## Results

### Cell-Type Composition of METABRIC Samples

A total of 10 METABRIC breast cancer samples were profiled by imaging mass cytometry (IMC), comprising 19,169 cells and 39 markers. Seven major cell types were annotated based on marker expression: Basal Cells, Endothelial Cells, Fibroblasts, Hypoxia-Related Cells, Immune Cells, Myoepithelial Cells, and Tumor Cells. As illustrated by the heatmap of cell counts (Figure 1), Tumor Cells were the predominant population in all 10 samples, whereas Myoepithelial and Hypoxia-Related subpopulations were comparatively rare. These findings are in line with previous reports of breast tumor heterogeneity where malignant epithelial cells usually represent the largest fraction within the tumor microenvironment (Ali et al., 2020).

### Identification of Tumor-Enriched Protein Markers

To pinpoint protein markers uniquely or strongly associated with tumor cells, a Wilcoxon test was conducted on the normalized intensities in the METABRIC dataset. Among the 39 proteins, 12 markers exhibited significantly higher expression (adjusted p < 0.01) in Tumor Cells compared with all other cell types: **GATA3, CK8/18, panCK, PR, Ki67, E\_cadherin, HER2, CK7, Sox9, p53, CD20**, and **CK5**. Notably, several of these proteins—such as Ki67, GATA3, panCK, and CK8/18—are well-established markers of tumor cell proliferation, luminal differentiation, or epithelial identity (Gown, 2008; Mehta et al., 2020). For instance, Ki67 is a hallmark of actively cycling cells and a routine clinical marker for tumor proliferation status in breast cancer. Likewise, GATA3 is commonly expressed in luminal breast tumors and plays a fundamental role in maintaining luminal epithelial differentiation (Chou et al., 2010).

### Random Forest Model Trained on the METABRIC Cohort

Using the 12 tumor-enriched markers as predictive features, we trained a binary Random Forest (RF) classifier to distinguish Tumor Cells from all non-tumor populations (basal, endothelial, fibroblast, immune, and other rare cell types) in the METABRIC dataset. Five-fold cross-validation on 19,169 cells yielded an **accuracy of 0.894** and a **kappa of 0.787**, indicating robust discrimination between tumor and non-tumor cells. Examination of variable importance scores (Figure 2) revealed that **Ki67** was the single most informative marker, followed closely by **GATA3**, **CK8/18**, **panCK**, and **PR**. These proteins align closely with canonical luminal markers (GATA3, PR), epithelial cytokeratins (CK8/18, panCK), and proliferation drivers (Ki67), underscoring their collective significance in breast tumor biology (Curtis et al., 2012).

### Classifier Performance on In-House IMC Samples

To assess the clinical and experimental generalizability of this marker panel, we applied the trained Random Forest classifier to 10 in-house breast cancer IMC samples (test set). Here, the model achieved an overall **accuracy of 0.815**. The **specificity** (ability to correctly classify non-tumor cells) was notably high at 0.927, indicating that only a small proportion of stromal or immune cells were misidentified as malignant. The **sensitivity** (correct identification of tumor cells) was 0.596, suggesting that a subset of tumor cells was misclassified as non-tumor. This lower sensitivity may reflect intratumoral heterogeneity, differential marker expression across distinct breast cancer subtypes, or variations in tissue staining procedures between the METABRIC and in-house cohorts (Ali et al., 2020).

### Summary of Key Findings

Overall, these results demonstrate that a condensed 12-marker panel—enriched for proliferation (Ki67), luminal differentiation (GATA3, PR), and epithelial lineage (CK8/18, panCK)—can reliably distinguish tumor cells from other cell populations in breast IMC datasets. While specificity was exceptionally high in both training and test cohorts, sensitivity was somewhat lower in the in-house dataset, indicating that further refinement or subtype-specific adjustments (e.g., basal-like vs luminal tumors) may be warranted. Collectively, this study highlights a focused set of breast cancer markers that recapitulate well-known hallmarks of malignancy and can serve as a foundation for systematic, reproducible tumor cell identification in imaging mass cytometry experiments (Ali et al., 2020; Curtis et al., 2012).

**References**

* Ali HR, Jackson HW, Zanotelli VRT, et al. (2020) Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. **Nat Cancer** 1: 163–175.
* Chou J, Provot S, Werb Z. (2010) GATA3 in development and cancer differentiation: Cells GATA have it! **J Cell Physiol** 222(1): 42–49.
* Curtis C, Shah SP, Chin SF, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. **Nature** 486(7403): 346–352.
* Gown AM. (2008) Current issues in ER and HER2 testing by IHC in breast cancer. **Mod Pathol** 21: S8–S15.
* Mehta A, Zhang L, Boufraqech M, et al. (2020) Understanding cancer signaling by single-cell proteomics: Current advances and future directions. **Proteomics** 20(5–6): 1900226.

**Figure Legends**

* **Figure 1.** Heatmap of Cell Counts per Sample in the METABRIC Cohort. Tumor cells were the most abundant population across all 10 samples, whereas myoepithelial and hypoxia-related cells were comparably underrepresented.
* **Figure 2.** Variable Importance Scores from the Random Forest Model. Ki67, GATA3, CK8/18, panCK, and PR were identified as the top five most informative markers for discriminating tumor and non-tumor cells.