Signals in the Noise: Uncovering the Biological Signatures of Ghost Cell Profiles in Human Breast Cancer

I. Introduction (308 words)

Understanding the biological significance of low RNA profile cells in tumors has lagged behind advances in single-cell RNA sequencing (scRNA-seq) [1]. While technologies have enabled detailed characterization of tumor ecosystems [1,2], cells exhibiting low overall RNA counts have often been treated as technical artifacts and excluded from analyses [3]. This approach risks overlooking rare or quiescent cell states that may hold critical biological importance, particularly in understanding cancer dormancy and resistance [4, 5].

Emerging evidence suggests that transcriptionally quiet cells, sometimes dismissed as “ghost cells”, [2] might represent viable, functionally distinct populations rather than debris or doublets [3]. Although standard quality control pipelines aim to minimize ambient RNA contamination [4], they can inadvertently filter out cells that are biologically meaningful [3]. In the context of breast cancer, where tumor heterogeneity[5] plays a role in treatment failure and relapse, failing to account for such populations could obscure important dynamics of progression and resistance [6, 7].

Existing research on low RNA profile cells has largely been descriptive, lacking formal causal inference approaches to directly test their functional roles or impact on tumor behavior. While causal inference methods are increasingly being advocated for in genomic research, they have rarely been applied in single-cell analyses at this granularity [8]. Thus, this study aims to develop causal inference strategies to assess the biological relevance of ghost-like cell profiles in human breast tumors.

Specifically, we[6] hypothesize that low RNA profile cells are not merely technical noise but instead represent distinct biological states that contribute to tumor heterogeneity and potentially influence disease progression. By combining proxy measures for “ghost-like” characteristics with robust causal inference strategies, we seek to uncover signals obscured in traditional scRNA-seq pipelines. This work addresses a critical gap by reframing ghost cells from artifacts to subjects of biological inquiry, offering a foundation for deeper understanding of human breast cancer mechanisms and providing novel therapeutic insights.

II. Literature review (333 words)

Advances in scRNA-seq have transformed the ability to resolve cellular heterogeneity within tumors through analysis of different cell states, revealing interactions that shape the tumor microenvironment [1, 2]. Despite these strides, analytical pipelines have often prioritized high DNA content cells, treating cells with low total RNA[7] counts as technical artifacts to be excluded from downstream analyses [3]. This approach, while effective for minimizing noise from ambient RNA contamination, risks eliminating biologically meaningful rare or quiescent populations.

Recent evidence suggests that transcriptionally quiet or “ghost-like” cells may represent viable functionally distinct cell states rather than debris or damaged cells [4, 5]. In the context of cancer, such 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 transcriptional profiles often overlap with cells flagged for exclusion by conventional scRNA-seq quality control [4, 5]. In breast cancer specifically, tumor heterogeneity and the persistence of minimal residual disease represent major clinical challenges [6, 7]. Thus, refining our approaches to capture and study low RNA profile cells could offer new insights into the mechanisms driving disease progression and persistence.

While descriptive studies have characterized aspects of low RNA cell populations, few have attempted to formally assess their functional relevance using causal inference techniques. Causal reasoning approaches have gained traction in genomics for disentangling complex biological relationships, but their application at the single-cell level, particularly at the granularity of rare or transcriptionally quiet states, remains limited [8]. Integrating causal inference methods with single-cell data holds promise for moving beyond correlative observations toward mechanistic (causal) insights.

Building upon these gaps, we propose a structured analysis plan to evaluate the biological significance of ghost cells within human breast tumors. This literature review highlights the importance of methodological rigor in inferring causality from observational single-cell data. By integrating robust statistical approaches with biologically informed feature engineering, our analysis aims to provide new insights into ghost cells' role in breast cancer biology.

III. Analysis Plan

Hypotheses

Hypothesis 1

- **Null hypothesis (H_0):** Ghost cells do not have a causal relationship with immune evasion in breast cancer.
- **Alternative hypothesis (H_1):** Ghost cells have a causal relationship with immune evasion in breast cancer

One of the more successful cancer treatments has been immune checkpoint inhibitors [9]. This treatment works by preventing cancer cells from sending false “checkpoint” signals to masquerade as a healthy cell. Specifically, as the name suggests, it inhibits the immune checkpoints, effectively unmasking the cancer cell. Given the effectiveness of targeting immune evasion, we investigate whether a causal relationship exists between ghost cells and immune evasion. We hope to provide some additional insight into developing strategies for breast cancer immunotherapies.

Hypothesis 2

- **Null hypothesis (H_0):** Ghost cells do not have a causal relationship with tumor proliferation in breast cancer.
- **Alternative hypothesis (H_1):** Ghost cells do have a causal relationship with tumor proliferation in breast cancer.

With the continued advancement of single-cell RNA sequencing, researchers have identified quiescent cancer cells (QCC) as important therapeutic targets. In a 2023 review [4] of strategies that target the dormant cancer cells, one of the challenges to the therapy is a lack of reliable detection for these low activity cell populations. While dormant cancer cells and ghost cells are different, they do share many similarities such as low activity and resistance to therapy. We investigate whether a causal relationship exists between ghost cells and tumor proliferation. We hope to provide some additional insights that can be applied to targeting QCC.

