On the illusion of a pre-treatment chemoresistance signal in HR+ breast cancer

Krishay Uthayasegar

Abstract

Small single cell RNA sequencing cohorts are increasingly used to search for pre-treatment cell states that predict chemotherapy resistance, but such studies are vulnerable to under-appreciated failure modes: cell-level pseudoreplication, patient batch effects masquerading as shared biology, and gene signatures that inadvertently capture proliferation rather than resistance. We reanalyzed a paired single cell RNA sequencing dataset of pre- and post-chemotherapy tumor biopsies from five HR+ breast cancer patients (three resistant, two sensitive) to test whether pre-treatment tumor cells resemble post-treatment surviving cells, then subjected the signal to four robustness checks. An initial per-cell analysis produced a seemingly positive result (one-sided Mann-Whitney U p = 0.069): resistant patients pre-treatment cells clustered with the post-treatment compartment more often than sensitive patients cells. This did not survive scrutiny. Harmony batch correction for patient identity showed the association was driven almost entirely by one patient contributing six pre-treatment tumor cells, whose post-treatment-compartment membership fell from 100% to 0% after correction (corrected p = 0.1665). A parallel signature derived from post-treatment upregulated genes scored higher, not lower, in pre-treatment sensitive-patient cells (p = 1.000); gene set enrichment analysis confirmed this signature was dominated by cell-cycle and mitotic pathways (Hallmark G2-M Checkpoint, adjusted p = 1.4x10^-9) rather than a resistance-specific program. A patient-balanced pseudobulk reanalysis found a consistent direction of effect across all three resistant patients but did not reach significance (paired Wilcoxon p = 0.125, n = 3). These results show that a superficially promising resistance signal in a small cohort can arise from patient-specific batch structure and proliferation confounding rather than reproducible biology, and that patient-level pseudobulk analysis, batch integration, and pathway enrichment testing are necessary, low-cost checks before such signals are reported as candidate biomarkers. We provide this as a case study and checklist for evaluating single cell resistance signatures in small cohorts.

Keywords

Breast cancer, Single cell RNA sequencing, Chemotherapy resistance, Batch effects, Harmony integration, Pseudobulk analysis, Gene set enrichment analysis, Reproducibility, Pseudoreplication, Hormone receptor positive

1. Introduction

Breast cancer is the most frequently diagnosed malignancy among women worldwide, with approximately 2.3 million new cases reported globally in 2020 (1). Among its molecular subtypes, hormone receptor positive (HR+) breast cancer accounts for nearly seventy percent of diagnoses and is routinely treated with neoadjuvant chemotherapy (NAC) prior to surgery to reduce tumor burden (2). Pathological complete response, defined as the absence of residual invasive tumor in the breast and axillary lymph nodes at surgery, is the standard readout of chemotherapy sensitivity in the neoadjuvant setting. However, rates of pathological complete response to NAC remain as low as 0.3% in luminal A subtypes, and the molecular mechanisms underlying this chemoresistance are poorly understood (3).

Standard neoadjuvant chemotherapy for HR+ breast cancer is not a single agent but a multi-drug regimen delivered over several cycles, typically combining an anthracycline such as doxorubicin or epirubicin, an alkylating agent such as cyclophosphamide, and a taxane such as docetaxel or paclitaxel. Response varies substantially with regimen and molecular subtype: HER2-positive tumors treated with concurrent HER2-targeted therapy respond most often, triple negative tumors respond intermediately, and HR+ tumors respond least often, with luminal A HR+ tumors having the lowest pathological complete response rates of any subtype. Because response varies so markedly by subtype and by patient within a subtype, identifying molecular features that predict response before treatment begins is a persistent goal of translational breast cancer research.

A major obstacle to understanding resistance is intratumoral heterogeneity. Individual tumors comprise molecularly distinct subpopulations with divergent transcriptional programs and drug sensitivities (4). Conventional bulk RNA sequencing averages gene expression across millions of cells, masking rare subpopulations that may disproportionately drive treatment failure (5). Two models have been proposed to explain chemoresistance: acquired resistance, driven by treatment induced selective pressure, and pre-existing resistance, in which rare cell states harboring intrinsic resistance programs survive therapy because they were present before treatment began (6). Distinguishing between these models is clinically significant, since pre-existing resistance could in principle be detected and targeted prior to treatment initiation.

Recent scRNA-seq studies in breast cancer have revealed subpopulations associated with chemotherapy resistance, including drug tolerant persister states mediated by chromatin remodeling (6). However, systematic characterization of pre-existing, intrinsically resistant cell states specifically in HR+ breast cancer, and the transcriptional programs that define them, remains limited.

Despite this prior work, whether pre-existing intrinsically resistant tumor cell states can be detected in HR+ breast cancer before neoadjuvant chemotherapy remains an open question, and studies attempting to answer it typically share a common structural vulnerability: cohorts are small (often fewer than ten patients), the analytical unit of evidence is frequently the individual cell rather than the patient, and candidate resistance signatures are rarely tested for confounding by generic processes such as cell cycle activity or by patient-specific batch structure. Each of these vulnerabilities can independently produce a signal that looks like reproducible biology but is not. Cell-level statistical tests silently treat thousands of correlated cells from one patient as independent observations, inflating apparent significance; patient identity and technical batch effects can masquerade as shared transcriptional states when not explicitly corrected for; and post-versus-pre differential expression signatures can inadvertently

capture cell-cycle or treatment-response biology rather than a resistance-specific program, particularly when chemotherapy itself preferentially eliminates actively cycling cells.

The objective of this study was twofold. First, we reanalyzed a paired pre- and post-neoadjuvant chemotherapy single cell RNA sequencing dataset of HR+ breast cancer (GSE205472) to test the hypothesis that chemotherapy resistant tumors harbor transcriptionally distinct tumor cell states, present before treatment, that resemble cells surviving chemotherapy. Second, and more centrally to the contribution of this study, we subjected the resulting signal to four independent robustness checks, patient-balanced pseudobulk analysis, Harmony batch integration, cell-cycle regression, and gene set enrichment analysis, that are inexpensive to perform but not yet standard practice in small-cohort single cell resistance studies. We report the outcome of each check in full, including where it overturns the initial result, with the goal of providing a concrete, reproducible case study of how such checks can be used to evaluate whether an apparent resistance signal reflects biology or analytical artifact.

