A065 · Discovering Biological Signals in the Noise

Presenter: Sophia K. Cheng — University of Michigan

Authors: Sophia K. Cheng

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Like astronomers who have only ever known a star by its light, biologists studying individual cells face a fundamental observational constraint – they cannot directly witness a cell's decisions, only interpret its signals. Single-cell RNA sequencing provides researchers with a snapshot of individual cells from which they can infer what the cell is, what it is doing, and what it might do next. This research focuses on the preprocessing that occurs before analysis begins. Current standards flag cells as noise to be filtered out if they fall outside any single quality control threshold. From a data quality standpoint, this is sound practice. However, the thresholds are often defined using the same set of assumptions that are being analyzed, introducing causal ambiguity. To investigate this further, I performed my own analysis on the discarded noise data from a high-impact 2021 paper (Pal et al., 2021). Using their provided supplementary data, I separated the samples into cells that would be discarded (noise) or analyzed (real). I identified 3 potential biological signals (PBS) when analyzing the latent space of the noise. Using these signals, I then built a decision tree classifier on the noise cells. The classifier has a high sensitivity to ER+ tumors (91.6%) with a moderate overall accuracy (85.4%). This research provides preliminary evidence that there are signals in what is discarded. Investigating these cells more carefully may matter most in cancers where heterogeneity is highest and treatment resistance is hardest to explain, such as breast cancer.

Discovering Biological Signals in the Noise

Sophia K. Cheng*, MS, MADS

School of Information, University of Michigan, Ann Arbor, MI

Like astronomers who have only ever known a star by its light, biologists studying individual cells face a fundamental observational constraint – they cannot directly witness a cell's decisions, only interpret its signals. Single-cell RNA sequencing provides researchers with a snapshot of individual cells from which they can infer what the cell is, what it is doing, and what it might do next.

This research focuses on the preprocessing that occurs before analysis begins. Current standards flag cells as noise to be filtered out if they fall outside any single quality control threshold. From a data quality standpoint, this is sound practice. However, the thresholds are often defined using the same set of assumptions that are being analyzed, introducing causal ambiguity.

To investigate this further, I performed my own analysis on the discarded noise data from a high-impact 2021 paper (Pal et al., 2021). Using their provided supplementary data, I separated the samples into cells that would be discarded (noise) or analyzed (real). I identified 3 potential biological signals (PBS) when analyzing the latent space of the noise. Using these signals, I then built a decision tree classifier on the noise cells. The classifier has a high sensitivity to ER+ tumors (91.6%) with a moderate overall accuracy (85.4%).

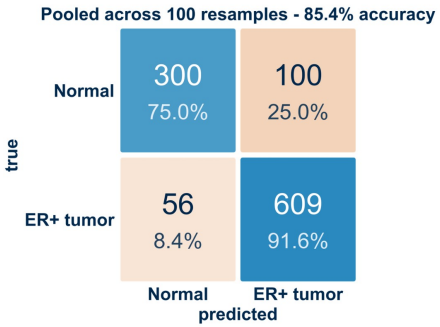

Figure 1. Pooled confusion matrix of the decision tree classifier across 100 resamples (85.4% overall accuracy). Cells are labeled with counts and row-wise percentages; sensitivity to ER+ tumor is 91.6% (609/665).

This research provides preliminary evidence that there are signals in what is discarded. Investigating these cells more carefully may matter most in cancers where heterogeneity is highest and treatment resistance is hardest to explain, such as breast cancer.

References

Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, Bryant VL, Penington JS, Di Stefano L, Tubau Ribera N, Wilcox S, Mann GB; kConFab; Papenfuss AT, Lindeman GJ, Smyth GK, Visvader JE. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. EMBO J. 2021 Jun 1;40(11):e107333. doi: 10.15252/embj.2020107333. Epub 2021 May 5. PMID: 33950524; PMCID: PMC8167363.