Single-Cell RNA Sequencing in Human Breast Tissue: A Comparative Study of Biological Signals and Technical Artifacts

Sophia K. Cheng
Team 13 - Signal Seekers
SIADS 699 - Spring/Summer 2025

Sophia K. Cheng
Team 13 - Signal Seekers
SIADS 699 - Spring/Summer 2025

Introduction

Single-Cell RNA Sequencing

Single-cell RNA sequencing (scRNA-seq) has played a pivotal role in advancing the understanding of biology by enabling researchers to measure gene expression at the resolution of individual cells. Through scRNA-seq analyses, researchers have created comprehensive cell atlases[1] and identified rare and/or previously unrecognized cellular subpopulations. Unlike bulk RNA sequencing (bulk RNA-seq), which uses whole tissue or or bulk-sorted cells[2] as inputs, scRNA-seq further breaks down the tissue samples into individual cells as inputs (Figure 1, Panel 1).

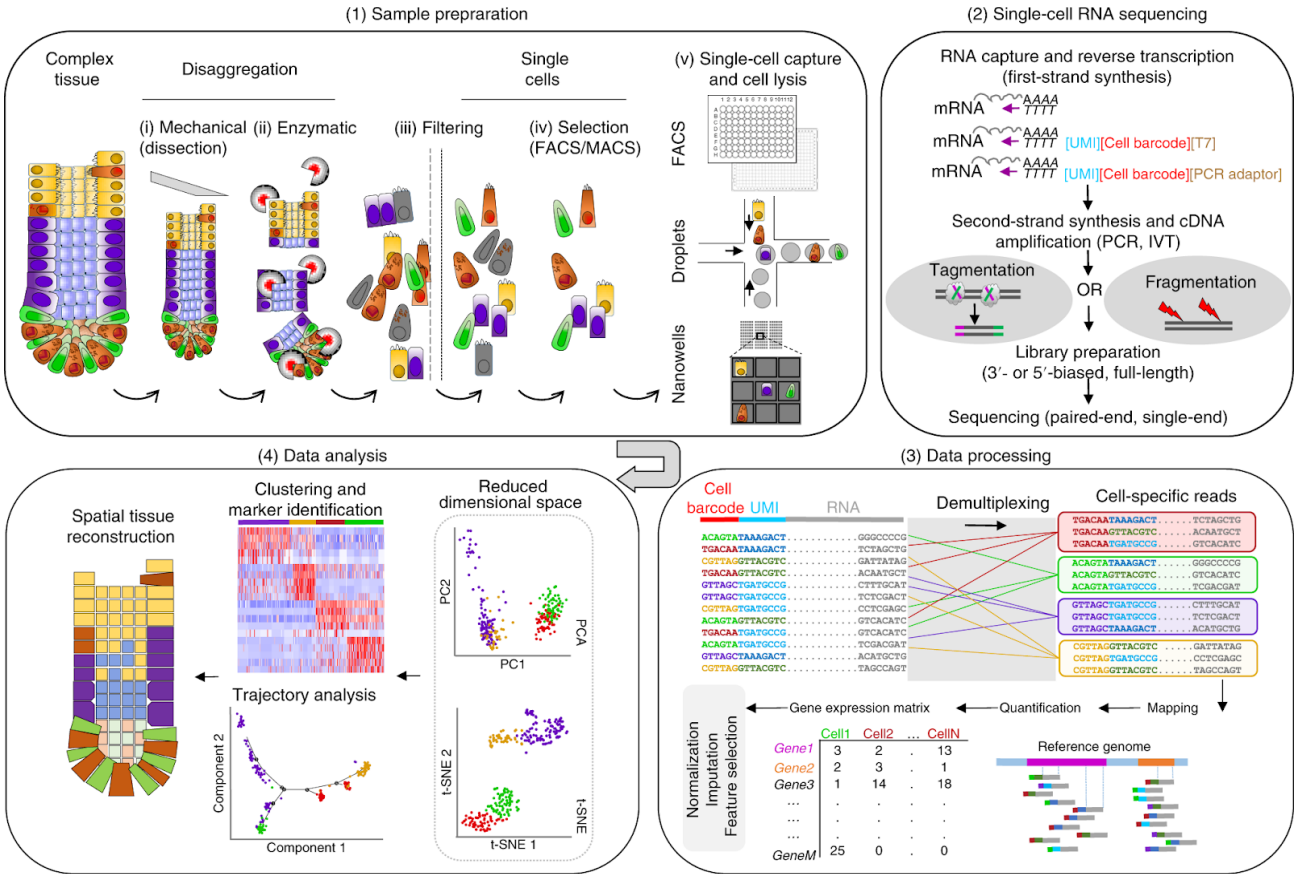


Figure 1. The workflow for scRNA-seq [1].
(1) Individual cells are isolated from the sample. (2) Cells are sequenced by tagging RNA molecules with a barcode. Barcodes are unique to each cell. (3) RNA molecules are collected by barcode to produce the sequence for the cell. (4) Typical analysis performed on scRNA data.

[1] www.humancellatlas.org
[2] **Bulk-sorted** refers to groups of cells that have been pooled together for analysis.

A necessary byproduct of this level of resolution is a dramatic increase in the number of observations, often by 3 to 4 orders of magnitude, resulting in significantly more data for downstream analysis. Another challenge is that the process of tagging mRNA may incorrectly tag mRNA from multiple cells with the same barcode or fail to tag anything at all. These two challenges highlight the importance of verifying the quality of the reads and filtering out noise. This is commonly done by calculating metrics such as total number of genes, percentage of genes that are for mitochondria[3], and total number of barcodes (cells)[4] that contain a gene. Thresholds are then determined for the dataset and cells that fall outside the threshold are filtered out from further analysis (Figure 2).

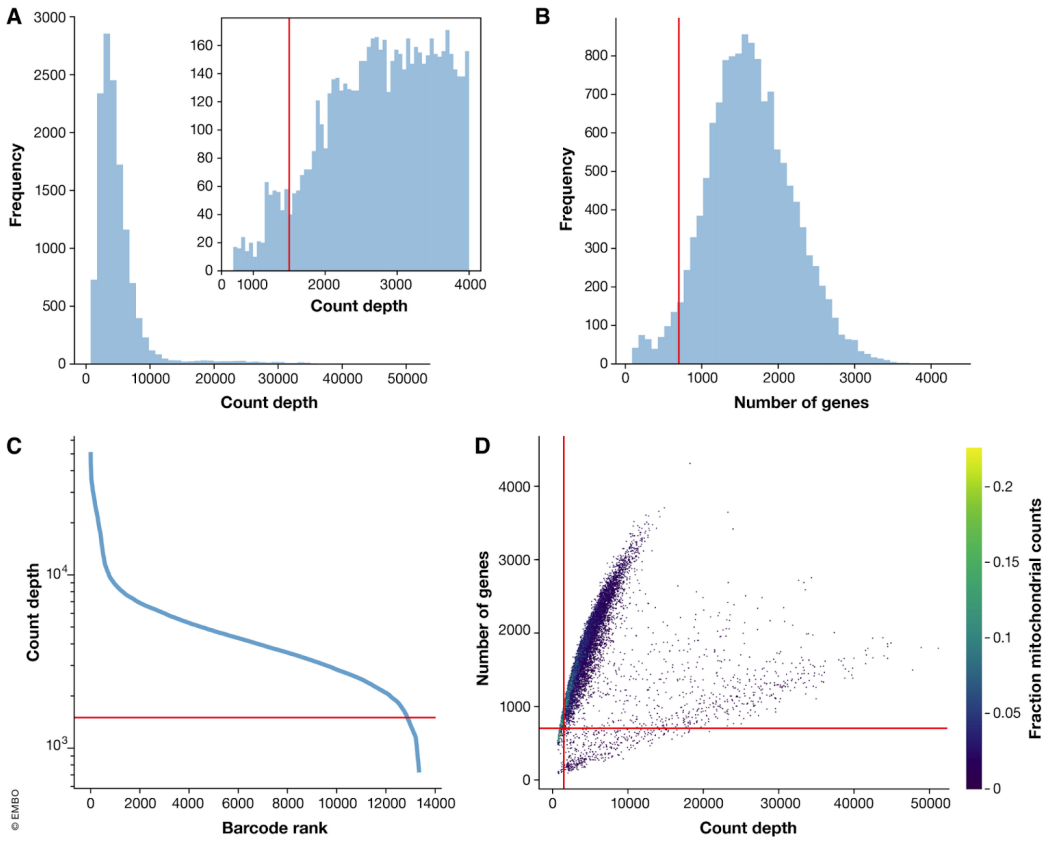


Figure 2. Plots of quality control metrics with filtering decisions for scRNA-seq dataset [2].

Motivation

This project is motivated by the causal ambiguity of identifying thresholds for quality control (QC) metrics in the pre-processing workflow. Specifically, thresholds for scRNA-seq are set using biological assumptions, while those same or related assumptions are being evaluated by scRNA-seq. One such biological assumption is that cells with higher total RNA are metabolically healthy. As a result, the QC

[3] **Mitochondria** are responsible for generating energy for cells to function.
[4] Each row of data corresponds to a unique molecular identifier (UMI) and usually referred to as a read. For readability, from this point forward, “cell” and “barcode” will be used synonymously with UMI.

process often prioritizes these cells, while treating cells with low total RNA counts as technical artifacts to be filtered out [3, 7]. This approach, while effective for minimizing noise from ambient RNA contamination, risks eliminating biologically meaningful signals.

Consider the DNA damage response (DDR) pathway, which is a mechanism used by cells to detect and repair their own DNA. If a cell's DNA is damaged beyond repair, the DDR pathway can trigger the cell to self-destruct and avoid propagation of harmful mutations. A damaged cell may leak RNA, which could result in lower than expected number of genes, or its stress response is activated. Notably, this RNA leakage can be quantified as a low total RNA count, one of the QC metrics used for thresholding. In the context of cancer, deficiencies in the DDR pathway are being targeted to develop novel therapies [4].

