Machine Learning in Computational Biology Final Project Report

Psallidas Kyriakos, Alvanakis Spyros

[1]Department of Computer Science and Telecommunications, National and Kapodistrian University of Athens

Abstract

In this study, we have successfully replicated and built upon the analysis conducted in a highly influential paper in the field of breast cancer single-cell RNA sequencing (scRNA-seq). Our research not only validates the feasibility of distinguishing between tumor and immune cells at the single-cell level but also sheds light on their characterization using targeted gene sets and unsupervised machine-learning techniques. These findings hold significant promise for advancing personalized medicine strategies in breast cancer. To enhance our analysis, we developed a preprocessing pipeline tailored to our specific research goals. By employing distinct and well-evaluated clustering approaches at each stage of cell separation, we aimed to optimize the accuracy and reliability of our results. Moreover, we incorporated advanced visualization techniques such as UMAP and t-SNE. These additions provided invaluable insights into the underlying organization and relationships within the data, enriching our overall analysis.

Introduction

Breast cancer displays significant heterogeneity, which implies that tumors can manifest genetic and molecular variations across various molecular subtypes such as Liminal-A, Luminal-B, HER2-positive, and triple-negative breast cancer (TNBC). Additionally, this heterogeneity can also be observed within individual tumors themselves. Single-cell RNA sequencing (scRNA-seq) is a powerful technique that allows researchers to examine the gene expression profiles of individual cells. Since such a method can provide a precise separation and characterization of patient cells and results in therapeutic targets. By enabling the precise separation and characterization of individual patient cells, this technique provides a comprehensive view of the cellular composition and heterogeneity within tumors. Through this approach, the researchers aimed to identify novel therapeutic targets and unravel the underlying mechanisms driving tumor progression and treatment resistance. The research project encompassed a diverse group of breast cancer patients from the highly cited paper "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer" [1], representing various molecular sub-types, including Luminal-A, Luminal-B, HER2-positive, and TNBC. While bulk genomic profiling is commonly employed to describe tumors in patients, cancer cells show variability within the tumor itself which could impact the effectiveness of a personalized treatment approach. By focusing on the transcriptome of individual cells, the researchers aimed to clarify the gene expression patterns that define the distinct cells in each molecular sub-type, thereby enhancing the understanding of the molecular drivers that contribute to tumor heterogeneity

The analysis extension proposed in this report, encompassing clustering analysis and dimensionality reduction, holds the potential for advancing our understanding of breast cancer heterogeneity. By critically evaluating the effectiveness of the clustering approach used in the original paper across multiple tasks, we can identify areas for improvement and enhance the accuracy of cell distribution, immune cell grouping, and pathway activation categorization. Additionally, by exploring alternative dimensionality reduction methods, such as UMAP and t-SNE, we can overcome the limitations of linear assumptions and discover intricate relationships within the high-dimensional gene expression data. In this report, we will present a comprehensive overview of our replication methodology, outlining the steps taken to reproduce the RNA-seq analysis and the enhancements introduced through visualization methods and clustering techniques.

Methodologies

Data Set description

The referenced paper "Single-cell RNA-seq enables comprehensive tumor and immune cell profiling in primary breast cancer" [1] investigates a data set that comprises transcription counts per million (TPM) measurements. The data set encompasses 550 cells and 57,915 genes, which are derived from 10 breast cancer patients, including one patient who underwent treatment with the drug Herceptin. Furthermore, regional metastatic lymph nodes were collected

from two patients, resulting in a total of 13 tumor groups. These tumor groups correspond to four distinct breast cancer molecular subtypes: Luminal-A (ER+, PR+/HER2-), Luminal-B (ER+/HER2-), HER2 (ER-/HER2+), and Triple-negative or basal-like breast cancer (TNBC) (ER-/HER2-) [2]. The dataset also contains the 13 pooled samples for each tumor group.

It is noteworthy that the distribution of both cell numbers and subtype populations across the data set is uneven. Figure 1 visually depicts this uneven distribution. Patients BC01 and BC02 are associated