2. Methods

This study is explicitly retrospective and exploratory in design. Because chemotherapy response was known for all five patients before any analysis was performed, the resistance signature was necessarily derived using outcome information, and the subsequent test for that signature (or a related transcriptional compartment) in pre-treatment samples of the same patients is a test of retrospective self-consistency rather than an independent predictive test. This design can identify candidate patterns worth testing in future, adequately powered, independently validated cohorts; it cannot by itself establish that pre-treatment resistant-like cells exist and can be used prospectively to predict response. We adopt this framing throughout and report results, including negative and fragile ones, accordingly.

2.1 Dataset and preprocessing

Single cell RNA sequencing data were obtained from the NCBI Gene Expression Omnibus (accession GSE205472), comprising paired pre- and post-neoadjuvant chemotherapy tumor biopsies from five HR+ breast cancer patients (2). In the source study, patients received standard neoadjuvant chemotherapy administered over eight cycles, with per-patient regimens reported therein (2). Although the dataset comprised eight patients, five had paired pre- and post-treatment single cell samples passing quality control and were used here, because the post-treatment versus pre-treatment comparison requires both timepoints. Three patients (SC-P04, SC-P05, SC-P07) were chemotherapy resistant and two (SC-P01, SC-P02) were chemotherapy sensitive, based on the clinical resistance labels provided with the dataset. Raw count matrices were loaded with Scanpy's read_10x_mtx function (v1.9) (7), using gene symbols as feature identifiers. Sample identity and treatment state were mapped to cell barcodes using barcode suffixes.

Quality control was applied in two stages. Cells expressing fewer than 200 genes and genes detected in fewer than three cells were removed, and cells exceeding 20% mitochondrial content were excluded as putative apoptotic or damaged cells. Following quality control, 76,118 cells were retained. Counts were normalized to 10,000 reads per cell and log transformed using Scanpy's normalize_total and log1p functions. The 2,000 most highly variable genes were used for dimensionality reduction and clustering, while the full gene matrix was retained for scoring and visualization.

2.2 Dimensionality reduction and clustering

Principal component analysis was performed on the highly variable gene matrix using the ARPACK solver in Scanpy's PCA function with 30 principal components. A k nearest neighbor graph was constructed with k = 15 using Scanpy's neighbors function, UMAP embeddings were computed with Scanpy's UMAP function (8), and Leiden clustering was performed with Scanpy's Leiden clustering function (9). Clustering was performed on the full multi patient dataset without batch correction, as all samples were provided as a single pre merged count matrix.

2.3 Automated cell type annotation and removal of immune cells

Automated cell type annotation was performed with CellTypist v1.3 using the Cells_Adult_Breast reference model and majority voting (10). Cells assigned to immune subtypes (T cell, B cell, natural killer, macrophage, monocyte, dendritic, mast, plasma, and granulocyte related labels) were removed. This was motivated by the observation that immune and non-immune populations form highly distinct groupings that dominate the neighbor graph and obscure finer distinctions among non-immune cells, such as tumor versus non-tumor luminal epithelial cells.

2.4 Luminal subsetting and re-clustering

Because breast cancer arises predominantly from luminal epithelial cells, the immune depleted dataset was subset to cells assigned a luminal epithelial CellTypist label. Principal component analysis, neighbor graph construction, Leiden clustering, and UMAP embedding were recomputed on this luminal subset independently of the full dataset clustering, yielding a set of luminal clusters used for tumor cell identification.

2.5 Quantitative tumor versus stromal labeling

To label tumor versus stromal cells within the luminal compartment without manual, potentially biased cluster assignment, gene set scores were computed for every luminal cell using Scanpy's score_genes function. A tumor score was computed from the markers TSPAN1, CDC20, ID3, MKI67, TOP2A, TNFRSF18, and TNFRSF4, and a stromal score from COL1A1, COL1A2, VWF, PECAM1, ACTA2, VWA1, HSPG2, TIE1, and ANGPTL7. A tumor score threshold was selected using the biological expectation that post-treatment tumor cells should be depleted in chemotherapy sensitive patients, whose tumors were eliminated by therapy, while pre-treatment cells and post-treatment resistant cells should retain tumor cells. Cells above the threshold were labeled tumor and all others stromal; no ambiguous category was retained. This threshold choice is a limitation: because it was informed by the expected treatment outcome, the resulting tumor cell classification is not fully independent of chemotherapy response, and this outcome-informed step should be considered when interpreting the downstream resistance analyses. A threshold selected by an outcome-independent criterion, such as inferred copy-number variation or an unsupervised bimodality criterion on the score distribution, would avoid this dependency and is recommended for future work.

2.6 Resistance classification

Patient level chemotherapy resistance status was taken directly from the clinical labels provided with GSE205472 (2), classifying SC-P04, SC-P05, and SC-P07 as resistant and SC-P01 and SC-P02 as sensitive. These labels were used for all downstream comparisons.

2.7 Differential expression and resistance signature

For each resistant patient, differential expression between post-treatment and pre-treatment tumor cells was computed using the Wilcoxon rank-sum test via Scanpy's rank_genes_groups function, with the post-treatment group as the test group. To ensure that only genuinely upregulated genes entered the aggregation, results were filtered to genes with log-fold-change > 0 and adjusted $p < 0.05$ for each patient. Patient SC-P07 produced no significantly upregulated genes under this filter because only six pre-treatment tumor cells were available, and was excluded from signature derivation but retained as a resistant patient in all downstream hypothesis testing. Rank aggregation was then performed across SC-P04 and SC-P05: each gene's aggregate rank was the sum of its per-patient ranks (rank 1001 for a gene absent from a patient's list), and the 50 genes with the lowest aggregate rank were retained as the resistance signature.

2.8 Resistance scoring and hypothesis testing

Per cell resistance scores were computed across tumor cells with Scanpy's score_genes function, which compares mean expression of the signature genes to a control gene set matched for expression level (11) using a fixed random seed. To test the hypothesis, two metrics were compared between sensitive and resistant patients on pre-treatment tumor cells: the fraction of a patient's pre-treatment tumor cells with a resistance score above 0.3, and the fraction falling within the post-treatment compartment, defined as tumor subclusters that were at least 90% post-treatment. Group differences were assessed with the one-sided Mann-Whitney U test (12).