Further, recent evidence suggests that transcriptionally quiet cells may represent viable, functionally distinct cell states rather than technical artifacts [5, 6, 7]. These low-activity states may play important roles in dormancy, resistance to therapy, and metastatic reactivation. Dormant cancer cells, for example, can survive therapeutic treatments and later contribute to relapse, yet their scRNA-seq profile often overlaps with cells that have been filtered out by QC.

Feature to Threshold	**Filtered by QC Metric**	**Targeted by DDR**	**Dormant Cells**
Low total RNA content	✅ Damaged cell	⚠️ Depends	✅ Viable but quiet cell
High total RNA content	✅ Degraded cell	⚠️ Depends	✅ Limited active gene expression
Low number of genes	✅ Technical artifact	⚠️ Depends	✅ Limited active gene expression
Low mitochondrial RNA %	❌ Not filtered out	✅ Damaged cell	✅ Limited energy needs
High Mitochondrial RNA %	✅ Damaged cell	✅ Damaged cell	❌ Not dormant

Table 1. Summary of QC metric thresholds and how they correspond to different kinds of cells. (✅) Is a characteristic of the cell. (❌) Is not a characteristic of the cell. (⚠️) Might be a characteristic, but it depends.

Objective

The goal of this study is to perform a comparative scRNA-seq analysis of cells classified as biological signals ("real") versus those labeled as technical artifacts ("noise"), with the aim of evaluating whether current QC processes systematically exclude potentially informative cellular states. We[5] focus our analysis on human breast tissue for its heterogeneous cellular composition and biological complexity. The study is organized around two specific aims:

1. Baseline Validation - validate our classification of cells as either "real" or "noise" by reproducing one of the figures in the reference article.
2. Pathway Associated Gene Expressions - compare gene expression analysis of the pathways implicated in tumor proliferation and chemotherapy resistance.

[5] "**We**" and "**our**" are used for narrative flow, and not an indication of shared work.

Data Description

Sources

Our primary source of data is the scRNA-seq data provided by the authors of "A single-cell RNA expression atlas of normal, preoneoplastic and tumorigenic states in the human breast" [8], our reference article for this study. In addition to the size and variety of data, this article also provides a companion article that provides detailed information on the preprocessing and downstream analysis [9].

GEO[6] ID	GSE161529[7]
Number of patients	52
Number of cells	421,761
Profiles included	TNBC[8], BRCA1[9] TNBC, HER+[9], ER+[9], normal patients without breast cancer
Platform	10x Genomics Chromium

Table 2. Details of the scRNA-seq data provided by the reference article.

In addition to our primary source, we use gene signatures to reproduce epithelial cell typing and analyze pathway associated gene expression (Table 3).

Name	Purpose	Source
MaSC-enriched	Epithelial cell typing (basal)	Lim E, et al. [10]
Luminal progenitor	Epithelial cell typing (luminal progenitor)	Lim E, et al. [10]
Luminal mature	Epithelial cell typing (mature luminal)	Lim E, et al. [10]
Stroma	Epithelial cell typing (stromal)	Lim E, et al. [10]
Hallmark DNA Repair	Gene expression analysis	MSigDB[10] [11]
Hallmark E2F Targets	Gene expression analysis	MSigDB [11]
Hallmark G2M Checkpoint	Gene expression analysis	MSigDB [11]
Hallmark P53 Pathway	Gene expression analysis	MSigDB [11]

Table 3. Details of gene signatures used.

[6] **GEO** is the Gene Expression Omnibus, provided by the National Center for Biotechnology.
[7] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161529
[8] **TNBC** is a type of breast cancer known as triple-negative.
[9] **BRCA1, HER+, ER+** are proteins of interest in breast cancer research.
[10] **MSigDB** is the Human Molecular Signatures Database.

Finally, we make use of the supplementary data in the reference article to further annotate the dataset with additional features (Table 4).

Filename	**Purpose**
table_supplementary_1.xlsx	Tissue sample metadata
table_supplementary_2.xlsx	QC thresholds used and final cell count per sample
table_ev_4.xlsx	Tissue sample phenotype details

Table 4. Details of the supplementary data used.

Preprocessing

To facilitate analysis, we load the data as anndata[11] objects. Each tissue sample is provided as a folder with the cell barcodes and count matrix. An additional file contains the feature annotations for all tissue samples. We encapsulate the logic to reconstruct anndata's expected structure of a single folder per tissue sample that includes the barcodes, count matrix, and corresponding feature annotations.

After loading the data, we subset the anndata objects to the normal epithelial samples the authors used for profiling the normal breast epithelium (Appendix 2) for the rest of this study.

Feature Engineering

Target Feature (binary)

In order to perform a comparative analysis of "real" and "noise" cells, we first reverse-engineered the QC thresholds reported in the reference article. The supplementary table[12] details the initial counts for each sample, the threshold values, and final counts for each sample. We empirically determined whether the thresholds applied inclusively or exclusively and used this information to construct a binary feature `is_noise`. A value of `1` indicates the cell falls outside the threshold and would be filtered out as a technical artifact ("noise" cell), whereas a value of `0` indicates the cell falls within the threshold and would be considered a real biological signal ("real" cell). To validate our implementation of the thresholds, we compared our resulting counts of "real" cells with those reported in the supplementary table[12]. All but three samples matched exactly (table 5), and the mismatched samples were excluded from downstream analysis.

Sample Name	**Expected Count**	**Actual Count**
GSM4909296_ER-MH001	5754	4459
GSM4909313_ER-MH0064-T	3603	3604

[11] **anndata** is a python package to access and store annotated data matrices [12].
[12] **table** refers to the file table_supplementary_2.xlsx.

GSM4909319_mER-PM0178	7274	7674

Table 5. Samples that did not have the expected counts after applying QC thresholds.

Epithelium Cell Typing (categorical)

Next, we score each cell against the epithelium gene signatures to classify it as basal, luminal progenitor, mature luminal, stromal, or other (Figure 3). For each gene signature, we use the average log fold-change values to generate separate lists of upregulated and downregulated genes. A cell's score for a given signature is computed as the difference between the downregulated genes score and the upregulated genes score. We then apply a simple classification rule: the signature with the highest score is assigned as the predicted cell type (`predicted_type`). In the cases where the highest score is a negative value, the cell is predicted to be "`other`", limiting the scope to upregulated expressions of the signatures.

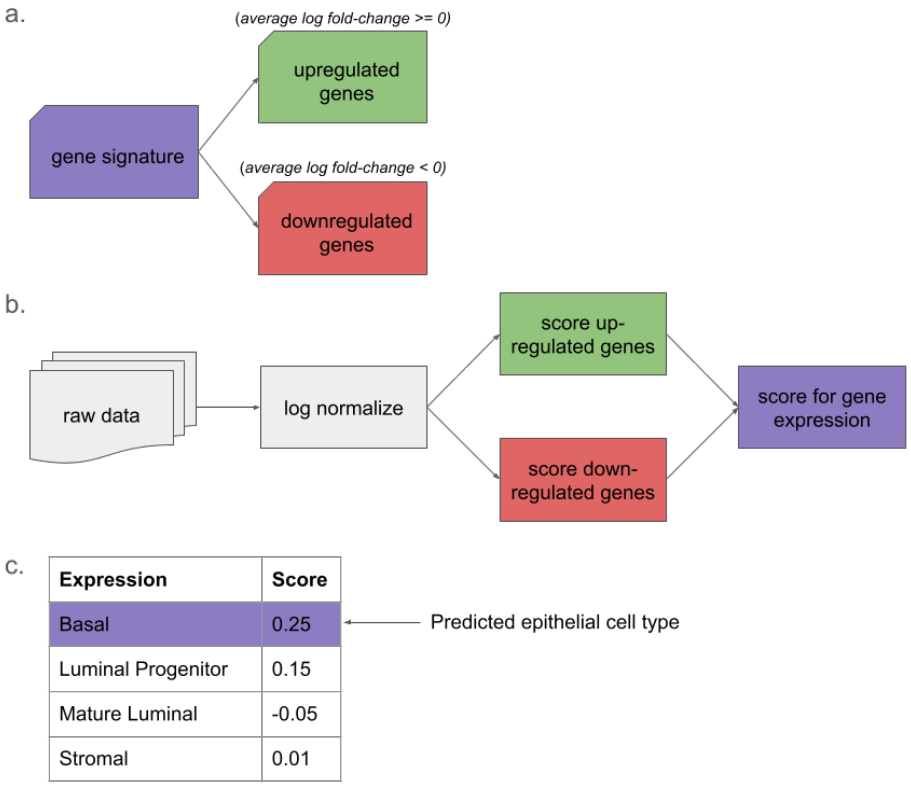


Figure 3. Workflow to predict a cell's epithelial type. Scores are illustrative and not indicative of actual values.

Specimen Metadata (categorical)

Finally, we annotate each cell with the following metadata using the supplementary data:

- specimen_id - Unique identifier for each tissue sample

- cancer_type - Indicator for type of cancer patient has
- cell_population - Indicator of tissue cell population is from

Baseline Validation

To mitigate our lack of domain expertise, we grounded our understanding by reproducing the t-distributed stochastic neighbor embedding (t-SNE) results from the reference article's analysis of epithelial cells from normal breast tissue samples (Appendix 1). To reproduce this analysis, we further split the normal epithelial dataset into "real" cells and "noise" cells, based on the values of `is_noise`, and removed the stromal cells as specified in the article [8]. Finally, we use a series of unsupervised learning techniques to cluster the data (Appendix 3).

Validate clustering approach

We validate our clustering approach by reproducing the t-SNE plot colored by cell type (Figure 4).