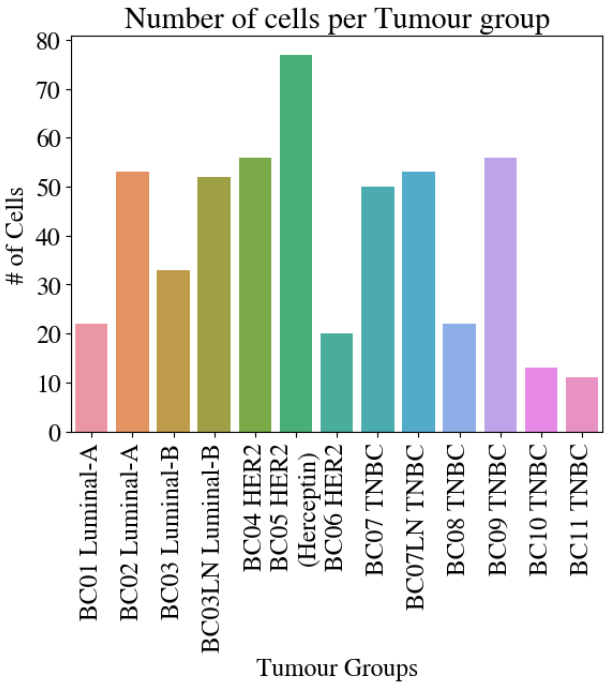


Figure 1: *Cell counts per tumor group*

with Luminal-A subtypes, with BC02 having approximately 30 more cells than BC01. Patient BC03, along with BC03 Lymph Nodes (LN), represents the sole instance of the Luminal-B subtype. Patients BC04 and BC06 belong to the HER2 subtype, while BC05 denotes the HER2 patient treated with Herceptin. BC05 exhibits a significantly larger number of cells compared to BC04, whereas BC04 contains more cells than BC06. Lastly, the TNBC subtype is represented by patients BC07, BC09, BC08, BC10, and BC11. BC07 and BC09 each have approximately 50 cells, while the latter three patients possess around 10-20 cells. Additionally, we have obtained a second dataset that includes the cell separation labels extracted from the reference paper analysis. This dataset will be utilized for evaluating the outcomes of our study.

Data Preprocessing

To begin, we preprocessed the dataset by eliminating the gene id and gene type columns. We then set the gene names as the index of the data frame and transposed it. Afterward, we converted the data frame into an Anndata object [3], where the observations corresponded to cell IDs and the variables represented the gene names. Subsequently, we divided the dataset into two subsets: pooled samples and single cells. To the single cells subset, we appended a tumor population tag (e.g., BC01) to each cell of patient BC01.

Cell filtering followed our own pipeline which consists of the calculation of three distributions: the log-transformed overall count of all gene expression in the cells, the log-transformed number of genes expressed, and the percentage of the total expression of the 20 most expressed genes for all cells. These distributions are showcased in Figure 14. We then removed cells that are five absolute deviations from the median in any of the above covariates.

$$MAD = median(|X_i - median(X)|)$$

This resulted in 518 remaining cells after prepossessing, compared to the researchers' 515.

Gene filtering was conducted using the preprocessing pipeline established in the reference paper. Initially, all genes with expression levels below 1 TPM (Transcripts Per Million) were set to 0 expression. Subsequently, the expression values were transformed using a logarithm base 2 after adding a value of 1. The inclusion of an absent normalization step was deemed crucial by us for the subsequent dimensionality reduction techniques we will employ. Finally, genes expressed in 10% or less of all tumor groups were removed from the analysis. The latter proved to be challenging to interpret, thus we utilized two different methodologies.

With the initial method, the expression counts were categorized based on tumor group values. The expression values were then converted into binary values (0 and 1) and added together. As a result, genes with a cumulative binary expression value lower than $13 \times 0.1 = 1.3 \approx 1$ were eliminated from both the single cell and the pooled dataset. This resulted in a remaining 27461 genes compared to the 17779 remaining genes of the researchers.

The second method incorporated the removal of genes that are expressed in 10% of the cells of each tumor group for all tumor groups. The fraction of cells expressing each gene for each tumor group was calculated utilizing scanpy's [4] `rank_genes_groups` as illustrated in Figure 2. Following the application of this method, 17,851 genes were retained, which is very close to the 17,779 genes reported by the researchers. However, since there was a substantial difference in the unique genes between the two sets,

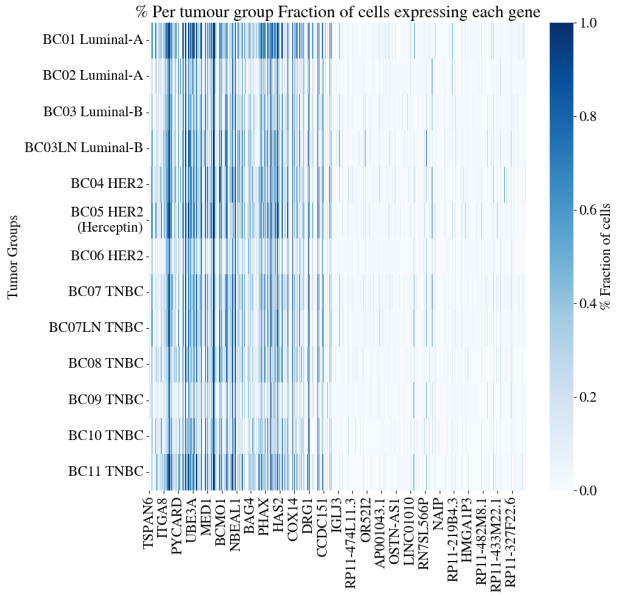


Figure 2: *Percentage of cells expressing each gene for each tumor group*

we opted for method one which besides the inclusion of more genes, it included those deemed important for the subsequent stages of the analysis.