Hypothesis 3

- **Null hypothesis (H_0):** Ghost cells do not have a causal relationship with chemotherapy resistance in breast cancer.
- **Alternative hypothesis (H_1):** Ghost cells do have a causal relationship with chemotherapy resistance in breast cancer.

Studies have shown drug resistance is a significant challenge when treating metastatic breast cancer, “One of the main clinical issues is the development of drug resistance, which accounts for failure of treatment leading to death in more than 90% of patients with MBC and affects all classes of agents…” [6]. Further, a 2020 mini-review [7] indicates that cancer stem cells (CSC) may play a role in

chemotherapy resistance in general. Given that CSC also shares similar properties with ghost cells as QCC does, we investigate whether a causal relationship exists between ghost cells and chemotherapy resistance.

Data Description

We will be combining two openly available data sources to conduct this analysis.

Source Data - Samples

For this analysis, we will be looking at single-cell RNA expression data for normal, precancerous, and cancerous human breast tissue. In addition to a diversity of cancerous states, this dataset also exhibits a diverse set of cell types and clusters.

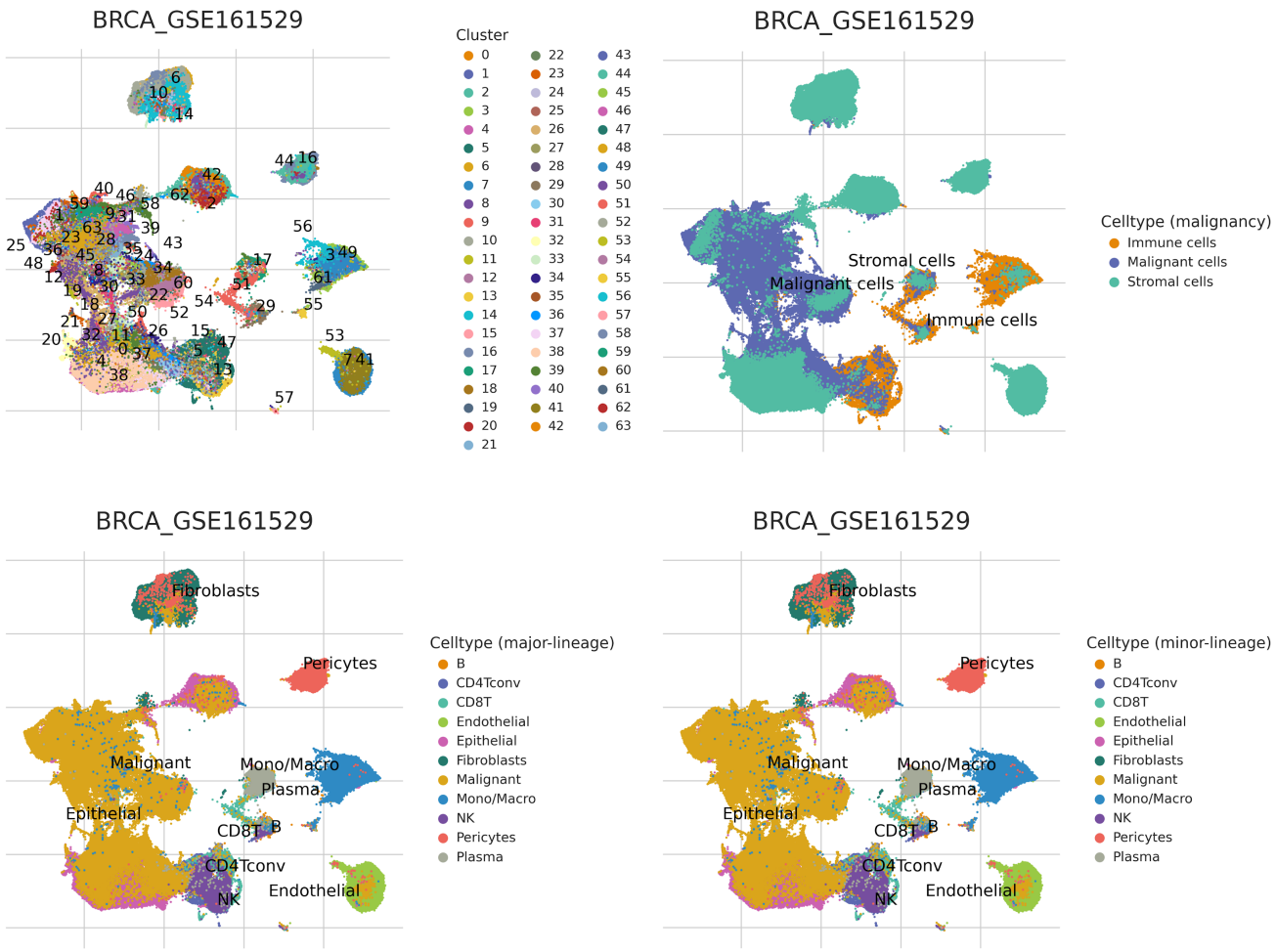

Fig 1. Images downloaded from TISCH2's web interface for browsing datasets showing the diversity of data in (top-left) clustering; (top-right) cell type by malignancy; (bottom-left) cell type by major-lineage; (bottom-right) cell type by minor-lineage;

The dataset we will be using can be downloaded from the Tumor Immune Single-cell Hub 2 (TISCH2) [1, 2].

