## Results

### Data Exploration and Preprocessing

The dataset, a subset of the METABRIC breast cancer cohort, comprised imaging mass cytometry data from 10 samples, encompassing 19,169 cells and 39 markers. Seven major cell types were represented within the dataset. Prior to analysis, the raw intensity data was normalized using a combination of 'trim99', 'minMax', and PC1 methods to ensure comparability across cells and markers.

Initial exploration of cell type distribution revealed that tumor cells were the predominant cell type across all samples. Other cell types, including Myoepithelial Cells and Hypoxia-Related cells, were present at lower frequencies. The distribution of cell types across samples is shown in Figure 1, demonstrating the variability in cellular composition between samples.

### Identification of Tumor-Associated Markers

To identify markers with elevated expression in tumor cells, we performed Wilcoxon tests for each marker. Using a stringent adjusted p-value threshold of <0.01, we identified 12 protein markers that exhibited significantly higher expression in tumor cells compared to other cell types. These markers (GATA3, panCK, CK8\_18, PR, Ki67, E\_cadherin, HER2, CK7, Sox9, p53, CD20, and CK5) represent potential key factors in tumor cell biology. The full list of significant markers and their statistical significance is presented in Table 1.

### Random Forest Model for Tumor Cell Classification

To evaluate the predictive power of the 12 identified tumor-associated markers, we trained a Random Forest model. The METABRIC dataset was used as the training set, with cells categorized into "Tumor" and "Other" (a combination of Basal, Endothelial, Fibroblasts, Hypoxia-Related, Immune, and Myoepithelial cells).

The Random Forest model demonstrated strong performance in cross-validation, achieving an accuracy of 0.894 and a Kappa of 0.787 (Table 2). Feature importance analysis revealed that Ki67, GATA3, CK8\_18, panCK, and PR were the top five most important markers for tumor cell identification (Figure 2). This aligns with the known roles of Ki67 as a proliferation marker, GATA3 in luminal breast cancer, and CK8/18 and panCK as epithelial markers.

### Prediction of Tumor Cells in In-House Data

The Random Forest model was then applied to predict tumor cells within an independent in-house breast cancer IMC dataset. On this independent dataset, the model achieved an overall accuracy of 0.815, with a sensitivity of 0.596 and a specificity of 0.927 (Table 3).