Dimensionality Reduction & Initial Visualization

We utilized the following *scGmix* class [Assignment 3] method:

```
scGmix.dimreduction(\
pc_selection_method="screeplot")
```

to reduce the principal components to 9. This reduction was performed on a dataset consisting of 518 cells and 27,664 genes obtained through preprocessing. To visualize the data, we employed PCA on the resulting 9 principal components. Additionally, we employed the `scgmix(adata=adata,rand_seed=42,method="TSNE")` to determine the optimal parameters for t-SNE visualization. For UMAP visualization, we conducted extensive parameter experiments, specifically focusing on min_dist and n_neighbors, as there is no established standard methodology for parameter selection in UMAP visualization. Ultimately, we settled on n_neighbors=15 and min_dist=0.75 as the most suitable parameters. Subsequently, using the `scgmix(adata=adata,rand_seed=42,method="UMAP")` function, we generated a visual representation using UMAP for cancer and non-cancer cells based on the 9 principal components.

It is important to note that in both TSNE and UMAP visualizations, the distances between clusters do not necessarily have a direct interpretation. However, the clusters formed provide valuable information about the local and global structure of the data. Based on these findings, we hypothesize that tumor cells exhibit distinct clusters due to their unique gene expression patterns, while non-tumor cells tend to cluster together within a larger cluster. This hypothesis will be further explored during the clustering stage, specifically focusing on the separation of cancer and non-cancer cells.

Cell Separation methods

Tumor and non-tumor cell separation with clustering was performed using a selection of four distinct clustering algorithms implemented through the scikit learn Python library [5]. These algorithms consist of Gaussian Mixture Models (GMMs), Agglomerative Ward linkage, Average linkage, and Spectral clustering. For the non-deterministic nature of GMMs and Spectral clustering, the optuna library was utilized for Bayesian hyperparameter tuning, employing the Tree Parzen estimator (TPE) sampler [6]. These algorithms were specifically chosen due to their effectiveness in scenarios involving a predefined number of clusters, which in our case is two. The clustering process was carried out on the PCA representation of the TPM counts dataset, utilizing 100 principal components. To evaluate the quality of the clustering results, we employed a combination of one internal and three evaluation metrics that make use of the reference paperers' cell separation labels. For the internal evaluation metric we utilized the silhouette score [7]:

$$s = \frac{b - a}{\max(a, b)}$$

For comparison of our clustering methodologies with the results of the reference paper we utilized the Adjusted for chance Rand Index (ARI),

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]}$$

the Adjusted for chance Mutual Information (AMI)

$$AMI(U, V) = \frac{MI(U, V) - E(MI(U, V))}{avg(H(U), H(V)) - E[(MI(U, V))]}$$

and the Fowlkes-Mallows index (FMI).

$$FMI = \frac{TP}{\sqrt{(TP + FP) \cdot (TP + FN)}}$$

The Silhouette Coefficient for each clustering is calculated by taking into account the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. A Silhouette Coefficient value of $S = 1$ indicates perfectly well-defined clusters that

do not overlap, whereas a value of $S = -1$ represents the opposite scenario. The remaining criteria fall within the range of [0, 1] and correspond, in order, to the agreement in labeling pairs adjusted for random permutations, mutual information adjusted for random permutations, and the geometric mean of precision and recall between the results obtained from the reference papers and our clustering algorithms.

Estimate Package & Immune, Stromal cell seperation We utilized the ESTIMATE[8] package from R to predict the immune score, stromal score, and tumor purity. These scores are important in further separating immune and stromal cells after separating carcinoma and non-carcinoma cells. Additionally, the utilization of the ESTIMATE package allowed us to observe noticeable differences between cancer and non-cancer cells, providing clearer insights into the results of cancer/non-cancer cell separation.

The ESTIMATE package is a tool for estimating tumor purity and determining the presence of infiltrating stromal and immune cells in tumor tissues using gene expression data. The algorithm of ESTIMATE is based on single-sample Gene Set Enrichment Analysis and generates three scores: the stromal score, which captures the extent of stroma in tumor tissue; the immune score, which represents the infiltration of immune cells in tumor tissue; and the estimate score, which provides an inference of tumor purity. By using these scores, we gain valuable insights into the composition of the tumor-predicted scores.

Correlation analysis

To validate the correlation results, Pearson's correlation coefficient was applied to both the mixed dataset and the separated predictions for carcinoma and non-carcinoma cells. The objective was to assess the cell-to-cell correlations within the tumor samples. Also allowed us to evaluate the linear expression correlations and provided insights into the relationships among different cell types

Methods employed for cancer cell analysis

ER & HER2 module scoring with the genefu R package In order to accurately predict the cancer subtype of each cell by assigning ER,HER2 module scores, the researchers employed the Genefu package from R. This package was chosen due to its comprehensive set of functions specifically tailored for gene expression analysis, particularly in breast cancer studies.