Dataset name	BRCA_GSE161529 [23]
Number of patients	52
Number of cells	332,168 (our sample)
Time period	Unspecified, likely between 2020 - 2025 (dates of the National Breast Cancer Foundation grant)

Table 1. Details of the dataset. The samples for this study will be the cells.

Source Data - Supporting

In addition to our source samples, we will be engineering features for our dependent variables. Specifically, we will download hallmark gene sets from the Molecular Signature Database (MSigDB) [11]. Hallmark gene sets differ from the other gene sets offered by MSigDB because of the focus on gene expression change, improved signal-to-noise ratio and manual annotations. This makes them more interpretable and statistically robust for analyzing the results of hypothesis 2 and 3. We list our selected gene sets below, highlighting notable genes within the context of our analysis.

HALLMARK_G2M_CHECKPOINT

This gene set is focused on the genes that participate in the G2/M cell cycle checkpoint. This set of genes is highly relevant because checkpoints play a vital role in mitosis. The dysregulation of mitosis is known to cause cancers. The notable genes in this set are CDC25C; CDK1; PLK1; AURKA; WEE1. Kai Liu, et al. [12], discusses the role of these genes in cell cycle regulation.

HALLMARK_E2F_TARGETS

This gene set is focused on the genes that regulate E2F transcription factors, which are critical to the regulation of the cell cycle. When this cycle is disrupted, cells can begin to accumulate mutations, leading to cancer. The notable genes in this set are CDK1; CCNA2; E2F1; E2F2. Zeyu Xing, et. al., [14] and Yongbin Lu, et al., [13] show the association of CDK1, CCNA2, and E2F1 with breast cancer.

HALLMARK_DNA_REPAIR

This gene set is focused on genes that participate in repairing damaged DNA, which are potential targets to address chemotherapy resistance. A recent review article [16] discusses the impact of DNA damage tolerance on responding to chemotherapy. Specifically that an overexpression of these genes may be negating the effects of chemotherapy by repairing the cancerous cells. The notable genes in this set are BRCA1/BRCA2; RAD51; ATM; CHEK2; XRCC1. Elaine Gimore, et al., [15] provide a good summary of why these genes are notable for DNA repair.

HALLMARK_UNFOLDED_PROTEIN_RESPONSE

This gene set is focused on genes that participate in responding to misfolded and/or unfolded proteins in the cells. Similar to the overexpression of DNA repair genes,an abundance of unfolded protein response (UPR) genes negates the effects of chemotherapy by removing the drug before they've had a chance to work. The notable genes in this set are ATF6; XBP1; CHOP; GRP78 (BiP).; PERK.

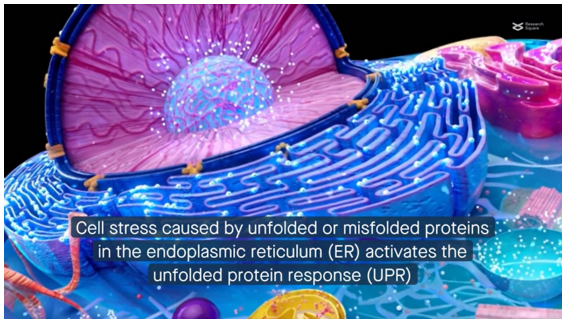

Fig 2. Screenshot of an animation to illustrate how UPR relates to cancer, as well as discussing the role of notable genes (https://pmc.ncbi.nlm.nih.gov/articles/PMC10832166/figure/MOESM1/) [17].

Dependent Variables

The dependent variables in this analysis are engineered to represent the biological outcome of each hypothesis and a placebo outcome.