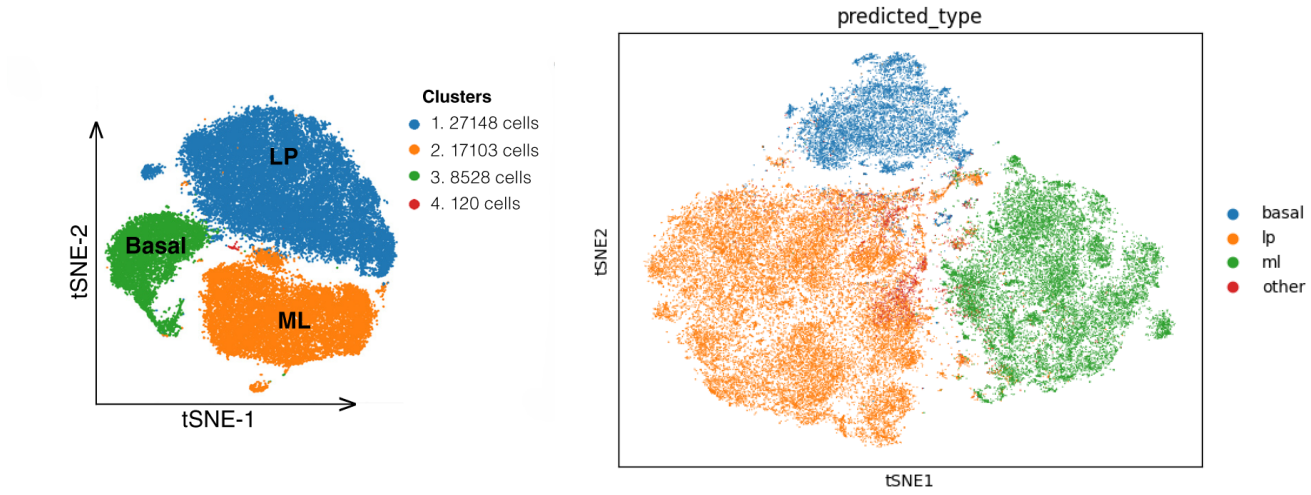


Figure 4. Visualization from the reference article [8] is shown on the left. Our reproduction is shown on the right.

Keeping in mind that the axes, orientations and positions of t-SNE are not informative, our clustering approach does approximate the article. In both plots, there is good separation between each of the cell types and the sizes of each cluster of cell types are comparable. One notable difference between the plots is the cluster density, which could be attributed to technical differences in the approach. In the article, the authors use R for their analysis whereas we are using Python, resulting in different implementations of the unsupervised learning algorithms used to compute the clusters. This supports that we have reverse-engineered a comparable clustering approach.

To further our understanding of the domain, we examined the clustering by visualizing the leiden cluster assignment, log-normalized total amount of RNA, log-normalized number of genes present in at least one cell, and the percentage of mitochondria (Figure 5). Examining the leiden cluster assignment (figure 6b) shows that cells along the interior spaces between the clusters may be in a transitional state from

one cell type to another. For example, the bottom right area of the basal cells (Figure 5a) shows cells that share a leiden cluster assignment while expressing the gene signature of a luminal progenitor and mature luminal cells. This finding is consistent with the article, “...suggesting that they represent transient intermediates prior to luminal lineage commitment.” [8]. Interestingly, the basal and luminal progenitor cells are considered the same leiden cluster. One naive interpretation is that they are more similar to each other than they are to mature luminal cells, which is supported by the developmental hierarchy of cells in breast tissue (Figure 6).

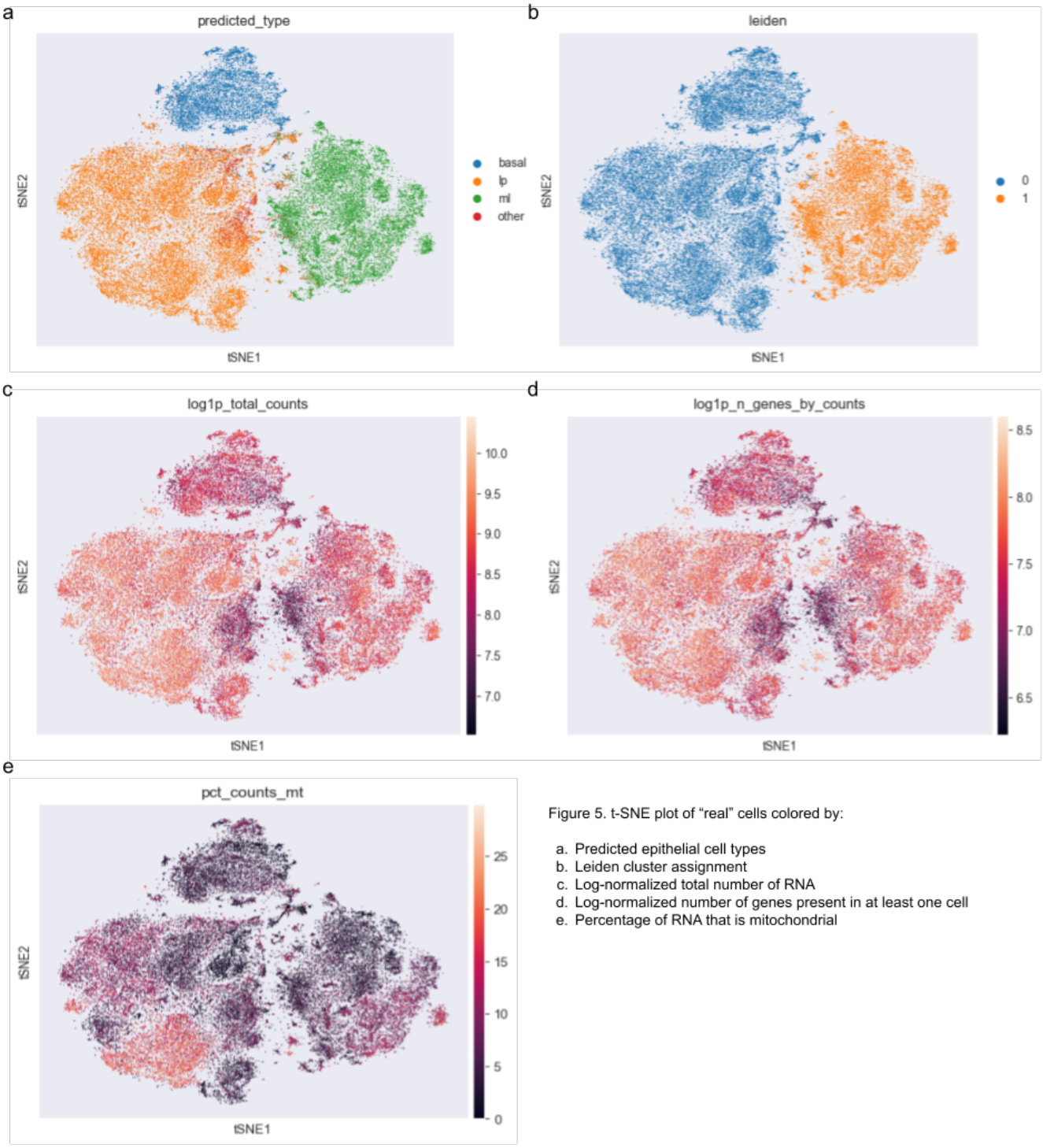


Figure 5. t-SNE plot of "real" cells colored by:

a. Predicted epithelial cell types
b. Leiden cluster assignment
c. Log-normalized total number of RNA
d. Log-normalized number of genes present in at least one cell
e. Percentage of RNA that is mitochondrial

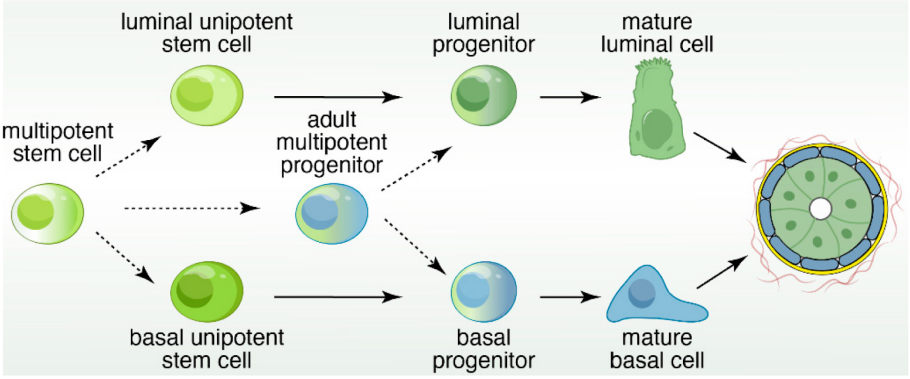


Figure 6. Developmental hierarchy in the mammary gland [13] to illustrate the relationship between the epithelial cell types.

Next, we examine the cell's QC characteristics with respect to its predicted cell type. On average, the total amount of RNA is proportional to the number of genes and is inversely proportional to the percentage of mitochondria. Specifically, the cells exhibit a low percentage of mitochondria and have a moderate amount of RNA and genes, indicating that these are viable, healthy cells.

Analysis of "noise" cells

Having established a basis for interpreting the embeddings for "real" cells, we look now at the embeddings for "noise" cells". For biological interpretability, we use the same workflow and parameters for clustering, despite the differences in amount of data (Table 6).

	"Noise" cells	**"Real" cells**
Number of cells	5014	53,136
Number of genes	51	233

Table 6. Size of the data for "noise" and "real" cells.

Looking first at the embeddings colored by cell type (Figure 7a), there are hints of a separation for each cell type suggesting some structures in the embedding, notably along the outer radius. Additionally, there appears to be a higher proportion of cells that are classified as other in the "noise" cells. Recall that if the highest epithelial cell type signature score is negative, we classify the cell as other. This may be an artifact we introduced by not tuning the parameters for the "noise" dataset. The algorithm [14] used for scoring calculates the average expression of a gene signature and a reference set of genes and returns the difference as the score. Given the smaller dataset for "noise" cells, the scoring algorithm is more sensitive to the individual fluctuations of each gene and may produce more negative scores.