Utilizing the Genefu package required a structured data frame containing both gene and cell information, similar to the requirements of the ESTIMATE package. However, an additional requirement for the Genefu package was the usage of the Enterez Genes ID, which was crucial for its functionality. To acquire these gene IDs, we utilized the API of the National Center for Biotechnology Information (NCBI) with the assistance of BioPython. It is worth noting that this process proved to be time-consuming due to the large number of genes involved. Within the Genefu [9] package, we utilized the subtype.cluster function, which employs Gaussian Mixture Models with prior knowledge derived from three distinct datasets available in R. These datasets encompassed significant genes associated with ER-/HER2-, HER2+, and ER+/HER2- subtypes. By applying this function, probabilities were generated for each cell, indicating its likelihood of belonging to one of the three previously mentioned classes.

GSVA for Subtype-specific gene expression In order to validate the researchers' findings, we performed Gene Set Variation Analysis (GSVA) at a single-cell level to identify differentially expressed marker genes among the different subtypes of carcinoma cells. This analysis was specifically conducted on carcinoma cells rather than bulk tumor cells with marker gene-sets extracted from the supplementary material of the reference paper and Scnapy to plot the expression heatmap. Additionally, we included patient BC05, who underwent Herceptin treatment, to observe the overall gene expression patterns.

GSVA for Subtype-specific pathways In the subsequent step, we performed Gene Set Variation Analysis (GSVA) specifically for subtype-related pathways, focusing exclusively on cancer cells. We investigated six distinct pathways extracted from the Gene Set Enrichment Analysis (GSEA) online portal [10]:

- smid_breast_cancer_luminal_b_up: Upregulated genes associated with Luminal B breast cancer, providing insights into the molecular characteristics of this subtype.
- halmark_estrogen_response_early and halmark_estrogen_response_late: Genes defining early and late response to estrogen, offering insights into the temporal dynamics of estrogen signaling in breast cancer.
- nikolsky_breast_cancer_17q11_q21 _amplicon: Genes related to copy number alterations specific to breast tumors, highlighting genomic instability in breast cancer.
- smid_breast_cancer_basal_up: Upregulated genes associated with Basal cancer (an aggressive form of triple-negative breast cancer, TNBC).
- mid_breast_cancer_relapse_in _brain_up: Upregulated genes associated with brain relapse

in breast cancer, providing insights into the metastatic behavior of breast tumors to the brain.

GSVA and correlation for aggressive cancer gene sets To perform the Gene Set Variation Analysis (GSVA) on gene sets associated with aggressive cancer, we obtained gene sets related to Stemness, Angiogenesis, and Epithelial-Mesenchymal Transition (EMT) from the GSEA online portal. For our final gene sets, we employed an intersection approach between the genes in our dataset and each of the aforementioned gene sets. Through this process, we removed a total of 4 out of 36 genes from the Angiogenesis gene set, 8 out of 200 genes from the EMT gene set, and no genes from the Stemness gene set. We employed Pearson's correlation coefficient (r) to assess the linear correlation between the expression levels of each gene-set pair across all cells. By creating scatter plots and color-coding the expression scores based on distinct tumor groups, we were able to visualize and identify tumor group-specific as well as patient-specific patterns of aggressiveness.

Methods employed for immune cell analysis

Immune cell separation into subtypes was similar in methodology to the clustering and evaluation methodology described for tumor and non-tumor cell separation, we conducted clustering analysis and evaluation to categorize immune cells into T-cells, B-cells, and macrophages. However, in this case, we employed cell expression values specifically associated with gene sets related to T-cells, B-cells, and macrophages, which were obtained from the supplementary material of the reference paper. By applying the set intersection method outlined earlier, we identified that 2 out of 34 genes from the first gene set, 2 out of 19 genes from the second gene set, and 1 out of 33 genes from the third gene set needed to be removed for gene-set and data-set match. The clustering algorithms employed in this stage, are: Agglomerative Ward Linkage, Non-negative matrix factorization, Gaussian Mixture Models, and Spectral clustering. Moreover, utilizing the heatmap function in scanpy, we were able to visualize the expression values of the clustered cells for the three genesets.

Clustering of T-cell signature pathways presented a challenge due to the absence of labels in the reference dataset, preventing the use of comparison metrics such as AMI, ARI, and FMI. Consequently, we relied solely on the silhouette score to evaluate the clustering results. Our evaluation focused on categorizing T-cells into specific pathway signatures, using gene sets obtained from the reference paper, which included Costimulatory, Cytotoxic, Exhausted, Naïve, and Regulatory signatures. Zero genes were removed utilizing the gene-set, data-set intersection method. For this stage, we employed several clustering algorithms, namely Agglomerative Average, Single linkage, Complete linkage, and Spectral clustering.