2.9 Robustness checks

To test whether the resistance signature and hypothesis-test results were supported at the level of the patient rather than the cell, gene expression was collapsed into one pseudobulk profile per patient per timepoint by averaging log-normalized expression across all tumor cells from that patient and timepoint. A gene was retained in a pseudobulk-derived signature only if its pseudobulk expression was higher in post-treatment than pre-treatment for every resistant patient with both timepoints available. A pseudobulk resistance score was computed per patient per timepoint as the mean expression of this reproducible gene set, and pre- versus post-treatment scores were compared across the three resistant patients using a paired one-sided Wilcoxon signed-rank test ($n = 3$ patients).

To test whether patient identity or technical batch effects could explain the clustering-based hypothesis test result, Harmony integration was applied to the tumor cell principal component embedding (50 components), correcting for patient as the batch variable (13). Neighbor graph construction, Leiden clustering, and UMAP embedding were recomputed on the Harmony-corrected embedding using the same parameters as the uncorrected analysis ($k = 15$ neighbors, resolution 1.0). The post-treatment compartment was redefined on the corrected clusters using the same 90% post-treatment threshold, and the hypothesis test was repeated on this corrected definition.

To test whether the resistance signature reflected cell-cycle activity rather than a distinct resistance program, tumor cells were scored for S-phase and G2M-phase using the marker gene sets of Tirosh et al. (11) as implemented in Scanpy's score_genes_cell_cycle function. S-phase and G2M scores were regressed out of the expression matrix for every gene using

Scanpy's regress_out function, and the resistance signature was re-derived on the corrected data using the same per-patient differential expression and rank-aggregation procedure described in Section 2.7. Because this dataset had been filtered to the top 2,000 most highly variable genes prior to this analysis (Section 2.1), only a subset of canonical cell-cycle marker genes were present for scoring; this limits the completeness of the correction and is addressed in the Discussion.

To characterize the biological pathways represented by the resistance signature, over-representation analysis was performed on the original 50-gene signature using the Enrichr web service via the gseapy package (14), testing against the MSigDB Hallmark, GO Biological Process, KEGG, and Reactome gene set libraries. Enrichment was also tested for the subset of cell-cycle-regression-derived signature genes that were not canonical cell-cycle genes, to assess whether a non-proliferation pathway signal was present. Adjusted p-values (Benjamini-Hochberg) were used to rank enriched terms.

3. Results

3.1 Isolation of a tumor cell population

Automated annotation with CellTypist assigned the 76,118 quality controlled cells from the five patients to fine grained breast cell types, with luminal hormone receptor and basal epithelial cells forming the largest populations alongside substantial immune infiltrate. Removing immune cells and re-clustering the luminal compartment produced clusters with clearer spatial separation than the original full dataset clustering, in which immune and non-immune populations had dominated the overall structure. Quantitative tumor labeling using gene set scores, rather than manual cluster assignment, produced a tumor cell population whose marker expression was consistent with malignant epithelium and which recovered tumor cells across patients and treatment timepoints. Consistent with the biological expectation used to set the score threshold, post-treatment tumor cells were depleted in the chemotherapy sensitive patients. The re-clustered luminal compartment, tumor and stromal marker expression, and the resulting tumor versus stromal assignments are shown in Figure 1.

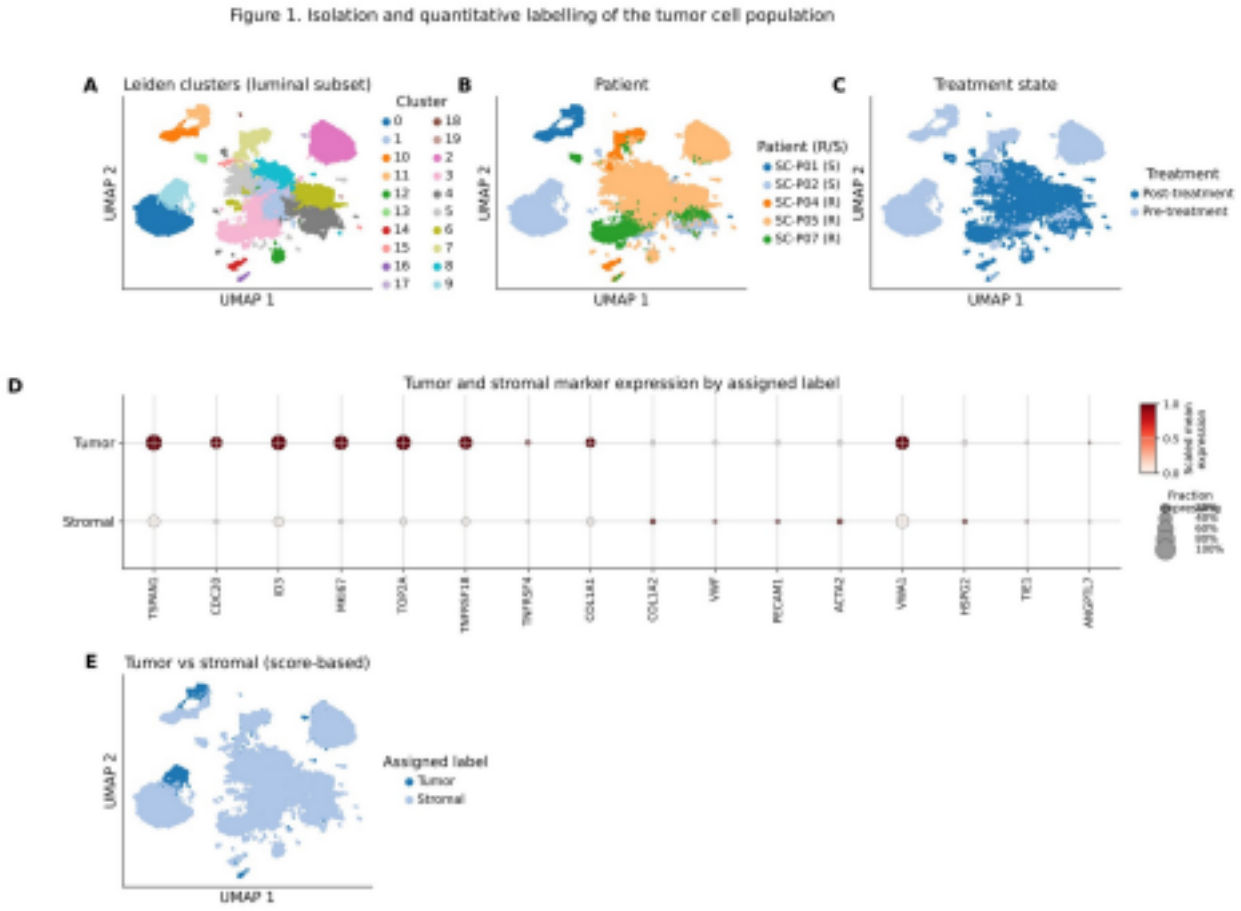

Figure 1. Isolation and quantitative labelling of the tumor cell population. (A) UMAP of the re-clustered luminal epithelial compartment coloured by Leiden cluster. (B) The same UMAP coloured by patient, with resistant (R) and sensitive (S) status annotated. (C) The same UMAP coloured by treatment state. (D) Dot plot of tumor and stromal marker expression by the score-based assigned label. (E) The luminal UMAP coloured by the score-based tumor vs stromal assignment. Tumor identity was assigned quantitatively using a tumor gene set score threshold rather than manual cluster annotation.