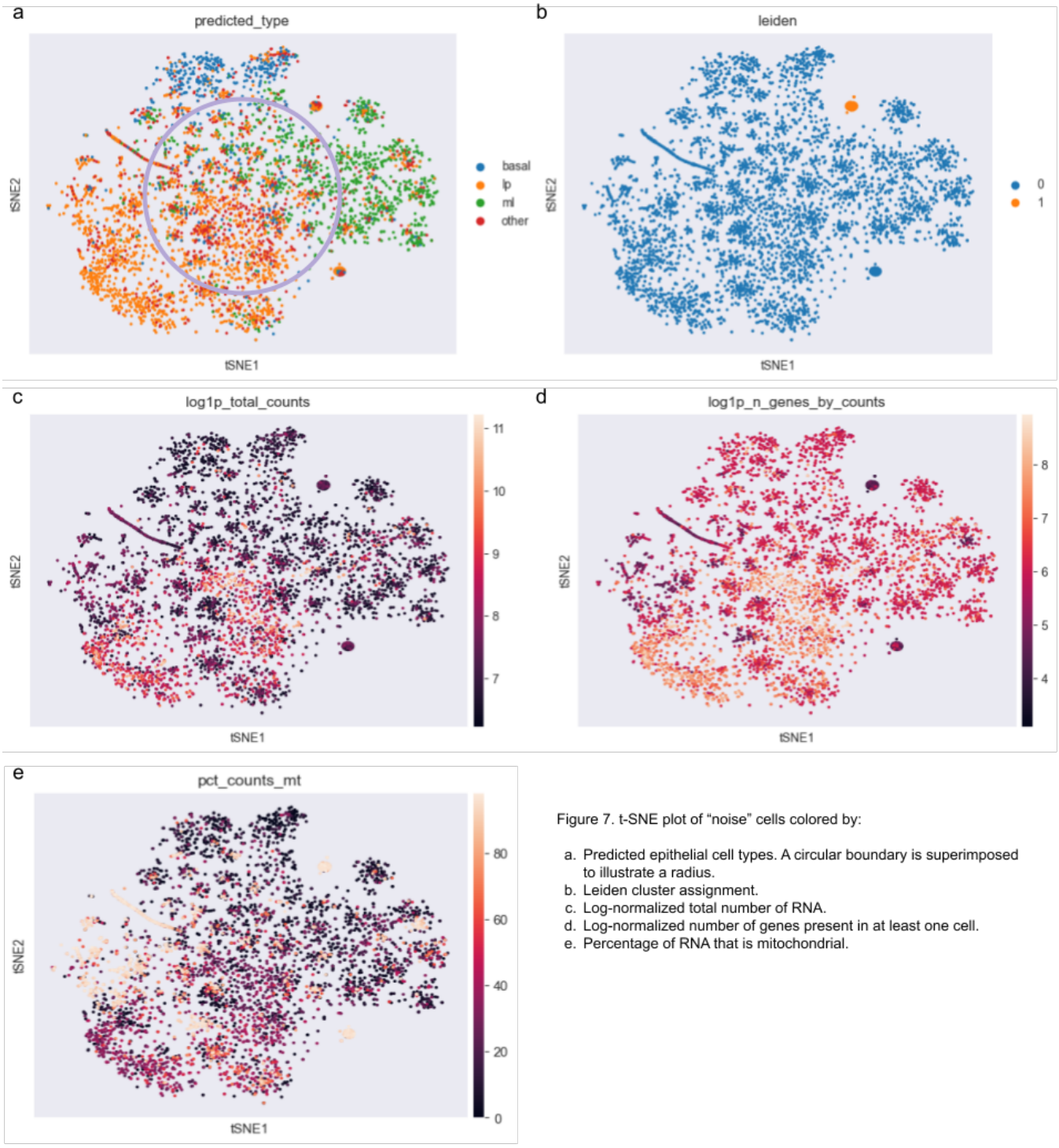


Figure 7. t-SNE plot of "noise" cells colored by:

a. Predicted epithelial cell types. A circular boundary is superimposed to illustrate a radius.
b. Leiden cluster assignment.
c. Log-normalized total number of RNA.
d. Log-normalized number of genes present in at least one cell.
e. Percentage of RNA that is mitochondrial.

The leiden clustering algorithm places the majority of the "noise" cells in the same cluster, with a relatively small secondary cluster of cells. The small secondary leiden cluster has a full mix of cell types that have an **extremely high percentage of mitochondria and low counts for RNA and genes**. This set of characteristics suggest that these cells are likely stressed, damaged, or dying. There are additional areas that are less well defined spread throughout the plot, suggesting the possibility of a real

biological signal of damaged cells being filtered out during QC. These “noise” cells may provide important information for researchers seeking to exploit the DDR pathway for novel therapeutic targets.

Within the primary cluster, there are two discernible patterns of QC characteristics. The first is having **extremely low to zero percentage of mitochondria and counts of RNA while expressing a moderate number of genes**. The low percentage of mitochondria suggests that these cells likely have intact membranes and are transcriptionally quiet. The low RNA count and moderate number of genes suggests that the cell may be dormant with reduced needs and may be functionally intact. Taken together, these characteristics suggest the cell may be dormant, which have been implicated in chemotherapy resistance and recurrence of cancer.

The second pattern is having a **moderate percentage of mitochondria and counts of RNA while expressing a high number of genes**.The moderate percentage of mitochondria and rna counts suggest the cells are healthy and active. The high number of genes suggest a diversity of functions in the cell. These may be healthy viable cells that happen to be part of a complicated system requiring a large number of genes.

We formulated a simple classification model from these findings. Values below the first quartile are considered low, values above the first quartile but below the 3rd quartile are considered moderate, and values above the 3rd quartile are considered high. In total, there are three potential biological signals that may be inadvertently filtered out (Table 7).

Potential Biological Signal	**Count (% of noise)**	**Matching QC Filter**
Damaged cells	214 (4.27%)	● Low number of genes ● High mitochondrial RNA %
Dormant cells	539 (10.11%)	● Low number of genes
Multifunctional cells	115 (2.29%)	● High mitochondrial RNA %

Table 7. Summary of the potential biological signals being filtered out as noise and the reason they were filtered out.

Univariate Manifold Approximation and Projection (UMAP)

While t-SNE has been a valuable first step for analyzing scRNA-seq data due to its strength in highlighting discrete cell-level resolution, it is important to remember that cells originate from complex tissues whose global structure may be lost. In the case of breast tissues, this is evident when the plot is colored by sample id (Figure 9). The local structure is preserved so well that sample specific clusters are visible within larger clusters. For completeness, we replot the figures 5 and 7 with UMAP (Appendix 4) to better assess global structures, and the findings in table 7 remain consistent.

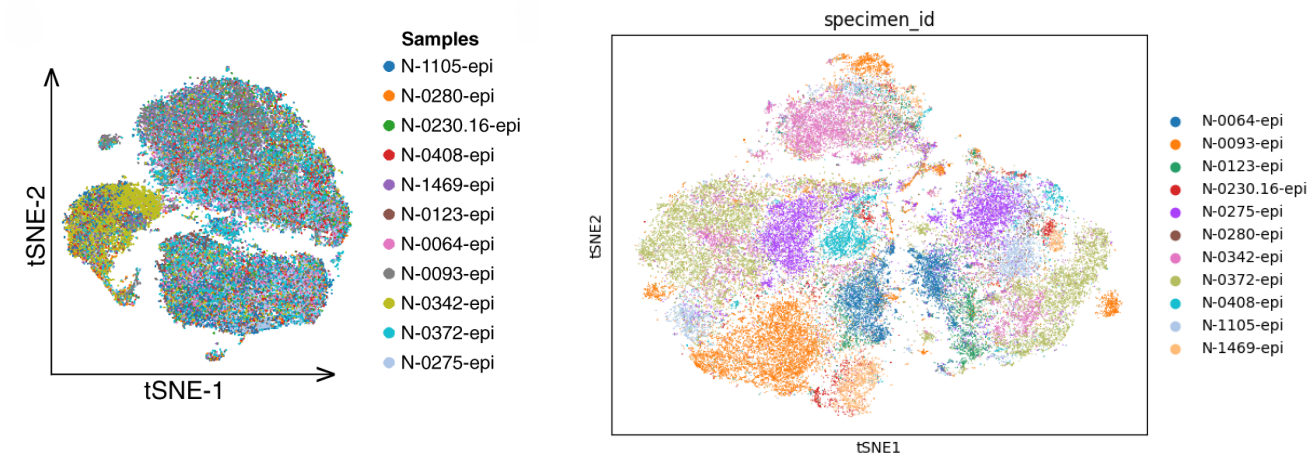


Figure 8. Reference article, figure 1C (left). Reproduced t-SNE (right).