Results & Discussion

Data-set dimensionality reduction & visualization

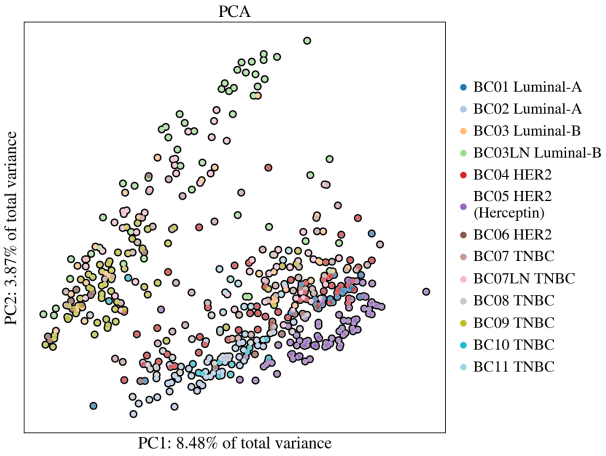


Figure 3: *PCA visualization of the TPM counts matrix.*

We employed dimensionality reduction techniques to reduce the number of principal components to 9, using the elbow rule as a guide. Initially, we visualized the data using PCA as seen in Figure 3 and observed that the first component accounted for 8.48% of the variance, an improvement over the researcher's initial value of 6.87%. The second component provided similar information. Through dimensionality reduction, we obtained visual results consistent with the researchers' findings, but with higher cumulative variance information.

UMAP and t-SNE revealed distinct clusters formed by cells of the same category. It is important to note that in both the UMAP 5 and t-SNE 4 plots, some cells appeared in two different regions. This occurrence is likely due to the combination of all the cell types in the input dataset. However, we anticipate that once we separate the carcinoma and non-carcinoma cells, the visualization results will significantly improve, providing clearer insights.

Cell separation

Figure 15 displays the t-SNE, UMAP, and PCA representations of the TPM counts matrix, with colors indicating the labels identified in the reference paper. The figures clearly illustrate a distinct visual

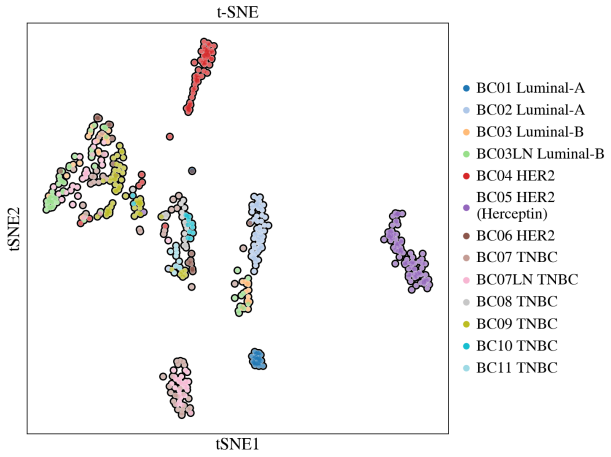


Figure 4: *T-SNE visualization of the TPM counts matrix.*

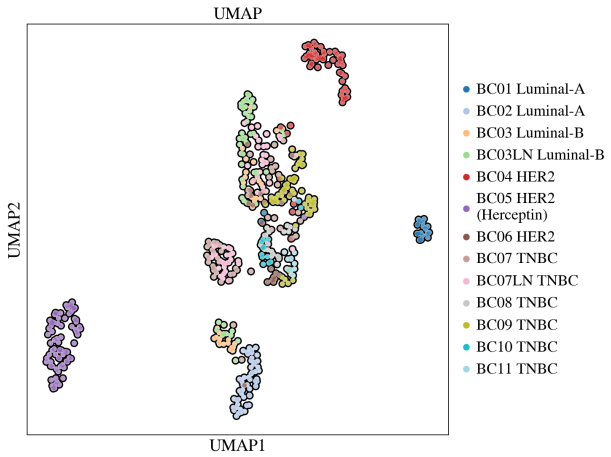


Figure 5: *UMAP visualization of the TPM counts matrix.*

cluster representing non-tumor cells, while the tumor groups appear as separate clusters in both t-SNE and UMAP spaces. In the 100-D PCA space, the silhouette coefficient score for the ground labels is calculated to be 0.134.

Among the clustering algorithms we employed, Agglomerative ward linkage, depicted in Figures 16(d), 16(e), and 16(f), closely resembled the researchers' labels and yielded the highest overall metrics. It achieved a silhouette score of 0.139, ARI of 0.892, AMI of 0.817, and FMI of 0.94. The spectral clustering algorithm, illustrated in Figures 16(j), 16(k), and 16(l), performed second best with a silhouette score of 0.143, ARI of 0.722, AMI of 0.615, and FMI of 0.867.