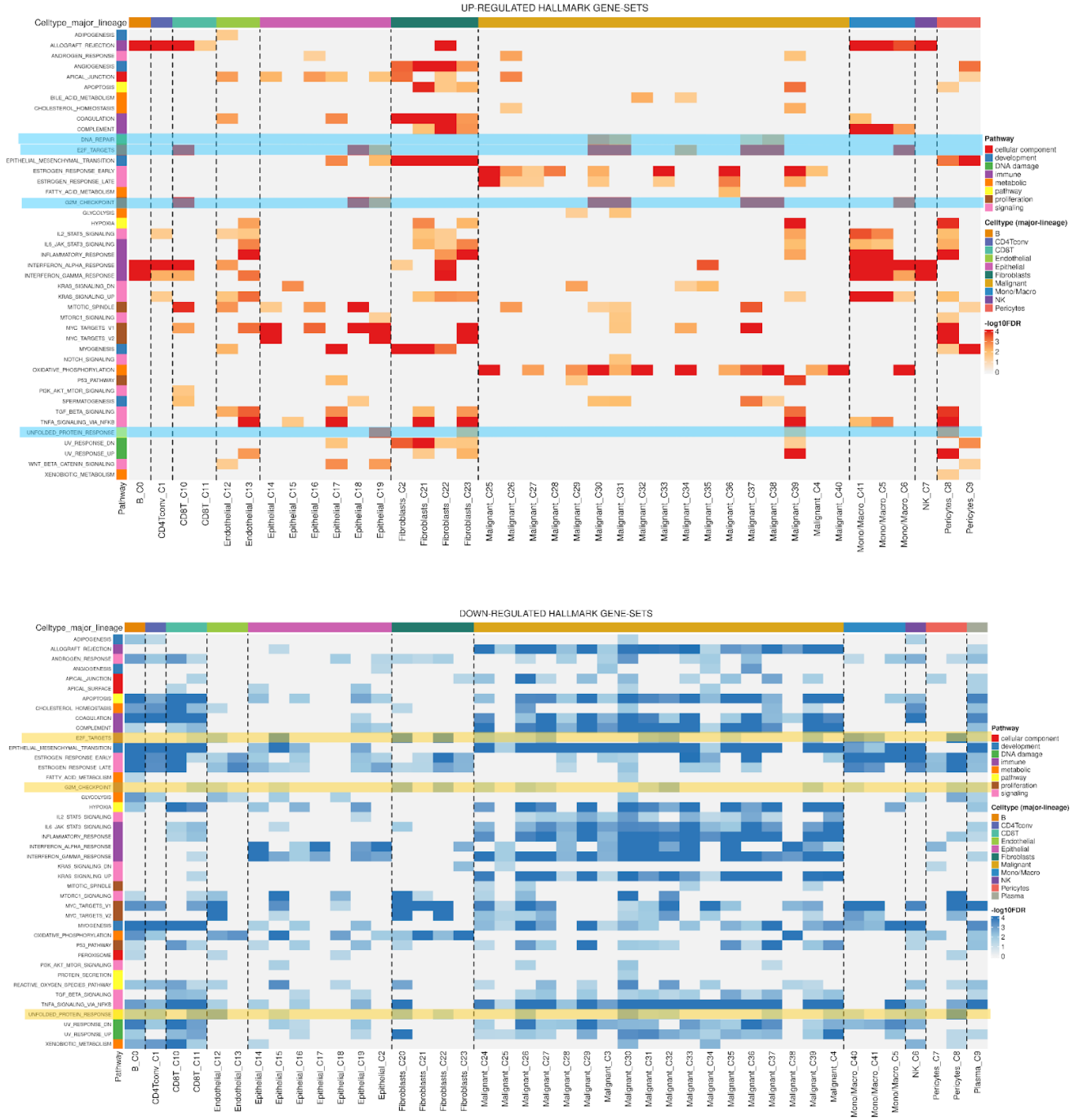

Fig 3. (Top) Up regulated expression of hallmark gene sets. Sets selected for this analysis are highlighted in blue. (Bottom) Down regulated expression of hallmark gene sets. Sets selected for this analysis are highlighted in yellow. Note that in this visualization, there is no down regulation of the DNA repair gene set. Images downloaded from TISCH2's web interface.

Hypothesis 1 - Immune Evasion

Given that there is no single hallmark gene set designated for studying immune response, an indication that the immune response is not as well characterized, we will instead use the expression of the CD274 gene as our dependent variable. CD274 is the gene that encodes the programmed death ligand-1 (PD-L1) [10], which has been studied extensively for immune checkpoint inhibitor (ICI) treatments.

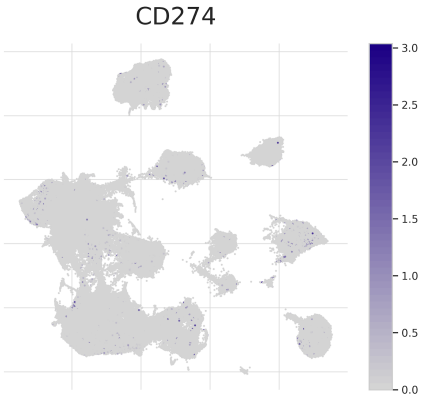

Fig 4. Expression of CD274 in the source samples. Images downloaded from TISCH2's web interface.

Hypothesis 2 - Tumor Proliferation

Using the HALLMARK_G2M_CHECKPOINT and HALLMARK_E2F_TARGETS gene sets, we will calculate proliferation scores using gene set enrichment analysis.

Hypothesis 3 - Chemotherapy Resistance

Using the HALLMARK_DNA_REPAIR and HALLMARK_UNFOLDED_PROTEIN_RESPONSE gene sets, we will calculate resistance scores using gene set enrichment analysis.