3.2 Derivation of a resistance signature

Differential expression between post-treatment and pre-treatment tumor cells identified 378 upregulated genes in patient SC-P04 and 482 upregulated genes in patient SC-P05 under the up-in-post filter, whereas patient SC-P07 produced no significantly upregulated genes and was excluded from signature derivation. Rank aggregation across SC-P04 and SC-P05 produced a shared 50 gene resistance signature. The signature was dominated by proliferation and cell-cycle genes, including TOP2A, MKI67, CDK1, CENPF, PTTG1, BIRC5, ASPM, TPX2, NUSAP1, and UBE2C, together with a smaller set of immune-related genes including TRBC1, CD7, HLA-DQA1, TNFRSF1B, TNFRSF4, and HERC5. Heatmap visualization confirmed that the displayed signature genes were more highly expressed in post-treatment than pre-treatment tumor cells in each of the two signature-derivation patients. The proliferation character of the signature has direct consequences for the hypothesis-testing results below and is examined

further in the Discussion. Differential expression volcanoes for SC-P04 and SC-P05 and the signature heatmap are shown in Figure 2.

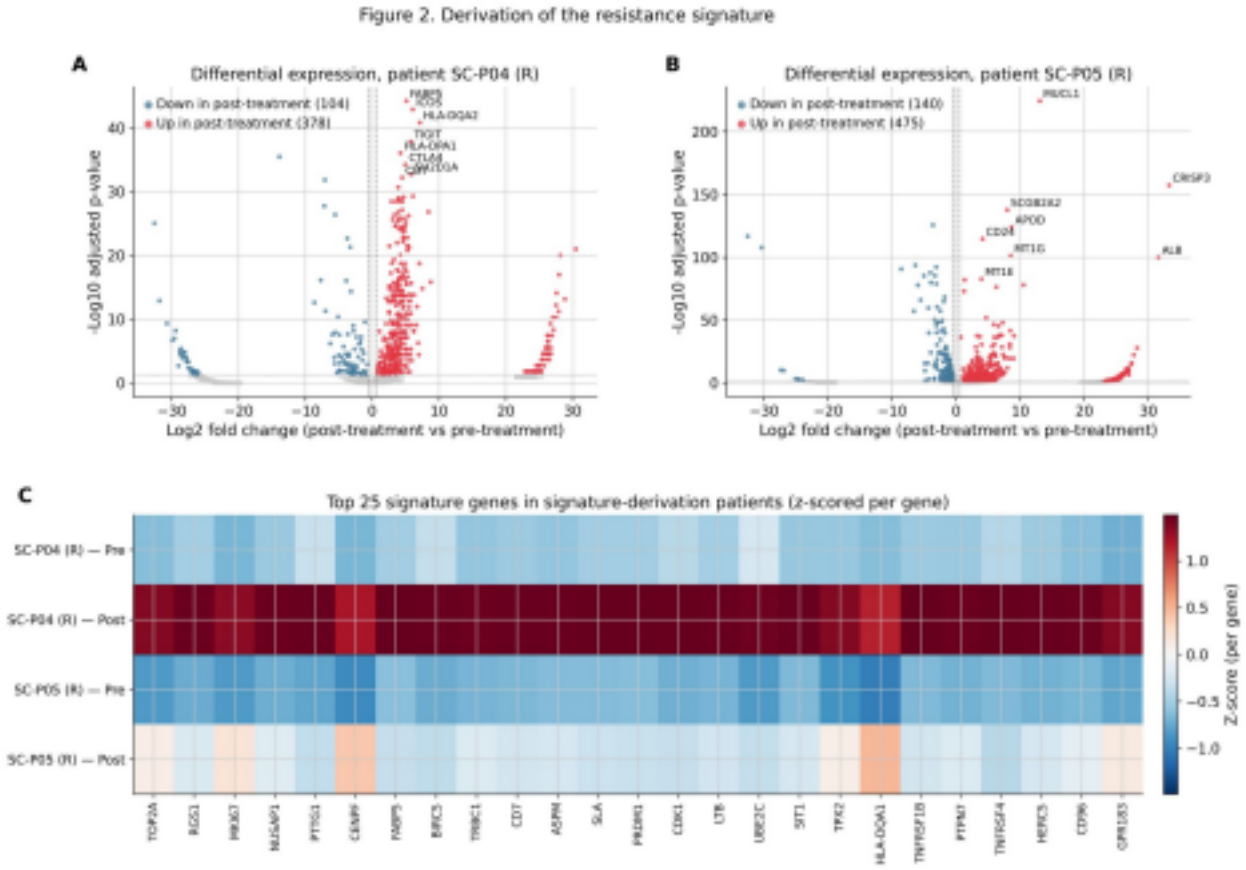

Figure 2. Derivation of the initial candidate resistance signature. (A, B) Volcano plots of post-treatment versus pre-treatment differential expression in resistant tumor cells for patients SC-P04 and SC-P05, the two resistant patients used to derive the signature, showing genes upregulated (red) and downregulated (blue) in post-treatment cells. (C) Heatmap of the top 25 signature genes across the two signature-derivation patients (SC-P04 and SC-P05) at pre-treatment and post-treatment timepoints, with expression averaged per patient by timepoint and z-scored per gene across the four rows. Patient SC-P07 was excluded from signature derivation because its six pre-treatment tumor cells produced no significantly upregulated genes under the up-in-post filter, and is therefore not shown in this display; SC-P07 is retained as a resistant patient in the hypothesis test (Figure 3). Gene set enrichment analysis (Section 3.4) subsequently confirmed that this signature is dominated by cell-cycle and mitotic pathways rather than a resistance-specific program.

