## Results

### Cellular Composition of Breast Cancer Samples

We first characterized the cellular landscape of the METABRIC breast cancer dataset, which comprised 10 samples containing 19,169 cells and 39 protein markers. Our analysis revealed a heterogeneous cellular composition, with tumor cells predominating across all samples (Figure 1). While tumor cells constituted the majority, other cell types such as Basal Cells, Endothelial Cells, Fibroblasts, Hypoxia-Related, Immune Cells, and Myoepithelial Cells were present in varying proportions.

### Identification of Tumor-Specific Protein Markers

To distinguish tumor cells from other cell types, we employed a rigorous Wilcoxon test to identify markers with significantly higher expression in tumor cells. At a stringent significance threshold (adjusted p-value < 0.01), we identified 12 protein markers that were differentially expressed (Table 1).

The top markers with the most significant differential expression included:

* GATA3 (adj. p-value = 0)
* panCK (adj. p-value = 0)
* CK8\_18 (adj. p-value = 1.18 × 10^-289)
* PR (adj. p-value = 6.50 × 10^-136)
* Ki67 (adj. p-value = 2.44 × 10^-65)

These markers exhibited varying fold changes in tumor cells compared to other cell types, with Ki67 showing the most substantial difference (fold change = 2.15).

### Machine Learning-Based Tumor Cell Classification

We developed a Random Forest classifier using the 12 identified markers to predict tumor cell identification. The model demonstrated robust performance:

* Cross-validation accuracy: 0.894
* Cohen's Kappa: 0.787, indicating substantial agreement beyond chance

Feature importance analysis revealed the most critical markers for tumor cell identification (Figure 2):

1. Ki67
2. GATA3
3. CK8\_18
4. panCK
5. PR

### Validation on Independent Dataset

When applied to an independent in-house breast cancer IMC dataset, our model showed promising predictive capabilities:

* Overall accuracy: 0.815
* Sensitivity: 0.596
* Specificity: 0.927

The confusion matrix indicated that while the model performed well in identifying non-tumor cells, there was moderate variability in tumor cell prediction, suggesting the complexity of cellular heterogeneity in breast cancer samples.

### Biological Implications

The identified markers, particularly Ki67, GATA3, and CK8\_18, represent potential key molecular signatures for breast cancer tumor cell characterization. Ki67, a proliferation marker, and GATA3, a transcription factor important in mammary gland development, align with known breast cancer biology, underscoring the biological relevance of our computational approach.