Pathway Associated Gene Expression Analysis

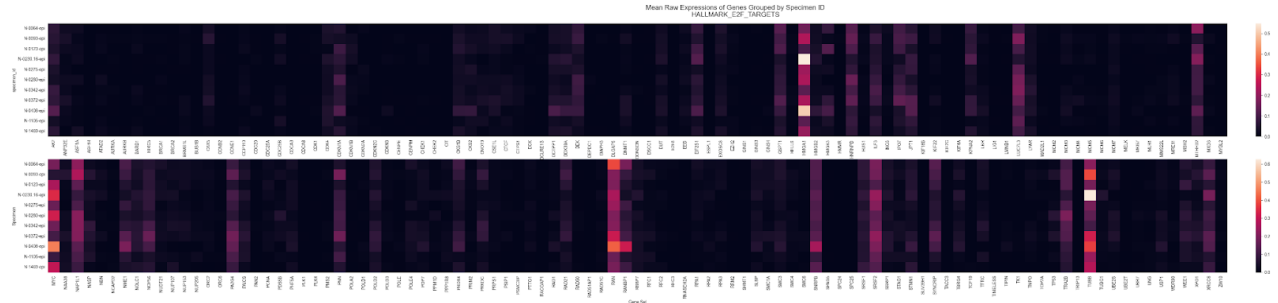
Dimensionality reduction techniques such as t-SNE and UMAP are powerful tools for identifying cells with similar overall expression profiles, without relying on prior biological knowledge. These methods reveal which cells may be similar, but not why they are similar. Pathway associated gene expression analysis addresses this “why” by visualizing the expression patterns of predefined sets of genes associated with specific biological pathways, providing insight into the functional processes that may be driving the observed similarities. Our earlier findings suggest that dormant and/or damaged[13] cells (not to be confused with ambient RNA) exist in the “noise”. To explore this, we selected pathways that are representative of these states and calculated the mean log-normalized expression of the gene signature across all cells for each specimen. If the “noise” cells are truly technical artifacts, we would expect the resulting heatmap to show no discernible patterns across specimens.

Currently, there are no widely accepted gene signatures for cell dormancy. Instead, we select pathways implicated in tumor progression, representing the opposing cell state of interest. The Hallmark E2F Targets gene set regulates transcription factors that are critical to cell cycle regulation, while the Hallmark G2M Checkpoint gene set plays a vital role in mitosis[14] [7]. As a proxy for cell dormancy, we assessed these pathways for minimal expression.

The heatmap for these pathways in “noise” cells (Figure 9) reveal a subtle pattern across specimens, with most genes showing minimal expression (shades of purple and black) and only a small subset exhibiting higher expression. While the absence of pathway activation does not confirm dormancy, it also does not rule it out. The purple striations could reflect true ambient RNA containing target genes, averaging to low overall expression.

[13] Multifunctional cells are excluded from this analysis as there are no established gene sets.
[14] **Mitosis** is a part of the cell cycle where it divides itself into two.

a



b

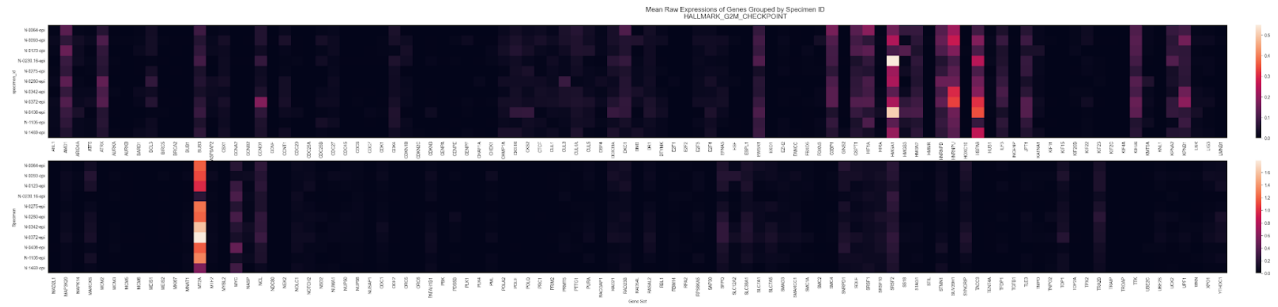
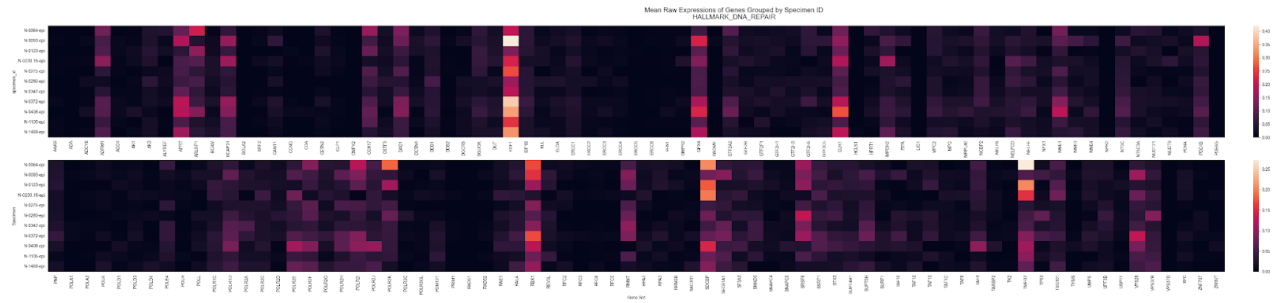


Figure 9. Mean log-normalized expression of pathway associated genes in the “noise” cells. Each signature is split into two rows. Genes are on the x-axis; specimens are on the y-axis.
(a) Hallmark E2F Targets with expression ranges of 0.00 to 0.60 (top) and 0.00 to 0.65 (bottom).
(b) Hallmark G2M Checkpoint with expression ranges of 0.00 to 0.60 (top) and 0.00 to 1.80 (bottom).

However, in the heatmaps for “real” cells (Appendix 5), there is a similar expression pattern, even though the log-normalized expression values for “noise” cells are, on average, four times lower. If “noise” cells were predominantly composed of ambient RNA, this would indicate that “real” cells, regardless of biological functions, are lysing[15] at the same rate to leave this shadowing effect.

For damaged cells, we select the pathways for the Hallmark DNA Repair, a gene set focused on genes that participate in repairing damaged DNA [7], and the Hallmark P53 Pathway, containing genes that are activated by damage and other stresses. In this case, we assess the pathways for significant expressions. Once again, there is a similar expression pattern between the “noise” cells (Figure 10) and “real” cells (Appendix 5).

a



[15] **Lysing** is the process of breaking down the cell membrane.

b

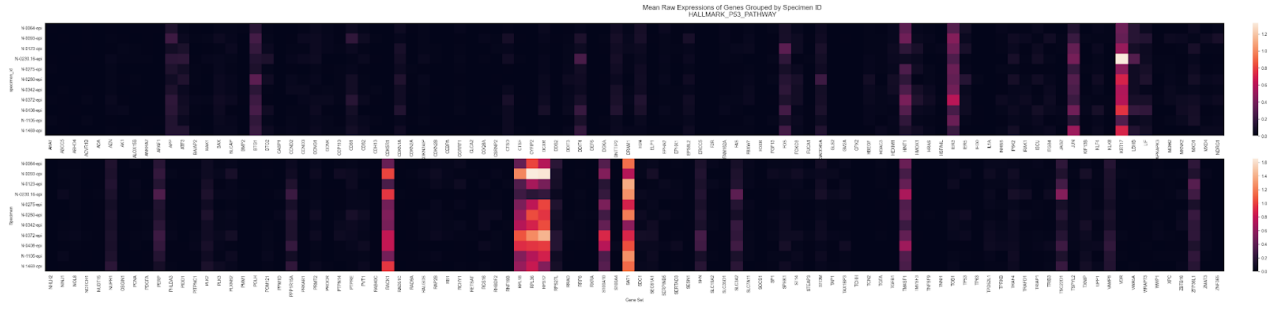


Figure 10. Mean log-normalized expression of pathway associated genes in the “noise” cells. Each signature is split into two rows. Genes are on the x-axis; specimens are on the y-axis.
(a) Hallmark DNA Repair with expression range of 0.00 to 0.42 (top) and 0.0 to 0.27 (bottom).
(b) Hallmark P53 Pathway with expression range of 0.00 to 1.40 (top) and 0.00 to 1.60 (bottom).

Discussion

We aimed to provide insights on the causal ambiguity surrounding the selection of QC thresholds in scRNA-seq pre-processing workflow. While our findings neither confirm nor refute the claim that current QC practices systematically exclude potentially informative cellular states, they do offer additional insights into the profiles of technical artifacts.

When selecting QC thresholds, both the sample origin (tissue type) and the intended purpose of the scRNA-seq analysis (cell state of interest) should be considered, as each influences which metrics are thresholded and how, particularly for cell size and mitochondrial content. Of these thresholds, the lower bound for the minimum number of genes per cell warrants special caution. In our dataset, approximately 14% of potential biological signals in the “noise” cells were flagged for a low gene count. Of the potential signals found, it is possible the multifunctional cells are true technical artifacts. Specifically, they could be two (or more) cells that are mistakenly recorded as a single cell.

The striking similarity in the pathway associated gene expression analyses between the “real” and “noise” cells suggests that scRNA-seq data may follow a bimodal distribution (Appendix 6). If a bimodal distribution does exist, downstream analysis and modeling of scRNA-seq data could benefit from being applied to both populations.

Implications

In the context of cancer research, these potential signals could offer valuable insights into tumor biology. Characterizing cells in their early, dormant states may help inform the factors that drive malignancy, advancing therapeutic discoveries. Alternatively, quantifying the proportion of “noise” cells could serve as a novel diagnostic indicator.

These potential signals may reflect the terminal stages of a cell’s life cycle, given the expression similarities across multiple pathways. The potential dormant cells may represent newly divided cells, which are expected to contain fewer genes than their mature counterparts. In contrast, the potential

damaged cells may be those that have naturally come to the end of their life cycle, rather than an indication of a rare unknown subpopulation.

To explore the potential impact of this study, we expanded the identification of potential biological signals to both precancerous and cancerous tissue specimens, focusing on datasets representing total cell populations. Interestingly, if the potential dormant signal indicates early cell stages, ER+ tumor tissue specimens have fewer young cells than normal tissue specimens, despite having twice as many “noise” cells (Figure 11). This finding strengthens the case for deeper exploration of “noise” cells as a potential source of biologically meaningful signals to further cancer research.