3.3 Pre-existing resistant-like cells in resistant but not sensitive patients

Two independent tests of the hypothesis were run on pre-treatment tumor cells. The first metric, the fraction of pre-treatment tumor cells with a resistance score above 0.3, did not support the hypothesis: sensitive patients had a mean of 15.13% cells above the threshold,

whereas resistant patients had a mean of 1.55%. The second metric, the fraction of pre-treatment tumor cells falling within the post-treatment compartment (defined as tumor subclusters that were at least 90% post-treatment), gave a per-patient pattern that must be reported at the individual level rather than as a group mean: patient SC-P07 showed 100% of its six pre-treatment tumor cells within the post-treatment compartment, whereas patients SC-P04 and SC-P05 showed 0.31% and 0.28% respectively, values indistinguishable from the 0.00% observed in both sensitive patients. The group mean of 33.53% for resistant patients is therefore driven almost entirely by one patient contributing only six pre-treatment tumor cells, and does not represent a pattern recurring across the resistant group.

One-sided Mann-Whitney U tests returned p = 1.000 for the score threshold metric and p = 0.069 for the clustering compartment metric. With three resistant and two sensitive patients, neither test has meaningful statistical power, and the p = 0.069 result should not be read as strong evidence: as shown above, it is generated almost entirely by a single patient with six pre-treatment tumor cells, and would not persist if that patient were reweighted or excluded from this metric. The reversal on the score threshold metric has a specific, identifiable cause rather than being attributable only to sample size, and its origin is examined in the Discussion. The violin distribution of resistance scores across pre-treatment tumor cells shows that sensitive patients exhibited systematically higher scores than resistant patients. Per-patient metrics and hypothesis-test results are shown in Figure 3 and Table 1, and we emphasize the per-patient table over the group means for both metrics.

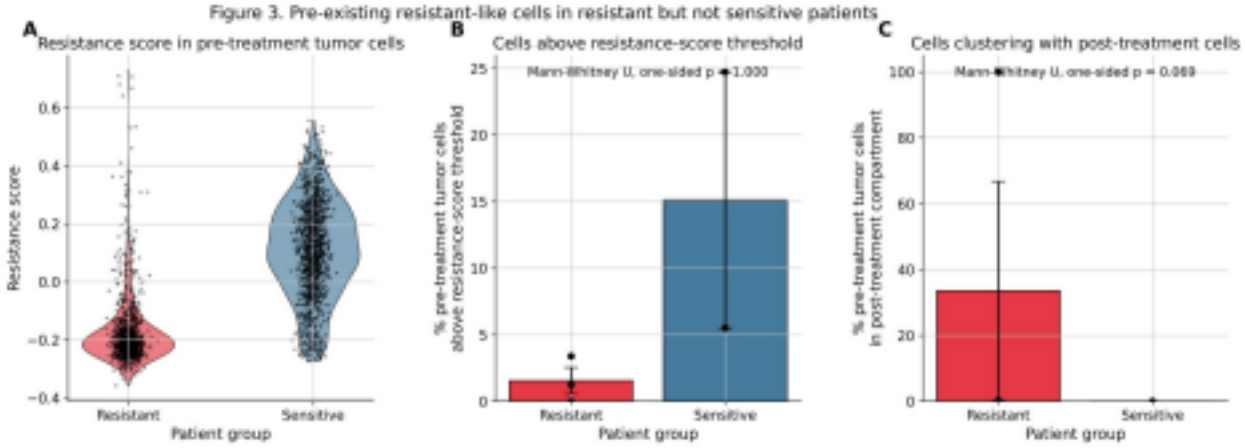

Figure 3. Initial hypothesis test for pre-treatment resistance-like cells, prior to robustness testing. (A) Violin distribution of resistance scores across pre-treatment tumor cells, by patient group. (B) Fraction of pre-treatment tumor cells with a resistance score above 0.3, compared between sensitive and resistant patients, with the one-sided Mann-Whitney U p value annotated. (C) Fraction of pre-treatment tumor cells clustering with the post-treatment compartment, compared between sensitive and resistant patients. Points indicate individual patients; bars indicate group means with standard error. As shown in Section 3.4, the result in panel B is in the opposite direction from the hypothesis, and the result in panel C did not survive Harmony batch correction for patient identity (corrected p = 0.1665); neither panel should be read as evidence of a reproducible pre-treatment resistance signal.

Patient	Resistance	Pre-treatment tumor cells	% above threshold	% in post-treatment compartment
SC-P01	Sensitive	477	24.74	0.00
SC-P02	Sensitive	852	5.52	0.00
SC-P04	Resistant	324	3.40	0.31
SC-P05	Resistant	717	1.26	0.28
SC-P07	Resistant	6	0.00	100.00

Table 1. Per-patient pre-treatment metrics. For each patient, the number of pre-treatment tumor cells and the fraction of those cells that were above the resistance score threshold or within the post-treatment compartment.

3.4 Robustness checks overturn the initial clustering result and confirm proliferation confounding

Because the initial results in Section 3.3 involved a small cohort and a result driven by one patient, four independent robustness checks were performed (Section 2.9) before drawing any conclusion about resistance biology.

Patient-level pseudobulk analysis.
Collapsing each patient's cells into one pseudobulk profile per timepoint, 367 genes were more highly expressed in post-treatment than pre-treatment pseudobulk profiles in all three resistant patients (SC-P04, SC-P05, SC-P07), including patient SC-P07 despite its low cell count. A pseudobulk resistance score based on these reproducible genes increased from pre-treatment to post-treatment in every resistant patient (SC-P04: 0.037 to 0.221; SC-P05: 0.021 to 0.083; SC-P07: 0.011 to 0.045), a consistent direction across all three independent patients. However, the paired one-sided Wilcoxon signed-rank test on these three patients did not reach significance ($p = 0.125$), which is expected given that three paired observations cannot produce a smaller p-value under this test regardless of effect size. This analysis, which respects the patient as the unit of evidence rather than the cell, found a directionally consistent but statistically inconclusive signal.