On the other hand, the GMMs and Average linkage methods, showcased in the remaining sub-figures of Figure 16, yielded much poorer results. The GMMs obtained a silhouette score of 0.02, ARI of 0, AMI of 0, and FMI of 0.514, while the Average linkage approach achieved a silhouette score of 0.233, ARI of 0, AMI of 0, and FMI of 0.72. A comprehensive visualization of these results is presented in the barplot of Figure 6 below. The reference paper mentions the usage of hierarchical clustering without specifying its type, however, based on the above information, we conclude that the authors employed the Agglomerative ward linkage method for tumor/non-tumor cell separation.

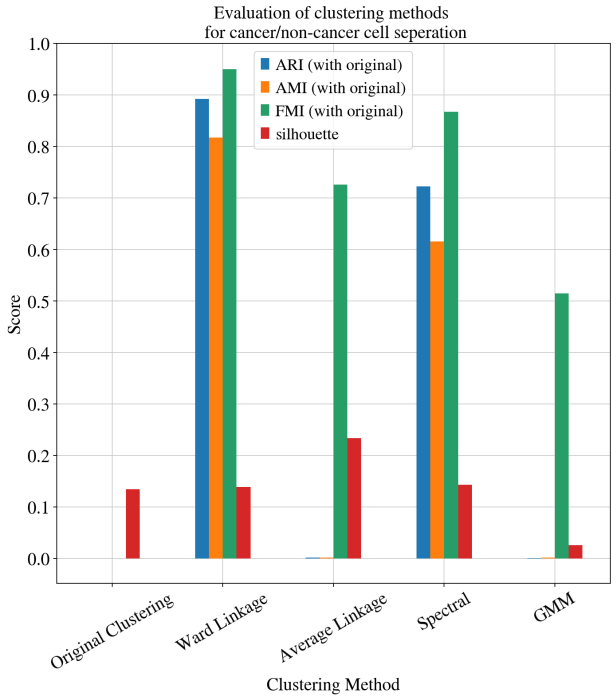


Figure 6: *Resulting clustering metric scores per clustering algorithm employed*

Estimate & Immune cells separation Results We utilized the ESTIMATE R package to predict stromal scores, immune scores, and tumor purity for each cell in our analysis. By examining the predictions of the separated carcinoma and non-carcinoma cells, we observed notable differences in the predicted scores. Specifically, the tumor cells resulted scores that were completely opposite to those of the non-tumor cells as seen in Figure 19.

To further distinguish between immune and stromal cells, we relied on the predicted immune scores. By examining the violin plot in Figure 17, we observed that the predicted immune scores were outside the mean range. We identified these outliers and classified them as stromal cells. Ultimately, we found that only two stromal cells had been misclassified as immune cells as showcased in Figure 23.

UMAP visual representation of the separated cell datasets Figure 8 provides interesting insights into the spatial distribution of non-tumor labeled cells in the UMAP 2D space. Unlike the tumor cells, which form distinct clusters, the non-tumor cells appear to be uniformly distributed without any clear clustering pattern. On the other hand, the UMAP plot for the

tumor cell 7 predictions reveals the formation of distinct clusters among the tumor cells. Notably, there is a noticeable clustering of TNBC subtype tumors, indicating their shared molecular characteristics. Additionally, in close proximity to the TNBC cluster, we can observe some HER2+ cells, suggesting a potential association or similarity between these two subtypes. The UMAP plot depicting the predictions of cancer and non-cancer cells further highlights the differences in their relationships. The cancer cells illustrates a higher level of organization and connectivity, forming organized clusters, while the non-cancer cells show a looser relationship with less evident clustering. This observation suggests that the cancer cells share more common molecular features, whereas the non-cancer cells may display a greater degree of variability or lack strong correlations in gene expression profiles.