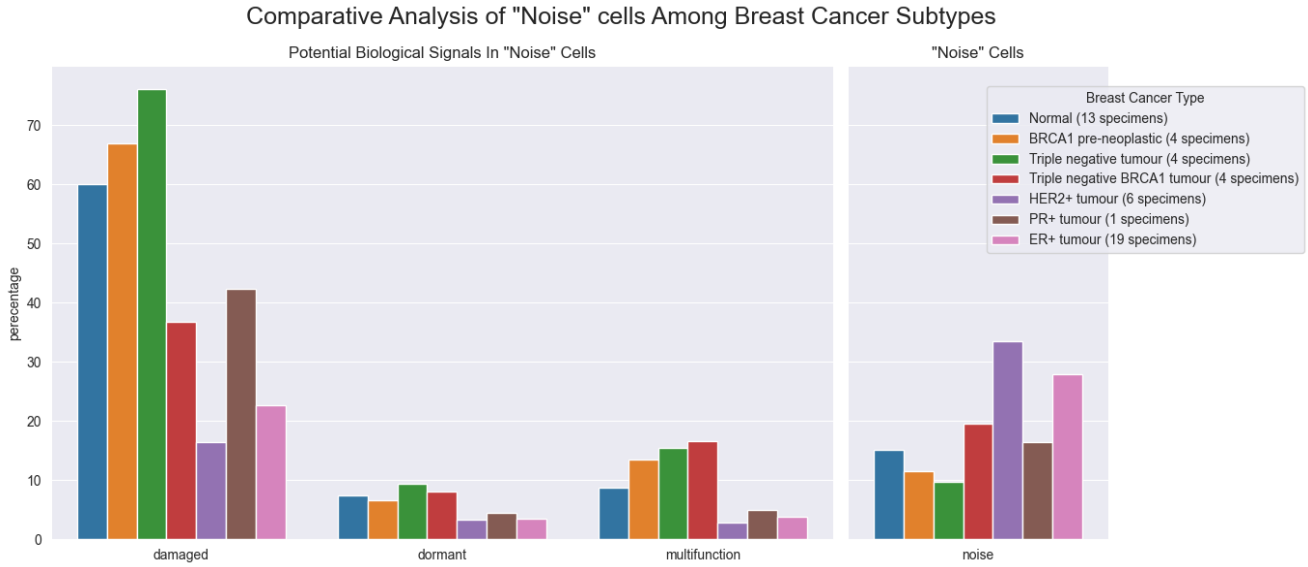


Figure 11. Datasets represent total cell populations, y-axis is in percentage.
(left) Percentage of “noise” cells that are potential biological signals for each cancer type.
(right) Percentage of all cells that are “noise” cells for each cancer type.

Beyond the technical implications of refining QC threshold selection, it is important to recognize that the underlying data carries inherent socioeconomic biases. For example, the dataset reflects patients who had access to healthcare providers and medical facilities to provide the tissue samples. Such biases should be carefully considered when developing potential therapeutic targets and diagnostic tools, to help ensure that marginalized populations are not excluded from the benefits of these advances.

Limitations

This study has some limitations. We are not trained medical researchers and lack the expertise to fully characterize the biological significance in the heatmaps and dimensionality-reduction plots. Our work represents an early, exploratory step toward a much deeper biological investigation that would require extensive experimental validation [7]. Additionally, this study cohort focuses on a single population, female mastectomy patients with no family history of breast cancer.

Future Work

Next steps include extending our findings to additional scRNA-seq datasets from other publications to assess robustness and generalizability. We would focus first on additional datasets for breast tissue before further extending to other tissue types. Ideally, partnering with a domain expert will be critical for accurate biological interpretation.

Statement of Work

All work for this project was done by Sophia K. Cheng.

References

1. Lafzi A, Moutinho C, Picelli S, Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nature protocols. London: Nature Publishing Group UK; 2018;13(12):2742–2757.
2. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology. London: Nature Publishing Group UK; 2019;15(6):e8746-n/a.
3. Young, Matthew D, Behjati, Sam. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. United States: Oxford University Press; 2020;9(12).
4. van Gent DC, Kanaar R. Exploiting DNA repair defects for novel cancer therapies. Molecular biology of the cell. United States: The American Society for Cell Biology; 2016;27(14):2145–2148.
5. Lindell, Emma, Zhong, Lei, Zhang, Xiaonan. Quiescent Cancer Cells-A Potential Therapeutic Target to Overcome Tumor Resistance and Relapse. International journal of molecular sciences. Switzerland: MDPI AG; 2023;24(4):3762-.
6. Yeh, Albert C, Ramaswamy, Sridhar. Mechanisms of Cancer Cell Dormancy--Another Hallmark of Cancer? Cancer research (Chicago, Ill). United States; 2015;75(23):5014–5022.
7. Cheng, Sophia K. Signals in the Noise: Uncovering the Biological Signatures of Ghost Cell Profiles in Human Breast Cancer. Dec 2025. Data Science for Social Good, University of Michigan, student paper.
8. Pal, Bhupinder, Chen, Yunshun, Vaillant, François, Capaldo, Bianca D, Joyce, Rachel, Song, Xiaoyu, Bryant, Vanessa L, Penington, Jocelyn S, Di Stefano, Leon, Tubau Ribera, Nina, Wilcox, Stephen, Mann, Gregory B, Papenfuss, Anthony T, Lindeman, Geoffrey J, Smyth, Gordon K, Visvader, Jane E. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO journal. London: Nature Publishing Group UK; 2021;40(11):e107333-n/a.
9. Chen Y, Pal B, Lindeman GJ, Visvader JE, Smyth GK. R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue. Scientific data. London: Nature Publishing Group UK; 2022;9(1):96–9.
10. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ, kConFab. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature medicine. New York: Nature Publishing Group US; 2009;15(8):907–913.
11. Liberzon, Arthur, Birger, Chet, Thorvaldsdóttir, Helga, Ghandi, Mahmoud, Mesirov, Jill P., Tamayo, Pablo. The Molecular Signatures Database Hallmark Gene Set Collection. Cell systems. United States: Elsevier Inc; 2015;1(6):417–425.
12. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Access and store annotated data matrices. Journal of open source software. 2024;9(101):4371-.
13. Samocha A, Doh H, Kessenbrock K, Roose JP. Unraveling Heterogeneity in Epithelial Cell Fates of the Mammary Gland and Breast Cancer. Cancers. Switzerland: MDPI AG; 2019;11(10):1423-.

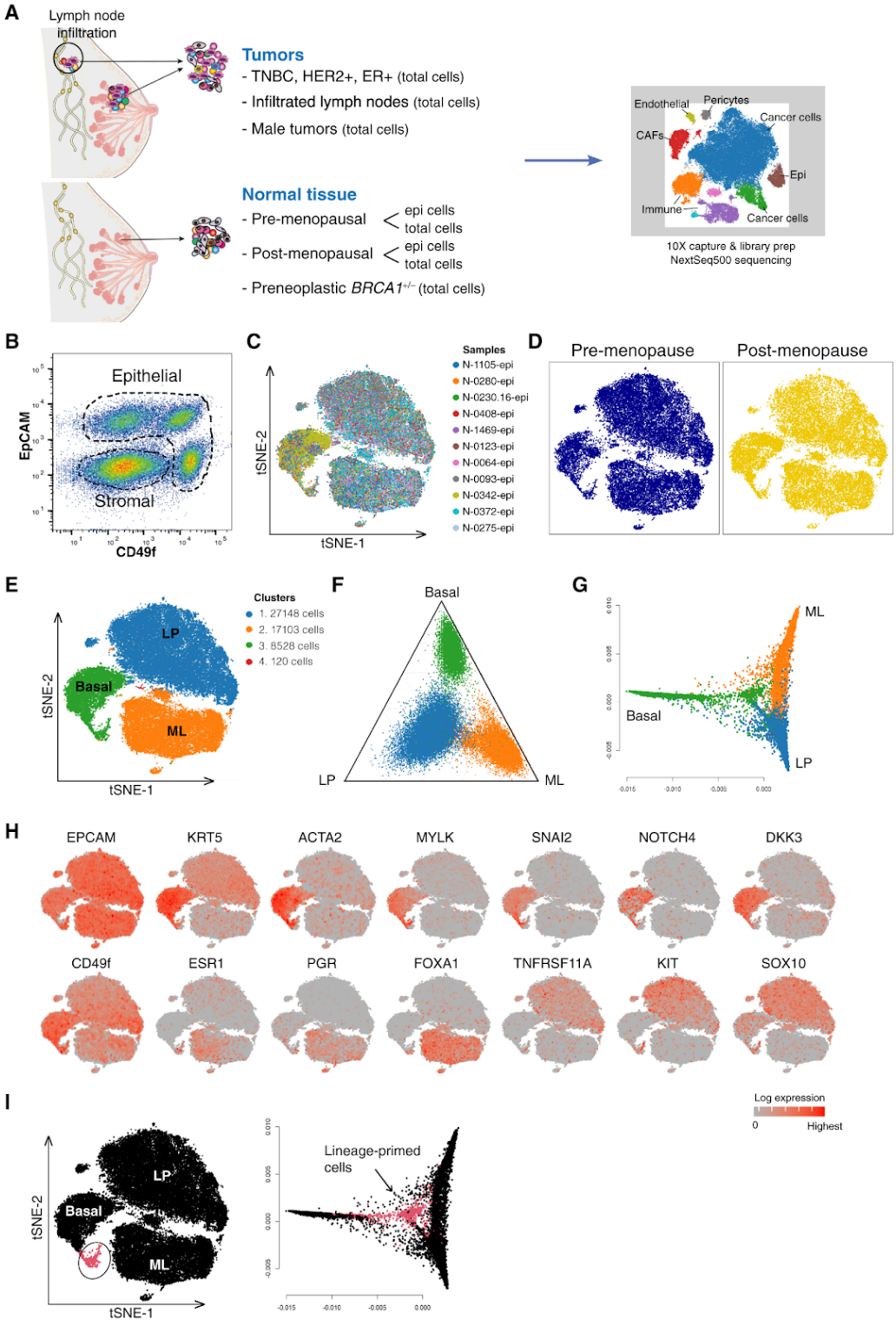
14. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. New York: Nature Publishing Group US; 2015;33(5):495–502.