Robustness Check - Placebo

Using the BRCA_GSE161529 dataset, we will calculate the expression of HPRT [20] for use as a placebo outcome in our robustness check.

Dependent Variable	**Hypothesis**
cd274_expr	(1) Immune Evasion
g2m_score	(2) Tumor Proliferation
e2f_score	(2) Tumor Proliferation
dr_score	(3) Chemotherapy Resistance
upr_score	(3) Chemotherapy Resistance

hprt_expr	n/a (placebo outcome)
Table 2. Dependent variables.	

Independent Variables

The independent variables in this analysis are engineered features that classify a cell's "ghost-ness". The ghost cells would normally be filtered out from analysis during the quality control process as ambient RNA (https://www.10xgenomics.com/analysis-guides/introduction-to-ambient-rna-correction). We will engineer continuous features representing the likelihood of the cell's "ghost-ness" using the raw count matrices of our source sample. We will have a total of 3 independent binary variables (see Table 3) whose value will be 1 (e.g. "treated") if the corresponding score is above a given threshold $TREATMENT_{THRESHOLD}$.

Independent Variable	**Score (Feature)**
doublet	$DOUBLET_{SCORE}$
mitochondrial	$MITOCHONDRIAL_{SCORE}$
ambient	$\left(DOUBLET_{SCORE} + MITOCHONDRIAL_{SCORE}\right) \div 2$
Table 3. Independent variables.	

For this analysis, we will use thresholds *lower* than what is recommended in literature for both (see Table 4), allowing us to intentionally bias our selection toward cells that naturally exhibit low RNA profiles, consistent with the expected behavior of ghost cells. By relaxing the cutoff, we increase the likelihood of capturing biologically relevant low-activity cells rather than filtering them out during preprocessing.

Feature	**Threshold**	**% Lower Than Recommended**
$DOUBLET_{SCORE}$	0.15	40% (0.25)
$MITOCHONDRIAL_{SCORE}$	0.15	50% (0.15 - 0.25)
Table 4. (First column) Engineered feature; (Second column) Threshold to be used; (Third column) percentage lower than recommended threshold which is shown in parenthesis.		

Omitted Variable Bias

One important consideration in this study is omitted variable bias, which can result in misleading causal inferences and false relationships. Omitted variable bias occurs when variables that are associated to both our independent and dependent variables are left out of the analysis. In our study,

there are several known confounders that we need to account for to mitigate our risk of omitted variable bias. We now list the main sources of our confounding variables and discuss them.

1. Patient Information - The demographics and treatment history of each patient affects the expression of genes. This information is not available in the TISCH2 dataset, making it an unmeasured confounder. We will use an instrumental variable approach to compensate for this missing information.

2. Tumor Subtype - One of the challenges of cancer research is that tumors come in many forms and subtypes. Different subtypes will exhibit distinct gene expression signatures, which affects the expression of genes. We do not have access to this level of granularity about the samples we will be using. We will use an instrumental variable approach to compensate for potential bias.

3. Cell Type - Another characteristic of our samples is their specific cell types (see figure 1 above). We *do have access* to this information in the metadata file for the TISCH2 dataset. We will control for the cell type in our analysis.

4. Cell Context - The extracellular matrix (e.g. the cell's environment) also plays a role in regulating gene expression. Understanding their role in cell behavior is still actively being studied [21]. Given that, we will use an instrumental variable to compensate.

Our main approaches for to mitigate omitted variable bias are to:

1. Control for them in the analysis, when possible.

2. Select genes that can be used as an instrumental variable in the analysis.

3. Use propensity score matching between treated and untreated samples.

Empirical Method

Feature Engineering

We will engineer two types of features using our data. From the TISCH2 data, we will calculate the expression level for our genes of interest. Using the MSigDB data, we will calculate a gene set enrichment analysis score with our hallmark gene sets.

Gene Expression Level

For each sample, we obtain this information directly from the dataset as a log-normalized count of the genes. This count is used as a proxy for gene activation, where more genes indicate a more active state ("on"). In addition to the standard data science libraries, we will be using HDF5 (https://www.h5py.org/) to work with the raw dataset.

Gene Set Enrichment Analysis Score

For each sample, we will calculate a gene set enrichment analysis (GSEA) score. GSEA captures the activity of specific biological pathways by evaluating the collective expression of multiple genes that participate in the pathway. In addition to the standard data science libraries, we will be using GSEApy (https://github.com/zqfang/GSEApy) [19] to perform GSEA in a python environment.

Treatment Score

For each sample, we will calculate a treatment score representing the likelihood of "ghost-ness." Using the log-normalized gene counts, we will apply Scrublet (https://github.com/swolock/scrublet) to compute a doublet score. To assess mitochondrial content, we will use Scanpy (https://scanpy.readthedocs.io/en/stable/), to calculate the percentage of mitochondrial genes, which are typically identified by the "MT-" prefix.