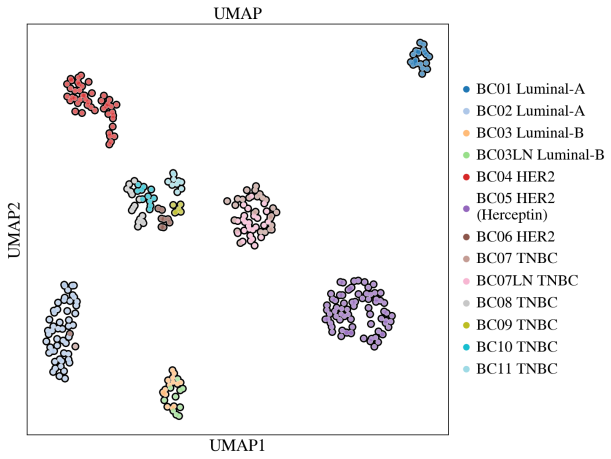


Figure 7: *UMAP, Ward linkage predicted as tumour cells*

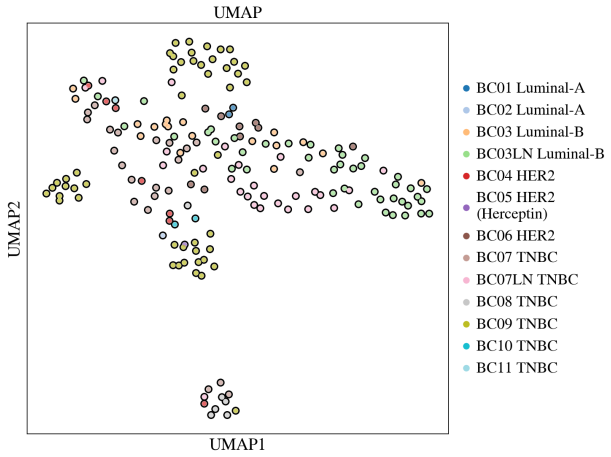


Figure 8: *UMAP, Ward linkage predicted as non-tumour cells*

Pearson's Correlations Results In Figure 24, we observe that Pearson's correlation provides some insights from the results. However, due to the presence of a mixture of carcinoma and non-carcinoma cells in the dataset, it becomes challenging to gain a clear understanding of the correlation between the cells. To address this, we separated the carcinoma and non-carcinoma cells and re-applied Pearson's correlation, Figures 25 26. As shown in Figure 26, this separation resulted in a denoised correlation plot. Notably, we can easily observe correlation patterns across almost all types of the cancer cells, except for cells from position 280 and after. These cells correspond to the TNBC type, and the absence of correlation indicates a high level of heterogeneity among TNBC cells. This finding highlights the unique characteristics and distinct gene expression patterns exhibited by TNBC cells compared to other types of cancer cells.

Cancer cell analysis

Cancer type predictions results In order to predict the cancer subtype, we utilized the Genefu package from R, which offers a range of functions for gene expression analysis. Specifically, we employed the subtype.cluster function, which leverages Gaussian mixture models and incorporates prior knowledge from three different datasets to identify highly correlated genes associated with each subtype. To run this function, the Entrez Gene ID was required. We obtained the gene IDs by utilizing BioPython and the NCBI API, generating a CSV file as input for the gene IDs.

The results 31 of the subtype predictions (with Gaussian Mixture Models) returned probabilities for each cell to be classified into one of the three classes: HER2+, ER+/HER2-, and ER-/HER2-. Notably, the function successfully classified cells into the ER+/HER2-. However, misclassifications to ER+/HER2- observed for all the non ER+/HER2- classes across the rest tumor types. The only non ER+/HER2- classification was the BC07 and BC07LN samples, which were misclassified as HER2+ instead. Overall, the probabilities assigned to the HER2+ and ER-/HER2- subtypes increased the probability of belonging to the correct class, although they were not dominant enough to ensure accurate classification.

This particular aspect of the study proved challenging to replicate. We speculate that the inclusion of approximately 10,000 genes during the preprocessing stage may have introduced some genes that are highly correlated with the ER+/HER2- genes from the R package. This potential correlation may have contributed to the discrepancies observed in the results. Despite this limitation, the findings from the rest of the study remain robust and provide valuable insights into the cancer subtypes under investigation.

GSVA for Subtype specific gene expression profiling at single-cell resolution results Conducting the analysis of differentially expressed genes across various subtypes of tumors at both the single-cell and bulk levels, we have successfully replicated the results obtained by the researchers. Figures 33 provide visual representations of the results obtained from the tumor cells and bulk tumors, respectively, further validating the tumor heterogeneity, especially between TNBC types.

Furthermore, focusing on patient BC05, who underwent herceptin treatment we observed significantly lower expression levels of marker genes associated with the HER2+ subtype in this particular patient compared to other HER2+ patients. This finding suggests a potential impact of the herceptin treatment on the gene expression patterns related to the HER2+ subtype.

GSVA for Subtype related pathways results Using the gene pathways smid breast cancer luminal b up, halmark estrogen response early, halmark estrogen response late, nikolsky breast cancer 17q11 q21 amplicon, smid breast cancer basal up, and smid breast cancer relapse in brain up, we validated of the researcher's results regarding the associations between these pathways and each tumor subtype as showcased in Figure **??**. The first three pathways, smid breast cancer luminal b up, halmark estrogen response early, and halmark estrogen response late, were found to be closely related to the ER+ subtype, providing support for the researchers' results. These pathways are associated with estrogen response, which aligns with the characteristics of the ER+ subtype that is known to be influenced by estrogen signaling. On the other hand, the pathway nikolsky breast cancer 17q11 q21 amplicon was found to be specifically related to the HER2+ subtype. This pathway likely plays a role in the amplification of genes located in the 17q11-q21 region, which is associated with the HER2+ subtype. Lastly, the pathways smid breast cancer basal up and smid breast cancer relapse in brain up were found to be associated with the ER-/HER2- subtype. These pathways indicate molecular characteristics specific to this subtype, potentially contributing to its basal-like nature and the propensity for relapse in the brain.