Sophia K. Cheng
Team 13 - Signal Seekers
SIADS 699 - Spring/Summer 2025

Appendices

1. Baseline validation reference figure

“Figure 1. Workflow for the breast atlas and scRNA-seq profiling of normal breast epithelium” [8]



2. Tissue samples for profiling normal breast epithelium

Using supplementary table 1 from the reference article, we can map the samples used in Figure 1 of the reference article [8] to their corresponding raw filenames, GEO IDs, and cached anndata filename.

Sample ID	**Raw Filename**	**GEO ID**	**anndata Filename**
N-1105-epi	N-N1105-Epi-*.*.gz	GSM4909260	GSM4909260_N-N1105-Epi.h5ad
N-0280-epi	N-N280-Epi-*.*.gz	GSM4909255	GSM4909255_N-N280-Epi.h5ad
N-0230.16-epi	N-N1B-Epi-*.*.gz	GSM4909264	GSM4909264_N-N1B-Epi.h5ad
N-0408-epi	N-NE-Epi-*.*.gz	GSM4909259	GSM4909259_N-NE-Epi.h5ad
N-1469-epi	N-NF-Epi-*.*.gz	GSM4909258	GSM4909258_N-NF-Epi.h5ad
N-0123-epi	N-MH0023-Epi-*.*.gz	GSM4909267	GSM4909267_N-MH0023-Epi.h5ad
N-0064-epi	N-MH0064-Epi-*.*.gz	GSM4909262	GSM4909262_N-MH0064-Epi.h5ad
N-0093-epi	N-PM0095-Epi-*.*.gz	GSM4909256	GSM4909256_N-PM0095-Epi.h5ad
N-0342-epi	N-PM0342-Epi-*.*.gz	GSM4909269	GSM4909269_N-PM0342-Epi.h5ad
N-0372-epi	N-PM0372-Epi-*.*.gz	GSM4909275	GSM4909275_N-PM0372-Epi.h5ad
N-0275-epi	N-MH275-Epi-*.*.gz	GSM4909273	GSM4909273_N-MH275-Epi.h5ad

3. Clustering scRNA-seq data

We use scanpy[16,17] to cluster the single-cell gene expression data, using their basic tutorial on clustering as a guide.

1. Feature selection
    a. Reduce the dimensionality of the data by only including the most informative genes.
    b. This first step is necessary to prevent the principal component analysis (PCA) from being dominated by uninformative genes.
    c. There are 33,538 genes for each cell before selecting for highly variable genes (HVG). The large number is due in part to how the data was stored, where all tissue samples share a single file for the features.
    d. After selecting for HVG, there are under 300 genes remaining.
2. Dimensionality reduction
    a. Use PCA to further reduce the dimensionality so that the nearest neighbor calculations are more stable, as well as reducing the computational resources needed.
3. Nearest neighbor graph construction and visualization
    a. Computed the k-nearest neighbors weighted graph for the clustering algorithm
4. Clustering
    a. Use the leiden algorithm to determine clusters.

[16] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology. England: BioMed Central; 2018;19(1):15–15.
[17] https://scanpy.readthedocs.io/en/stable/

4. UMAP Visualizations

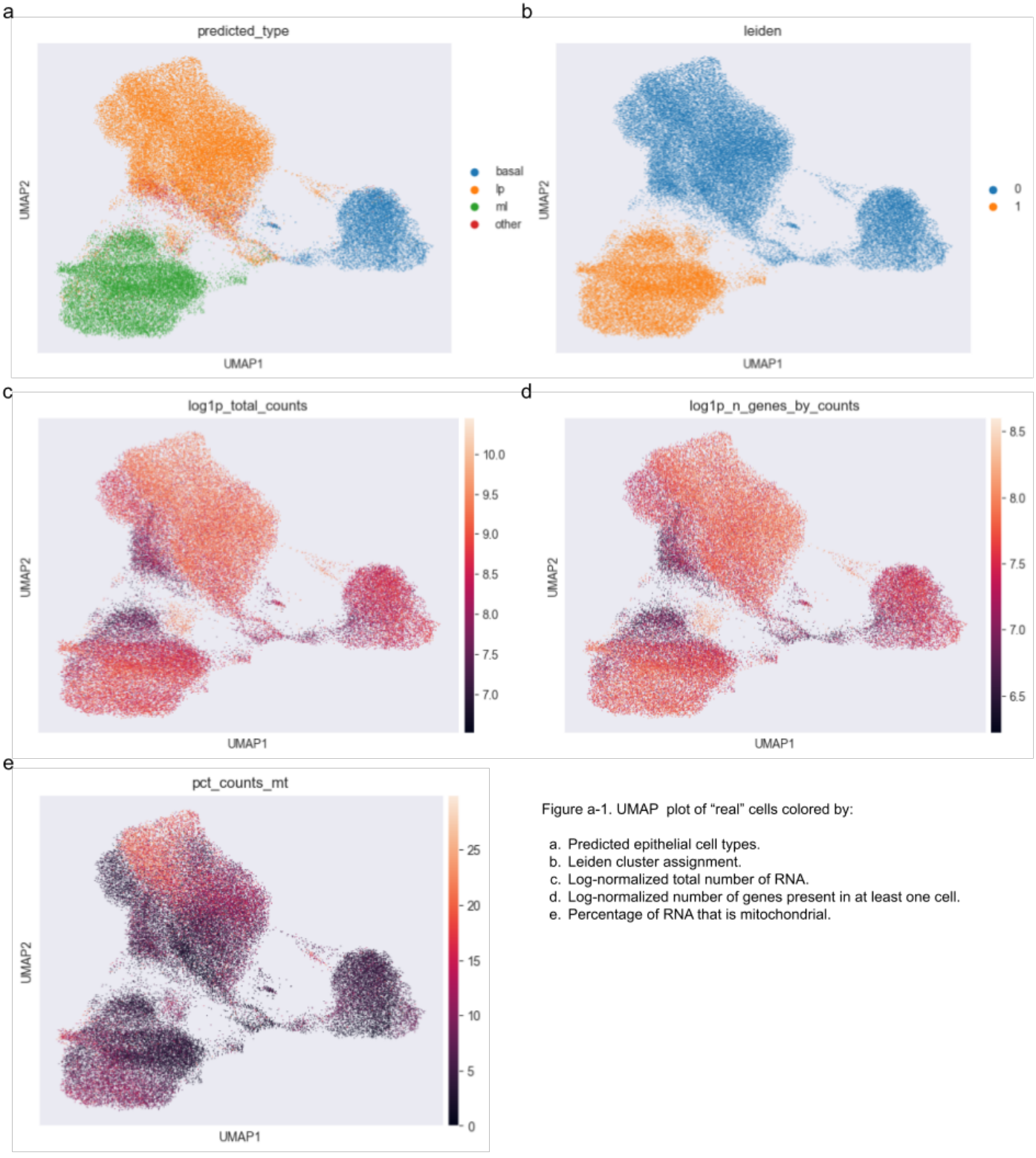


Figure a-1. UMAP plot of "real" cells colored by:

a. Predicted epithelial cell types.
b. Leiden cluster assignment.
c. Log-normalized total number of RNA.
d. Log-normalized number of genes present in at least one cell.
e. Percentage of RNA that is mitochondrial.