Causal Inference Techniques

We apply a series of causal inference techniques to investigate the causal relationship between ghost cell profiles and biological processes related to human breast cancer, which are proxied by our dependent variables, $OUTCOME_i$. We describe below the techniques that will be used for each outcome.

Ordinary Least Squares

First, we establish a baseline model for each of the dependent variables using ordinary least squares regression (OLS). This baseline will show the basic associations between ghost cells and immune evasion, proliferation, and resistance. We treat the subtype as a categorical variable and include it in our analysis as a dummy variable, D_{ji}.

$$OUTCOME_i = \beta_0 + \beta_1 GHOST_i + \sum_{j=1}^{k} \lambda_j D_{ji} + \epsilon$$

For each outcome, we will further conduct a two-sided t-test, where $\alpha = 0.05$ and report the 95% confidence intervals around the estimated coefficient β_1.

Propensity Score Matching

Next, we will address our observables, tumor subtypes, using propensity score matching (PSM). We do this by performing a logistic regression on the tumor subtypes, TS, and performing nearest-neighbor matching without replacement. We choose "without placement" because we anticipate far more controls (not-ghost-cells) than samples (ghost-cell).

$$logit(P(OUTCOME = 1)) = \beta_0 + \beta_1 TS_1 + \beta_2 TS_2 + \ldots + \beta_n TS_n$$

We will assess the balance of covariates between ghost-cells and non-ghost-cells by calculating the standardized mean differences for each covariate before and after matching. Specifically, we will calculate the standard mean difference between the two groups for each covariate before and after matching. Our threshold for balance will be a standard mean difference that is less than or equal to 0.1.

Two-Stage Least Squares

We will address our unobservables using the two-stage least squares regression with housekeeping gene expressions as our instrumental variables (IV). Housekeeping genes represent the core infrastructure for all cells to exist and function. While some genes have been implicated in cancer pathways [17], others have consistently not been associated with cancer pathways [18].

Stage 1

We begin first by predicting the likelihood the sample is a ghost cell, $GHOST_i$, using IVs, where the housekeeping genes are represented as HKG_1 & HKG_2. We control for the tumor subtype in the regression.

$$GHOST_i = \alpha_0 + \alpha_1 HKG_1 + \alpha_2 HKG_2 + \sum_{j=1}^{k} \lambda_j D_{ji} + \epsilon$$

Stage 2

Next, we estimate the causal effect of the ghost cell on the outcome, $OUTCOME_i$, for each of our hypotheses, again controlling for tumor subtypes.

$$OUTCOME_i = \beta_0 + \beta_1 \widehat{y_i} + \sum_{j=1}^{k} \lambda_j D_{ji} + \epsilon$$

We select genes that have consistently not been associated with cancer pathways and describe below their suitability for instrumentation.

Housekeeping Gene	Role
RPL13A	Ribosomal structure
TBP	TATA-binding protein for transcription

Table 5. Selected genes for use in our instrumental variable approach.

Relevance Criterion

The selected genes are linked to pathways involved in transcriptional activity, which is correlated with the amount of gene expression. Since transcriptional activity directly influences RNA levels, which is our proxy for "ghost" cell behavior. High relevance ensures that variation in the genes reflects variation in the treatments.

Independence Criterion

The selected genes have shown no known direct links to cancer pathways, satisfying the independence criterion by reducing the risk that the selected genes are independently associated with cancer outcomes. Without a direct link, observed effects are more likely mediated through "ghost" cells.

Exclusion Criterion

The selected genes have no direct causal links to cancer, ensuring that the genes affect cancer outcomes only through their influence on transcriptional activity and not through any other pathway.

Statistical Analysis Techniques

Control Group Selection

Once we have identified our target population (ghost-cells), we will stratify the entire population on tumor subtypes. For each stratum, we randomly select for our control population (not-ghost-cells). For example, if a given stratum has a total of 30 cells, 5 of which are ghost-cells, we will randomly select 5 non-ghost-cells from the remaining 25 cells.

Baseline Comparison

For each stratum, we will perform a permutation test with 5,000 iterations, where each iteration will randomly select ghost-cells and not-ghost-cells. Continuing the above example, where there are 5 ghost-cells for the stratum, we will randomly select 5 out of the 30 cells to be our $ghost^{perm}$ cells and label the remaining cells as $\overline{ghost}^{perm}$. Using the selected cells, we will then calculate the mean difference between the outcome for each of our dependent variables.This allows us to compare the observed mean difference to the distribution of mean differences under the null hypothesis. Our null hypothesis states that there is no difference in outcome between ghost-cells and not-ghost-cells.