Harmony batch correction.
After Harmony integration correcting for patient identity, the post-treatment compartment was redefined on the corrected clusters and the hypothesis test was repeated. The result changed substantially. Patient SC-P07's fraction of pre-treatment tumor cells within the post-treatment compartment fell from 100% (6 of 6 cells) before correction to 0% after correction. SC-P04 increased from 0.31% to 5.56%, and SC-P05 changed from 0.28% to 0.42%; both sensitive patients remained at 0.00%. The corrected one-sided Mann-Whitney U test gave $p = 0.1665$, weaker than the uncorrected $p = 0.069$. The complete reversal of patient SC-P07's contribution after batch correction indicates that its apparent 100% membership in the post-treatment compartment before correction reflected patient-specific transcriptional structure rather than shared biology with the post-treatment state, consistent with the concern that a

six-cell sample from a single patient cannot support a claim about a recurring resistant-cell phenotype.

Cell-cycle regression.

Cell cycle scoring was limited by the prior restriction of the dataset to the 2,000 most highly variable genes (Section 2.1): only 3 of 43 canonical S-phase marker genes and 13 of 54 canonical G2M-phase marker genes were present for scoring. With this restricted marker set, 1,285 tumor cells were assigned to S-phase, 751 to G2M-phase, and 1,256 to G1. After regressing out S-phase and G2M scores and re-deriving the signature, none of the resulting top 50 genes were canonical cell-cycle genes; however, this signature also lacked coherent biological composition, including six unannotated long non-coding RNA loci, two mitochondrial genes, and genes with no plausible relevance to breast tumor biology (for example, the taste receptor gene TAS2R31). Given the limited cell-cycle marker coverage available in this highly-variable-gene-filtered dataset, this analysis is considered inconclusive rather than evidence of a non-proliferation resistance signal, and is discussed further as a methodological limitation.

Gene set enrichment analysis.

Over-representation analysis of the original 50-gene signature confirmed strong enrichment for cell-cycle and mitotic pathways: Hallmark G2-M Checkpoint (10 of 200 genes, adjusted p = 1.4x10^-9), Hallmark Mitotic Spindle (7 of 199 genes, adjusted p = 3.1x10^-6), Hallmark E2F Targets (7 of 200 genes, adjusted p = 3.1x10^-6), and Reactome Cell Cycle, Mitotic (10 of 523 genes, adjusted p = 1.3x10^-4). Immune-related terms were also enriched, including Hallmark IL-2/STAT5 Signaling and GO Regulation of T Cell Activation (adjusted p = 3.1x10^-6 and 5.5x10^-6 respectively), reflecting the smaller immune-associated gene component of the signature. This enrichment analysis provides a formal, reproducible confirmation that the signature's dominant biological signal is cell-cycle activity rather than a coherent, resistance-specific mechanism.

Taken together, these four checks show that the initial clustering result does not withstand correction for patient batch structure, and that the initial signature-based score reflects proliferation rather than resistance biology, formally confirmed by pathway enrichment. The pseudobulk analysis, the only check that preserved a directionally consistent signal, did not reach significance at this sample size. None of the four checks provide support for a reproducible pre-existing resistance signal in this cohort. These four checks are summarized in Figure 4.

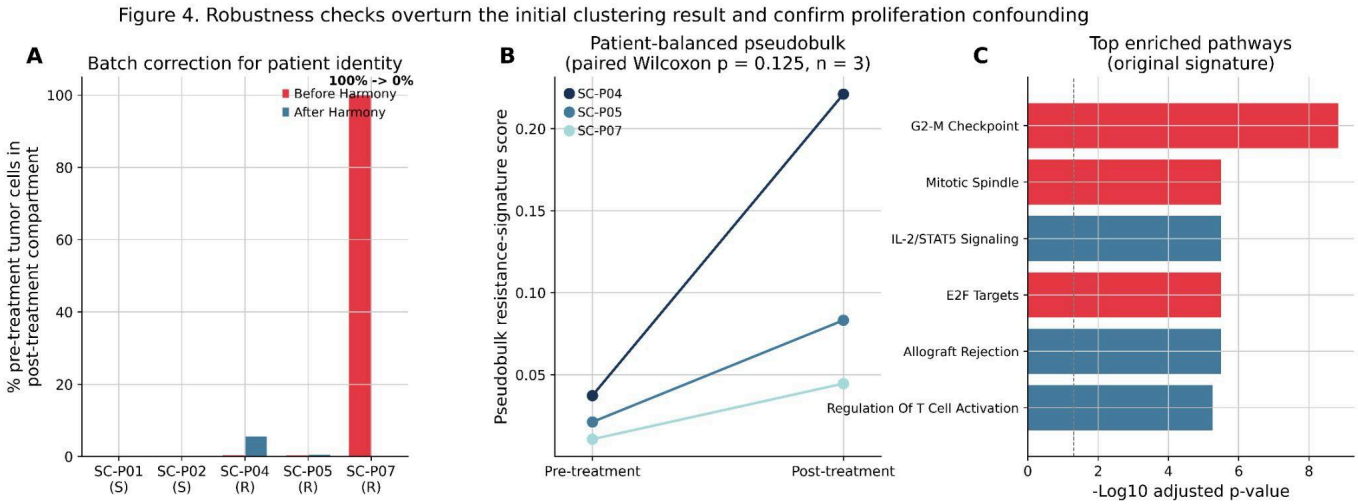

Figure 4. Robustness checks overturn the initial clustering result and confirm proliferation confounding. (A) Fraction of pre-treatment tumor cells within the post-treatment compartment, per patient, before and after Harmony batch correction for patient identity. Patient SC-P07's contribution fell from 100% to 0% after correction. (B) Patient-balanced pseudobulk resistance-signature score, per resistant patient, pre- versus post-treatment. All three patients show a consistent increase, though the paired one-sided Wilcoxon signed-rank test did not reach significance ($p = 0.125$, $n = 3$). (C) Top pathways enriched in the original 50-gene resistance signature by over-representation analysis (Enrichr), ranked by adjusted p-value. Bars in red indicate cell-cycle or mitotic pathways; bars in blue indicate immune-associated pathways. The dashed line marks the adjusted $p = 0.05$ threshold.

