## Results

### Characterizing Cell Type Composition in the METABRIC Breast Cancer IMC Dataset

We began our analysis by examining the METABRIC Imaging Mass Cytometry (IMC) breast cancer dataset, comprising 19,169 cells from 10 patient samples annotated with 39 protein markers. These cells were categorized into seven major cell types: tumor cells, basal cells, endothelial cells, fibroblasts, hypoxia-related cells, immune cells, and myoepithelial cells. A heatmap summarizing cell-type composition per sample revealed that tumor cells were consistently the most abundant population, while myoepithelial and hypoxia-related cells were notably rare across all samples (Figure 1).

**Figure 1**: Heatmap of cell-type counts across METABRIC samples. Tumor cells dominate across all samples; other cell types show variable and often low abundance.

### Identification of Tumor-Enriched Protein Markers

To identify tumor-specific markers, we conducted Wilcoxon rank-sum tests comparing marker expression levels between tumor and non-tumor cells. After adjusting for multiple testing (Benjamini-Hochberg, FDR < 0.01), we identified 12 protein markers significantly upregulated in tumor cells. These included canonical epithelial and proliferation markers such as **GATA3**, **panCK**, **CK8\_18**, **PR**, **Ki67**, and **HER2**, all of which have known associations with breast cancer pathology [1–3].

Notably, **Ki67** (FC = 2.15), a hallmark proliferation marker, exhibited the highest fold change among significant proteins, underscoring its role in distinguishing highly proliferative tumor populations. The full list of significant markers and their differential expression statistics is provided in Supplementary Table 1.

### Tumor Cell Classification Using a Random Forest Model

Using the 12 tumor-enriched markers, we trained a Random Forest classifier on the METABRIC dataset to distinguish tumor from non-tumor cells. Five-fold cross-validation yielded a high average accuracy of **89.4%** (Kappa = 0.787), indicating that this marker panel is both robust and generalizable for binary classification of tumor cells.

To interpret marker contributions, we computed variable importance scores from the trained model (Figure 2). **Ki67**, **GATA3**, **CK8\_18**, and **panCK** emerged as the top contributors, consistent with their known roles in breast cancer biology. Ki67’s prominence highlights its critical utility in tumor identification, aligning with its clinical use as a proliferation index [4].

**Figure 2**: Feature importance plot from the Random Forest model. Top markers include Ki67, GATA3, CK8\_18, and panCK.

### Validation on an Independent In-House IMC Dataset

To validate the predictive capacity of the trained classifier, we applied it to an independent in-house IMC breast cancer dataset comprising 10 additional samples. After normalizing the data using the same transformation pipeline, we extracted expression levels for the 12-marker panel and predicted cell identities using the trained Random Forest model.

Performance evaluation against manual annotations yielded an overall classification accuracy of **81.5%**, with a specificity of **92.7%** and a sensitivity of **59.6%** (Table 1). While tumor cells were identified with high precision (positive predictive value = 80.7%), the reduced sensitivity suggests the model occasionally misclassifies tumor cells with atypical expression patterns as non-tumor. This is expected given the biological heterogeneity of tumor phenotypes and the limited marker set.

**Table 1**: Confusion matrix for tumor classification on the in-house dataset

|  | Reference: Other | Reference: Tumor |
| --- | --- | --- |
| **Pred: Other** | 3,786 | 908 |
| **Pred: Tumor** | 2,572 | 11,556 |

### Minimal Marker Panel for Tumor Cell Identification

Taken together, these results suggest that a relatively small panel of 12 protein markers, particularly those linked to proliferation (Ki67), luminal identity (GATA3, CK8\_18), and epithelial status (panCK, PR), is sufficient to achieve robust classification of tumor cells in breast cancer IMC data. This reduced panel may be broadly applicable for computational cell annotation in future IMC-based clinical and research studies, with potential for streamlining multiplex panel design.

## References

1. Liu, S. et al. GATA3 and the Regulation of the Luminal Epithelial Cell Lineage in Mammary Gland. Cell Res (2015).
2. Kreike, B. et al. Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. Breast Cancer Res (2007).
3. Nielsen, T. O. et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res (2004).
4. Yerushalmi, R. et al. Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol (2010).