$$MEAN_{ghost} = \frac{1}{N_{ghost}} \sum_{i \in ghost} OUTCOME_i$$

$$MEAN_{\overline{ghost}} = \frac{1}{N_{\overline{ghost}}} \sum_{j \in \overline{ghost}} OUTCOME_j$$

$$\Delta OUTCOME = MEAN_{ghost} - MEAN_{\overline{ghost}}$$

Adjusted Causal Effect

To adjust for tumor subtypes, we will perform a regression analysis for each outcome and, again, apply permutation testing to obtain a null distribution of the treatment effect. This will allow us to assess the significance of any causal effects. Given the complexity of our analysis from unknown unknowns, each permutation for adjusted causal effect will have 10,000 iterations. We will use the same regression model as in the OLS analysis. With K being the number of iterations, β^k is the coefficient at permutation k, and β^{obs} is the observed coefficient, we perform a two-tailed test to calculate the p-value for each stratum:

$$p = \frac{1 + \sum_{k=1}^{K} 1\left(\left|\hat{\beta}^k \geq \hat{\beta}^{obs}\right|\right)}{1+K}$$

Robustness Checks

Recognizing that power limitations could impact the reliability of causal inference, we incorporated considerations of sample size and detectable effect sizes into our approach. To ensure that our analyses are adequately powered, we will monitor the distribution of treated and untreated samples post-matching and assess effective sample sizes following each causal inference technique.

Next, we will check the robustness of our analysis using placebo and overidentification tests. The placebo test will provide additional evidence against spurious relationships between ghost cells and our outcomes. We will use another housekeeping gene, HPRT [20], as our placebo outcome. We use an overidentification test to provide evidence on the validity of our choice of instrumental variables with a Wu-Hausman test.

IV. Conclusions and Limitations

This work aims to reevaluate the biological significance of ghost cells within human breast tumors, moving beyond assumptions of technical artifacts toward a more causally informed understanding. By integrating proxy measures for “ghost-ness” with structured causal inference methods, we seek to uncover signals that traditional scRNA-seq pipelines might obscure. Through this approach, we hope to contribute initial insights into how low RNA profile cells might influence key processes such as immune evasion, tumor proliferation, and chemotherapy resistance.

However, it is important to be transparent about the limitations of this analysis. We begin with the most significant: we are not trained medical researchers and lack the expertise to fully characterize the biological basis of ghost cells. Our work represents an early, computational step toward a much deeper biological investigation that would require extensive experimental validation. Additionally, we are constrained by the availability and nature of publicly available datasets. Without the ability to generate new data tailored specifically to our questions, we rely on observational, cross-sectional data, which inherently limits the strengths of causality. While causal inference tools offer valuable insights, they cannot substitute for direct experimental evidence.

Despite these limitations, we believe that applying causal methods to the analysis of ghost cells offers a meaningful conceptual shift, reframing them not as noise as potential signals in key processes. We view this analysis as an invitation for further biological and translational work to validate and expand upon these findings.

V. Footnotes

1. **“We”** is used for narrative flow, and not an indication of shared work.
2. **Single-cell RNA sequencing (scRNA-seq)** is a method that measures gene expression in individual cells, allowing researchers to study cell-to-cell differences within complex tissues like tumors.
3. **Ghost cells** is an informal term used for cells that have unusually low RNA activity, making them difficult to distinguish from debris.
4. **Debris** refers to dead cell fragments, and **doublets** are technical artifacts where two cells are mistakenly recorded as one, both of which can confound scRNA-seq analysis.
5. **Ambient RNA contamination** happens when free floating RNA molecules in the solution are mistakenly captured during sequencing, creating noise that can mimic real biological signals.
6. **Tumor heterogeneity** describes how different cells within the same tumor can have diverse characteristics, making the tumor harder to treat with a single strategy.
7. **Low RNA cell profiles** are cells that produce very little detectable RNA. They might be living but transcriptionally quiet, or they could be technical artifacts, so they are often mistakenly removed from data during analysis.
8. Highlighted text was added in response to instructor feedback on assignment.

VI. References

1. Han, Ya, Wang, Yuting, Dong, Xin, Sun, Dongqing, Liu, Zhaoyang, Yue, Jiali, Wang, Haiyun, Li, Taiwen, Wang, Chenfei. TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. Nucleic acids research. England: Oxford University Press; 2023;51(D1):D1425–D1431.

2. Sun, Dongqing, Wang, Jin, Han, Ya, Dong, Xin, Ge, Jun, Zheng, Rongbin, Shi, Xiaoying, Wang, Binbin, Li, Ziyi, Ren, Pengfei, Sun, Liangdong, Yan, Yilv, Zhang, Peng, Zhang, Fan, Li, Taiwen, Wang, Chenfei. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic acids research. England: Oxford University Press; 2021;49(D1):D1420–D1430.

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. 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-.

5. 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.

6. Yardley, Denise A. Drug Resistance and the Role of Combination Chemotherapy in Improving Patient Outcomes. International Journal of Breast Cancer. Cairo, Egypt: Hindawi Limiteds; 2013;2013(2013):26–40.

7. Abad, Etna, Graifer, Dmitry, Lyakhovich, Alex. DNA damage response and resistance of cancer stem cells. Cancer letters. Ireland: Elsevier B.V; 2020;474:106–117.

