = 0.69$, $p < 7e-6$, Wald test).

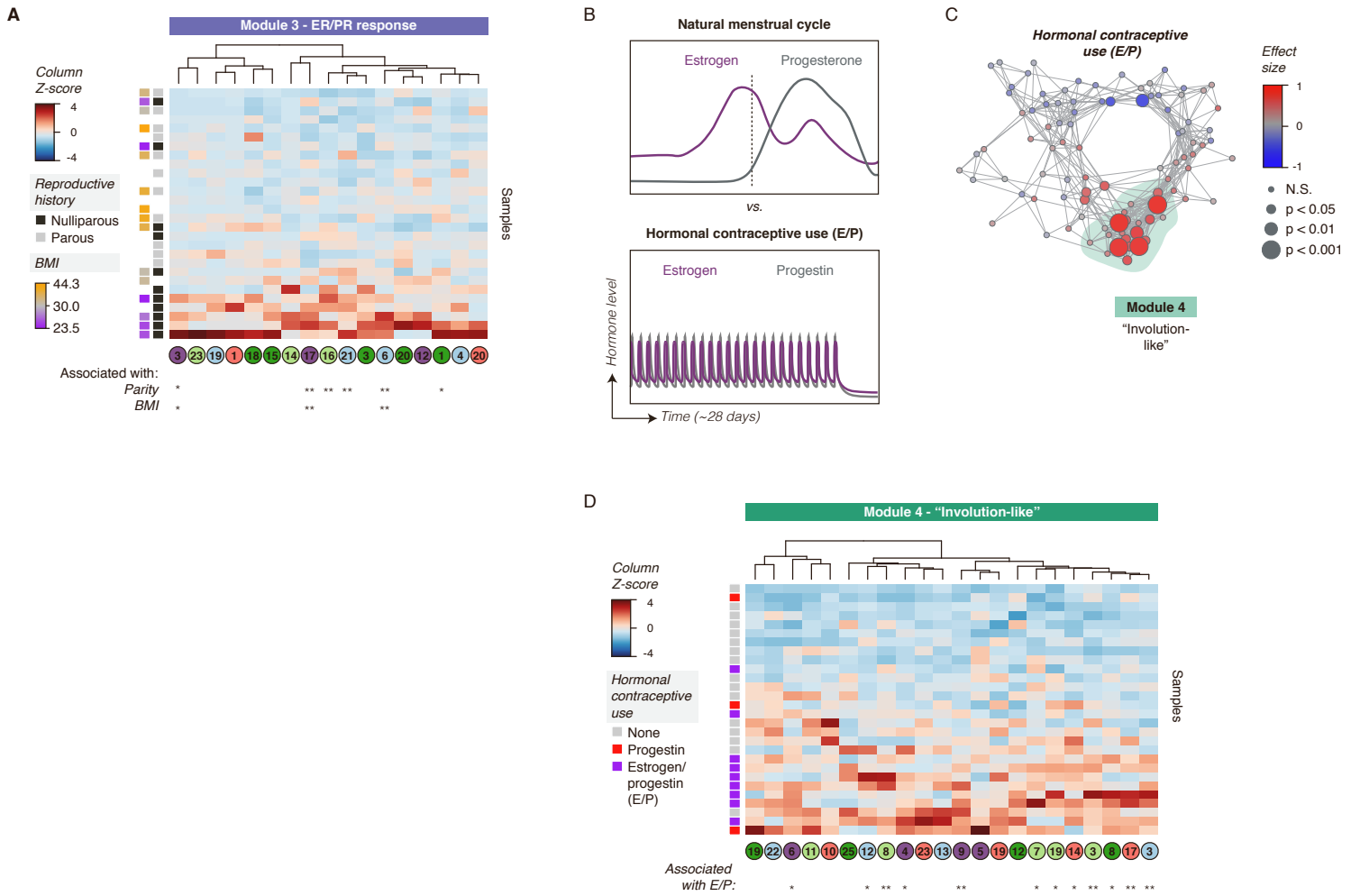


Figure S14 - Association of cell-cell interaction modules with reproductive history, body mass index, and hormonal contraceptive use

Figure S14. Association of cell-cell interaction modules with reproductive history, body mass index, and hormonal contraceptive use; Related to Figure 7

(A) Heatmap depicting the relative average expression score (column z-score, linear scale) of the indicated activity programs in each sample. Samples were ordered by the mean Z-score of all programs within the “ER/PR response” module. Activity programs were ordered by hierarchical clustering (complete linkage). Significant associations with parity or BMI are indicated by asterisks (* $p < 0.05$, ** $p < 0.01$; parity: Mann-Whitney test; BMI: Wald test).

(B) Schematic depicting relative estrogen and progesterin levels and dynamics across the natural menstrual cycle and in donors using combination (E/P) hormonal contraceptives. Samples from donors using hormonal contraceptives were used as a “virtual experiment” to test the effects of estrogen/progesterone treatment on downstream signaling pathways.

(C) Network graph of activity programs in the human breast, colored by the effect size of combined hormonal contraceptive use (Wilcoxon effect size, linear scale) on each activity program. Significant positive and negative associations are represented by larger nodes (Mann-Whitney test).

(D) Heatmap depicting the relative average expression score (column z-score, linear scale) of the indicated activity programs in each sample. Samples were ordered by the mean Z-score of all programs within the “Involution-like” module. Activity programs were ordered by hierarchical clustering (complete linkage). Significant associations with combined hormonal contraceptive use are indicated by asterisks (* $p < 0.05$, ** $p < 0.01$; Mann-Whitney test).