4. Discussion

This study set out to test whether chemotherapy resistant HR+ breast tumors show pre-treatment evidence of cell states resembling those that survive chemotherapy, and, more centrally, to subject any resulting signal to the kind of robustness testing that is inexpensive to perform but not yet standard in small single cell resistance studies. The initial analysis produced a superficially encouraging result: a directional association between resistance status and pre-treatment clustering with the post-treatment compartment ($p = 0.069$). Four independent checks, patient-level pseudobulk analysis, Harmony batch correction, cell-cycle regression, and gene set enrichment, were performed specifically to determine whether this result reflected reproducible biology. It did not survive that scrutiny. Harmony correction showed the result was driven almost entirely by patient-specific structure, not shared biology, since the one patient responsible for the positive signal fell from 100% to 0% compartment membership once patient identity was corrected for. The parallel signature-based score was in the opposite direction from the hypothesis throughout, and gene set enrichment formally confirmed that this signature reflects cell-cycle activity rather than a resistance-specific program. Only the pseudobulk analysis, which found a consistent direction of effect across all three resistant patients without reaching significance, offers any basis, and a weak one, for further investigation. We treat this combination of results as demonstrating that the initial finding does not constitute evidence of pre-existing resistance, rather than as motivation to select the most favorable of several discordant tests.

These observations bear on a longstanding question of whether chemoresistance is predominantly pre-existing or induced by treatment. Under the induced-resistance model, therapy actively reprograms surviving tumor cells into a resistant state; the drug-tolerant persister literature, for example, describes reversible tolerance mediated by chromatin remodeling that arises under drug pressure (6). Under the pre-existing model, resistant cell states are present before therapy and are selected by it. The source dataset itself distinguished innate from acquired chemoresistance programs in these tumors on the basis of copy number, identifying gene sets consistent with both (2). The present analysis does not adjudicate between these models. Given the fragility of the clustering result described above, and the absence of any independent validation, the data here should not be read as evidence favoring a pre-existing component over a purely induced one; they illustrate the kind of pattern such a component might produce and the specific pitfalls, patient-level dependence and outcome-conditioned analytical choices among them, that make this retrospective design unable to distinguish the two possibilities.

The approach used here is not specific to HR+ breast cancer. Any cancer type in which paired pre- and post-treatment tumor samples can be obtained is amenable to the same design: identify the transcriptional state of tumor cells surviving treatment, then ask whether pre-treatment tumors of eventual non-responders already contain cells resembling that state. The clustering-based test in particular does not require choosing a signature and could be applied to paired samples from patients receiving chemotherapy for triple negative breast cancer, targeted therapy for lung adenocarcinoma with actionable driver mutations, or immune checkpoint blockade for melanoma. In each of these settings, the same computational workflow could be applied to ask the same question, though our results here caution that such a workflow needs a larger, independently validated cohort and a signature-independent, patient-balanced analysis before any resulting pattern could be considered a candidate biomarker rather than a retrospective artifact of one or two patients.

The gene set enrichment results in Section 3.4 formally confirm what the signature's gene composition suggested. The highest-ranked signature genes are canonical markers of cell proliferation and mitosis, TOP2A, MKI67, CDK1, CENPF, PTTG1, BIRC5, ASPM, TPX2, NUSAP1, and UBE2C, and the full 50-gene signature was significantly enriched for Hallmark G2-M Checkpoint (adjusted p = 1.4x10^-9), Mitotic Spindle, and E2F Targets pathways, all consistent with a cell-cycle rather than resistance-specific program. Because chemotherapy sensitivity in HR+ breast cancer is associated with higher tumor proliferation at baseline (chemotherapy preferentially kills cycling cells), a signature built from post-versus-pre differential expression in resistant patients alone can end up measuring a proliferation axis that happens to run in the opposite direction from resistance when applied to a cohort that also includes more proliferative sensitive tumors. A smaller set of immune-associated genes and pathways (Hallmark IL-2/STAT5 Signaling, GO Regulation of T Cell Activation) was also enriched but did not dominate the signature. This is a demonstration, not merely a suspicion, that a signature derived by this common method can reflect the biology of chemotherapy targets rather than the biology of chemotherapy resistance.

The central practical lesson of this study is that each of its four robustness checks would have been straightforward to omit, and that omitting any one of them would have left a materially more encouraging, but unsupported, conclusion in place. Omitting Harmony correction would have left patient SC-P07's batch-driven signal standing as apparent biology. Omitting gene set enrichment would have left the proliferation confound as an informal observation about gene

names rather than a statistically confirmed pathway signal. Omitting patient-level pseudobulk analysis would have left the original cell-level statistics, which treat thousands of correlated cells as independent evidence, unchallenged. Based on this case study, we suggest that small-cohort single cell studies proposing pre-treatment resistance or response signatures should, at minimum: (1) report hypothesis tests at the patient level, either through pseudobulk analysis or by explicitly stating how many independent patients support a claimed effect; (2) apply a batch-integration method such as Harmony, scVI, or BBKNN and report whether clustering-based claims survive correction for patient identity; and (3) subject any candidate signature to pathway-level enrichment analysis rather than reporting a gene list alone, so that claims of a specific biological mechanism are tied to a reproducible pathway signal rather than to gene names that may reflect a generic process such as proliferation. These checks are inexpensive relative to the single cell sequencing itself and, as shown here, can change the paper's conclusion entirely.

4.1 Limitations

This analysis has several specific limitations that bear directly on how its results should be interpreted. First, the effective unit of inference is the patient, not the cell: with only three resistant and two sensitive patients, thousands of cells per patient do not provide thousands of independent observations, and neither hypothesis test is adequately powered. Second, the study design is retrospective and post-hoc: the resistance signature was derived from patients whose outcomes were already known, and the same patients' pre-treatment samples were then examined for that signature, so the analysis tests retrospective self-consistency rather than independent predictive validity. Third, the tumor-versus-stromal classification threshold (Methods 2.5) was chosen partly using the expected direction of treatment response, which means an outcome-informed step feeds into the downstream resistance comparison; an outcome-independent classifier such as inferred copy-number variation would avoid this dependency. Fourth, the clustering compartment result ($p = 0.069$) is driven almost entirely by one resistant patient contributing six pre-treatment tumor cells, all of which fell in the post-treatment compartment; the other two resistant patients showed values (0.31%, 0.28%) indistinguishable from the sensitive patients (0.00% each). This result should not be described as a recurring resistant-cell phenotype across resistant patients, and we do not consider it more reliable than the signature-based test simply because it does not depend on gene choice. Fifth, an attempt to isolate a non-proliferation resistance signal by regressing out cell-cycle scores (Section 3.4) was limited by the prior restriction of the dataset to the 2,000 most highly variable genes, which retained only 3 of 43 canonical S-phase and 13 of 54 canonical G2M-phase marker genes; the resulting corrected signature lacked coherent biological composition and is therefore inconclusive rather than informative about a non-proliferation resistance program. A dataset processed without this variable-gene restriction, or with cell-cycle genes explicitly retained regardless of variability, would be needed for a reliable cell-cycle-controlled analysis. Sixth, patient-level batch correction (Harmony, Section 3.4) was applied and did not support the clustering result: the patient responsible for the positive signal showed 0% post-treatment compartment membership after correction, compared with 100% before correction, indicating the original result likely reflected patient-specific structure rather than shared biology. Seventh, no independent validation cohort was used: the discovery and evaluation of both metrics occurred within the same five patients, and no frozen signature or classifier was tested on pretreatment samples from patients not used in discovery. A genuinely predictive test would