8. Bonev, Boyan, Castelo-Branco, Gonçalo, Chen, Fei, Codeluppi, Simone, Corces, M. Ryan, Fan, Jean, Heiman, Myriam, Harris, Kenneth, Inoue, Fumitaka, Kellis, Manolis, Levine, Ariel, Lotfollahi, Mo, Luo, Chongyuan, Maynard, Kristen R., Nitzan, Mor, Ramani, Vijay, Satijia, Rahul, Schirmer, Lucas, Shen, Yin, Sun, Na, Green, Gilad S., Theis, Fabian, Wang, Xiao, Welch, Joshua D., Gokce, Ozgun, Konopka, Genevieve, Liddelow, Shane, Macosko, Evan, Ali Bayraktar, Omer, Habib, Naomi, Nowakowski, Tomasz J. Opportunities and challenges of single-cell and spatially resolved genomics methods for neuroscience discovery. Nature neuroscience. New York: Nature Publishing Group US; 2024;27(12):2292–2309.

9. Wei, Jing, Li, Wenke, Zhang, Pengfei, Guo, Fukun, Liu, Ming. Current trends in sensitizing immune checkpoint inhibitors for cancer treatment. Molecular cancer. England: BioMed Central Ltd; 2024;23(1):279–25.

10. Budczies, Jan, Bockmayr, Michael, Denkert, Carsten, Klauschen, Frederick, Gröschel, Stefan, Darb-Esfahani, Silvia, Pfarr, Nicole, Leichsenring, Jonas, Onozato, Maristela L., Lennerz, Jochen K., Dietel, Manfred, Fröhling, Stefan, Schirmacher, Peter, Iafrate, A. John, Weichert,

Wilko, Stenzinger, Albrecht. Pan-cancer analysis of copy number changes in programmed death-ligand 1 (PD-L1, CD274) - associations with gene expression, mutational load, and survival. Genes chromosomes & cancer. United States: Blackwell Publishing Ltd; 2016;55(8):626–639.

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. Liu, Kai, Zheng, Minying, Lu, Rui, Du, Jiaxing, Zhao, Qi, Li, Zugui, Li, Yuwei, Zhang, Shiwu. The role of CDC25C in cell cycle regulation and clinical cancer therapy: a systematic review. Cancer cell international. London: BioMed Central; 2020;20(1):1–213.

13. Lu, Yongbin, et al. "E2F1 Transcriptionally Regulates CCNA2 Expression to Promote Triple Negative Breast Cancer Tumorigenicity." *Cancer Biomarkers : Section A of Disease Markers*, vol. 33, no. 1, 2022, pp. 57–70, https://doi.org/10.3233/CBM-210149.

14. Xing, Zeyu, Wang, Xin, Liu, Jiaqi, Zhang, Menglu, Feng, Kexin, Wang, Xiang. Expression and prognostic value of CDK1, CCNA2, and CCNB1 gene clusters in human breast cancer. Journal of international medical research. London, England: SAGE Publications; 2021;49(4):300060520980647–300060520980647.

15. Gilmore E, McCabe N, Kennedy RD, Parkes EE. DNA Repair Deficiency in Breast Cancer: Opportunities for Immunotherapy. J Oncol. 2019 Jun 19;2019:4325105. doi: 10.1155/2019/4325105. PMID: 31320901; PMCID: PMC6607732.

16. Cybulla, Emily, Vindigni, Alessandro. Leveraging the replication stress response to optimize cancer therapy. Nature reviews Cancer. London: Nature Publishing Group UK; 2023;23(1):6–24.

17. Wang, Jin, Yu, Xueting, Cao, Xiyuan, Tan, Lirong, Jia, Beibei, Chen, Rui, Li, Jianxiang. GAPDH: A common housekeeping gene with an oncogenic role in pan-cancer. Computational and structural biotechnology journal. Elsevier B.V; 2023;21:4056–4069.

18. Jo, Jihoon, Choi, Sunkyung, Oh, Jooseong, Lee, Sung-Gwon, Choi, Song Yi, Kim, Kee K, Park, Chungoo. Conventionally used reference genes are not outstanding for normalization of gene expression in human cancer research. BMC bioinformatics. England: BioMed Central Ltd; 2019;20(Suppl 10):245–245.

19. Fang, Zhuoqing, Liu, Xinyuan, Peltz, Gary. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics (Oxford, England). England: Oxford University Press; 2023;39(1).

20. de Kok, Jacques B, Roelofs, Rian W, Giesendorf, Belinda A, Pennings, Jeroen L, Waas, Erwin T, Feuth, Ton, Swinkels, Dorine W, Span, Paul N. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Laboratory investigation. New York: Elsevier Inc; 2005;85(1):154–159.

21. Lu, Pengfei, Takai, Ken, Weaver, Valerie M, Werb, Zena. Extracellular matrix degradation and remodeling in development and disease. Cold Spring Harbor perspectives in biology. United States: Cold Spring Harbor Laboratory Press; 2011;3(12).

22. 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.

23. Cybulla, Emily, Vindigni, Alessandro. Leveraging the replication stress response to optimize cancer therapy. Nature reviews Cancer. London: Nature Publishing Group UK; 2023;23(1):6–24.