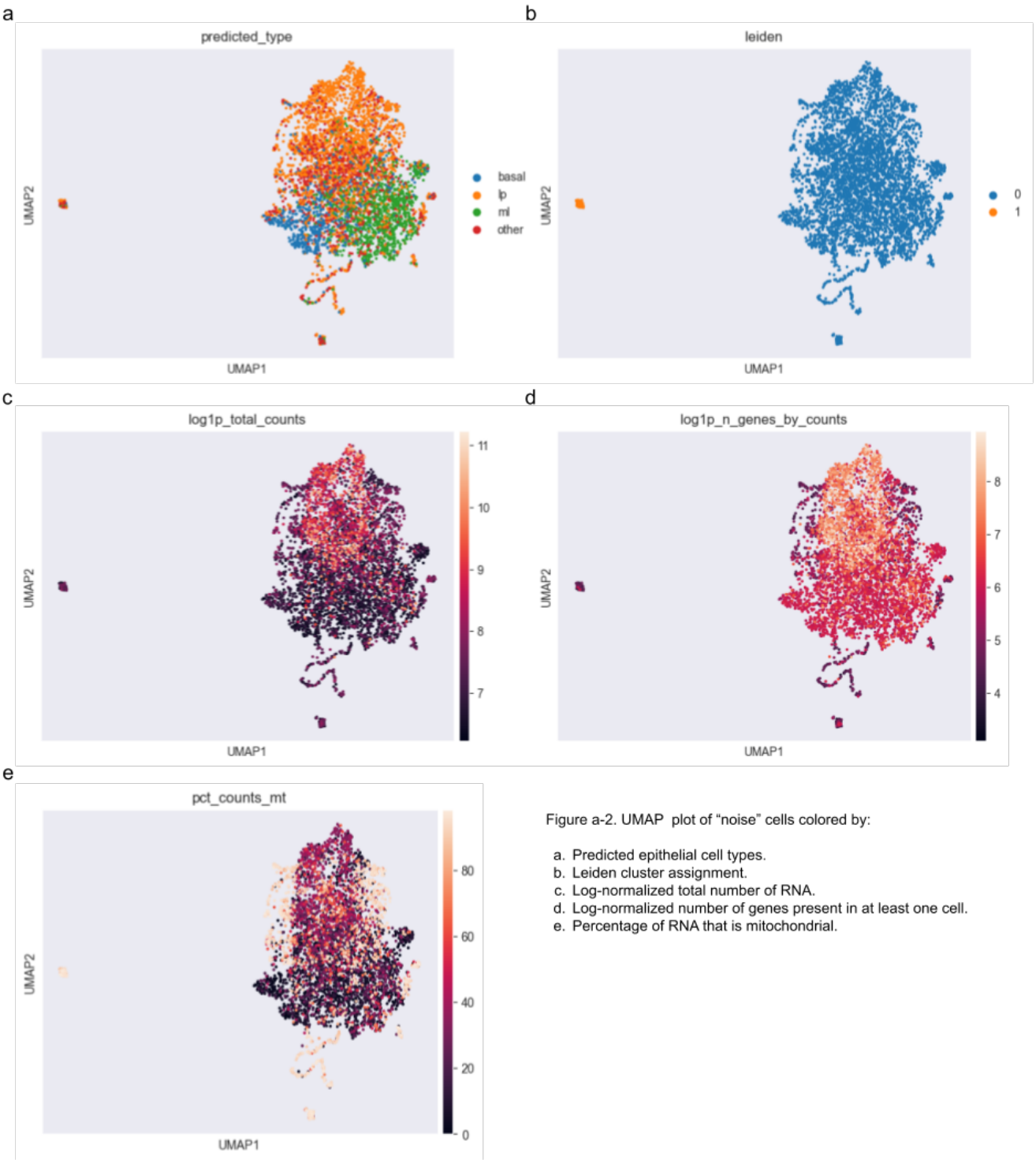
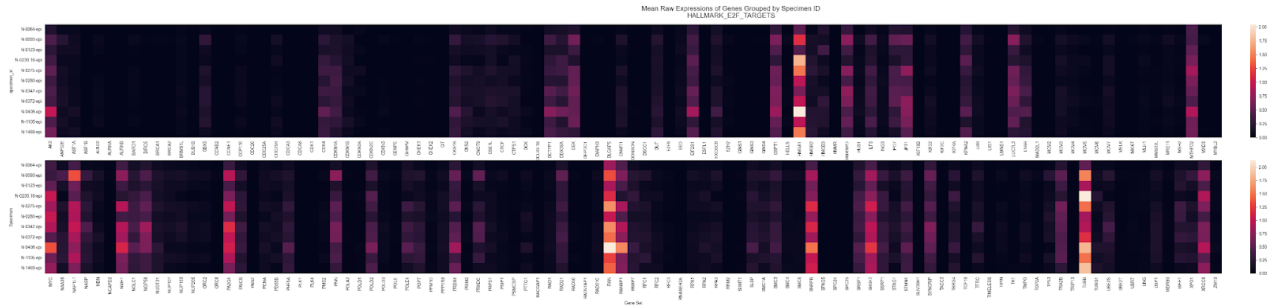


Figure a-2. UMAP plot of "noise" cells colored by:

a. Predicted epithelial cell types.
b. Leiden cluster assignment.
c. Log-normalized total number of RNA.
d. Log-normalized number of genes present in at least one cell.
e. Percentage of RNA that is mitochondrial.

5. Pathway Associated Gene Signature Analysis

a



b

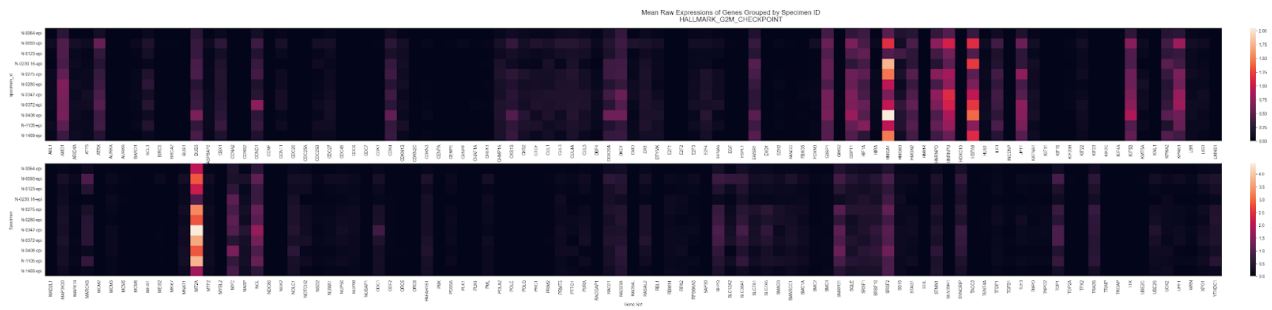
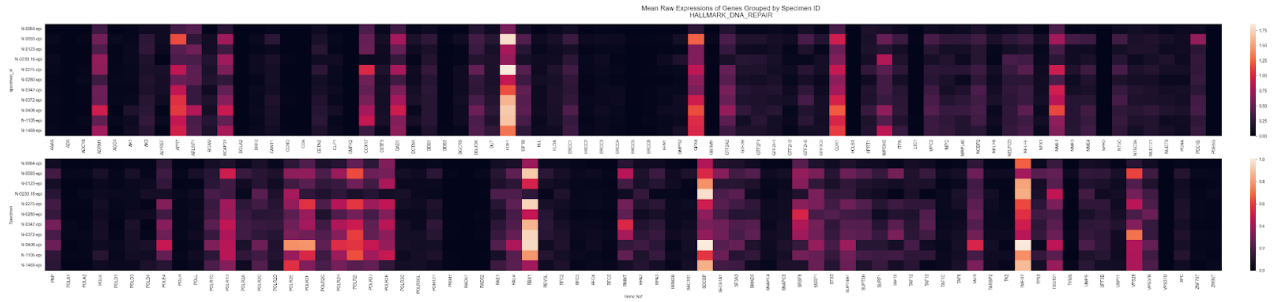


Figure a-3. Mean log-normalized expression of pathway associated genes in the “real” cells. Each signature is split into two rows. Genes are on the x-axis; specimens are on the y-axis.
(a) Hallmark E2F Targets with expression ranges of 0.00 to 2.00 (top) and 0.00 to 2.15 (bottom).
(b) Hallmark G2M Checkpoint with expression ranges of 0.00 to 2.00 (top) and 0.00 to 4.50 (bottom).

a



b

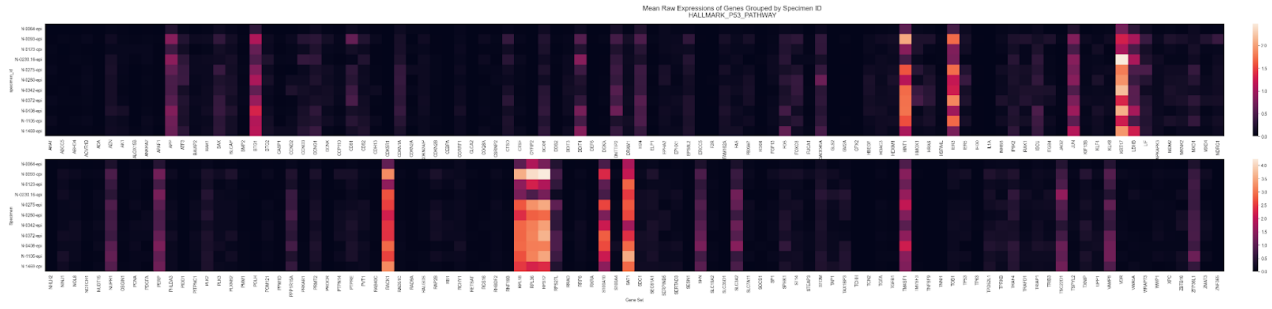


Figure a-4. Mean log-normalized expression of pathway associated genes in the “real” cells. Each signature is split into two rows. Genes are on the x-axis; specimens are on the y-axis.
(a) Hallmark DNA Repair with expression ranges of 0.00 to 1.80 (top) and 0.00 to 1.00 (bottom).
(b) Hallmark P53 Pathway with expression ranges of 0.00 to 2.50 (top) and 0.00 to 4.25 (bottom).

6. Cell typing gene expression distribution scores

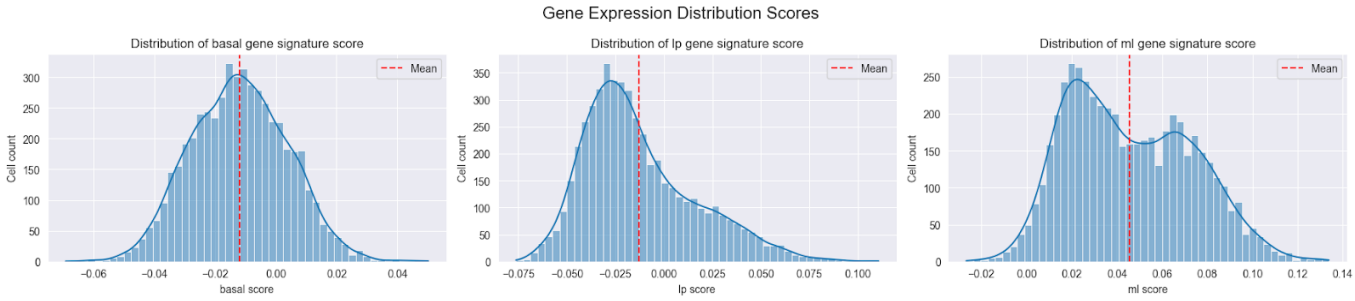


Figure a-5. Distribution of log-normalized gene expression scores for an arbitrary specimen in the dataset.
(left) Basal gene signature score.
(middle) Luminal progenitor gene signature score.
(right) Mature luminal gene signature score.

Figure a-6. Distribution of log-normalized gene expression scores for an arbitrary specimen in the dataset, filtered down to highly variable genes.
(left) Basal gene signature score.
(middle) Luminal progenitor gene signature score.
(right) Mature luminal gene signature score.