require estimating a resistance score on new pretreatment samples before examining outcome, and evaluating performance with patient-level ROC-AUC, sensitivity, specificity, and confidence intervals; this was not attempted here because no suitable independent cohort was identified. We searched public repositories for an independent pretreatment single cell cohort in HR+ breast cancer with known chemotherapy response and did not find one; to our knowledge this specific single-cell resource does not yet exist publicly, which is itself a constraint on this field rather than a gap unique to this study. Independent bulk RNA-seq or microarray cohorts with known pathological response do exist (for example GSE25066, GSE20194, GSE32646, GSE123845, and GSE163882) and could provide orthogonal validation of a frozen gene score in future work; however, because bulk measurements average expression across an entire tumor, they cannot directly test the rare-subpopulation hypothesis examined here, and any association found in bulk data would need to be interpreted as a tumor-average signal rather than confirmation of a rare pre-existing cell state. Eighth, no lineage tracing, clonal, or mutational information was available to establish that transcriptionally similar pre- and post-treatment cells are clonally related, so transcriptional resemblance cannot be interpreted as evidence that specific pre-treatment cells gave rise to the surviving post-treatment population. Finally, this study makes only a computational, associative claim; establishing a causal, mechanistic link between any candidate resistance program and chemotherapy sensitivity would require isolating or generating cells expressing the proposed program, exposing them to the relevant chemotherapy regimen in vitro, and testing whether perturbing a candidate driver gene or pathway alters sensitivity; this kind of experimental validation is outside the scope of the present computational reanalysis and is an important direction for future collaborative work. Taken together, these limitations mean this analysis should be read as identifying candidate patterns and concrete methodological pitfalls for future studies, not as evidence of a validated, reproducible pre-existing resistance state.

5. Conclusion

HR+ breast cancer is the most common breast cancer subtype, and although neoadjuvant chemotherapy is a mainstay of treatment, a substantial fraction of HR+ tumors respond poorly, and the transcriptional basis of that failure remains incompletely understood. This study conducted an exploratory, retrospective reanalysis of a paired pre- and post-chemotherapy single cell RNA sequencing dataset from five HR+ breast cancer patients to ask whether resistant tumors show any evidence, before treatment ever begins, of cells transcriptionally resembling those that survive chemotherapy.

An initial signature-independent clustering test showed a directional association in the hypothesized direction ($p = 0.069$), while a parallel signature-based scoring test showed the opposite direction. Rather than treating this discordance as a coin flip to be resolved in favor of the more encouraging result, four robustness checks were applied to determine which, if either, reflected reproducible biology. Harmony batch correction for patient identity showed that the clustering result was driven almost entirely by one patient, whose contribution fell from 100% to 0% compartment membership after correction, and the corrected test no longer supported the hypothesis ($p = 0.1665$). Gene set enrichment analysis formally confirmed that the score-based signature reflected cell-cycle activity (Hallmark G2-M Checkpoint, adjusted $p = 1.4x10^{-9}$) rather than a resistance-specific pathway. A patient-balanced pseudobulk reanalysis found a consistent direction of effect across all three resistant patients but did not reach significance ($p =$

0.125, n = 3). None of these checks provide convincing evidence for pre-existing resistance in this cohort.

This analysis does not establish that pre-existing resistant tumor cell states exist or can be detected before treatment in this cohort; the sample size, retrospective design, outcome-informed classification step, and the results of the four robustness checks preclude that conclusion. Its value instead lies in demonstrating, on real data, how specific and inexpensive checks can distinguish a reproducible biological signal from an analytical artifact, and in showing that omitting any one of these checks would have left a more encouraging but unsupported conclusion in place. Small-cohort single cell studies proposing pre-treatment resistance or response signatures should, at minimum, report hypothesis tests at the patient level rather than the cell level, apply a batch-integration method and report whether clustering-based claims survive patient correction, and subject any candidate signature to pathway-level enrichment analysis rather than reporting a gene list alone. Applied together, as in this study, these checks can overturn a result that otherwise appears statistically and biologically plausible, and we recommend they become standard practice before such signals are reported as candidate biomarkers.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 71: 209-249, 2021. https://doi.org/10.3322/caac.21660
2. Jia F, Sun S, Li J, Wang W, Huang H, Hu X, et al. Neoadjuvant chemotherapy-induced remodeling of human hormonal receptor-positive breast cancer revealed by single-cell RNA sequencing. Cancer Lett, 585: 216656, 2024. https://doi.org/10.1016/j.canlet.2024.216656
3. Dent SF, Cortes J, Bedard PL, Ellis P, Perez EA, Gelmon KA. Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy. Breast Cancer Res Treat, 170: 559-567, 2018. https://doi.org/10.1007/s10549-018-4801-3
4. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature, 501: 328-337, 2013. https://doi.org/10.1038/nature12624
5. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature, 541: 331-338, 2017. https://doi.org/10.1038/nature21350
6. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell, 141: 69-80, 2010. https://doi.org/10.1016/j.cell.2010.02.027
7. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol, 19: 15, 2018. https://doi.org/10.1186/s13059-017-1382-0
8. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv, 1802.03426, 2018. https://doi.org/10.48550/arXiv.1802.03426
9. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep, 9: 5233, 2019. https://doi.org/10.1038/s41598-019-41695-z

10. Dominguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science, 376: eabl5197, 2022. https://doi.org/10.1126/science.abl5197
11. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352: 189-196, 2016. https://doi.org/10.1126/science.aad0501
12. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat, 18: 50-60, 1947. https://doi.org/10.1214/aoms/1177730491
13. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods, 16: 1289-1296, 2019. https://doi.org/10.1038/s41592-019-0619-0
14. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res, 44: W90-W97, 2016. https://doi.org/10.1093/nar/gkw377