In the analysis of aggressive cancer gene-sets related to epithelial-mesenchymal transition (EMT), stemness, and angiogenesis, our findings align with the reference paper with only slight variations due to the inclusion of the Herception treated patient BC05 in our analysis. Firstly, EMT characterizes the cellular process wherein cells undergo a transformative change, enhancing their migratory capacity. Stemness, on the other hand, relates to cells exhibiting stem cell-like properties, including the abilities of self-renewal and differentiation. Lastly, angiogenesis involves the vital role of cells in promoting tumor blood supply growth and bolstering immune defense mechanisms. Through Pearson's Correlation analysis, we discovered an r coefficient of 0.39 between EMT and Stemness in TNBC cells from patients BC11 and BC10, indicating a medium correlation. These patients exhibited the most aggressive cells within this gene-set pair, as showcased in Figure 9.

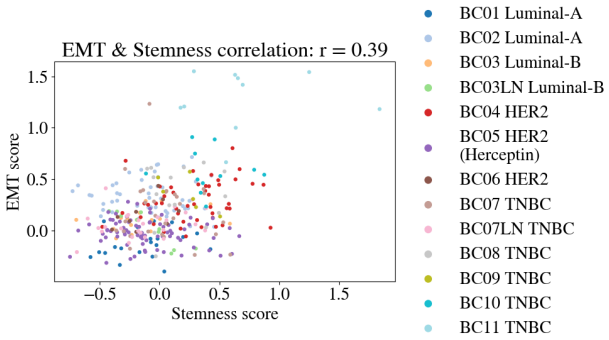


Figure 9: *EMT & Stemness correlation analysis*

For Stemness and Angiogenesis, the correlation was lower at 0.16, with patients BC11 and BC10 again displaying the most aggressive cells, as illustrated in Figure 10. Interestingly, Angiogenesis and EMT had

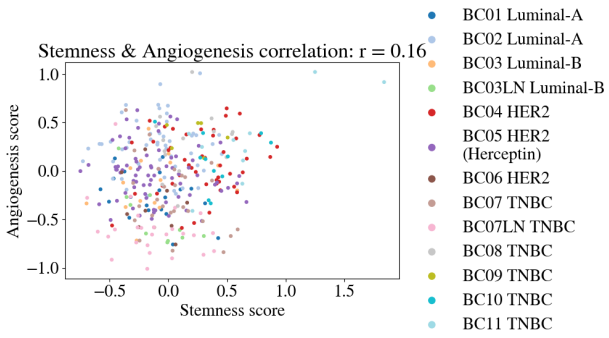


Figure 10: *Stemness & Angiogenesis correlation analysis*

the highest correlation among all pairs, equal to 0.46, once again highlighting patients BC11 and BC10 as having the most aggressive tumor cells for these gene signatures, results that cen be seen in Figure 11. It is noteworthy to mention that the HER2 patient BC05 (purple), who received herceptin treatment, exhibited relatively lower expression levels across all aggressive gene-sets compared to the non-treated HER2 patient BC04 (red). Moreover, the utilization of single-cell analysis highlights the evident intratumoral heterogeneity, allowing us to observe substantial differences in the expression of the gene-sets among for example aggressive triple-negative breast cancer (TNBC) pa-

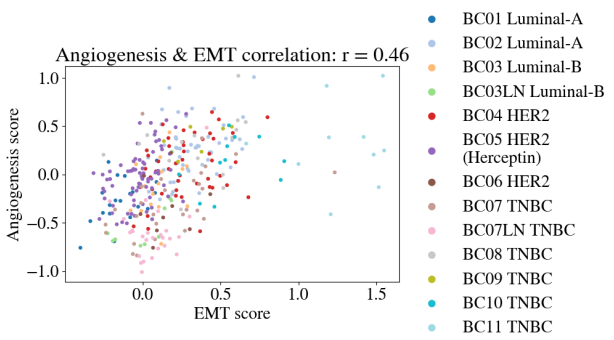


Figure 11: *Angiogenesis & EMT correlation analysis*

tients BC10 and BC11. These findings emphasize the importance of considering individual cell variations in our analysis.

Analysis of Immune cells

Immune cell type separation in the reference paper, was carried out with non-negative matrix factorization (NMF) to accurately differentiate cells into B-cells, T-cells, and macrophages. Their methodology produced a noteworthy silhouette score of 0.95 (supplementary material). Nonetheless, we encountered an issue. Specifically, during the earlier stages of analysis, we mistakenly classified three tumor cells and two stromal cells as immune cells, as showcased in Figure 28, which had a significant impact on the silhouette score when employing the labels provided in the reference paper and utilizing the sub-type specific genesets. Despite the decline in the values of clustering evaluation metrics, the comparison remains relative, and it did not have an impact on the ranking of the results. When compared to other methods, NMF clustering showcased in Figures 29(a), 29(b) and 29(c) proved to be the best clustering method, achieving a silhouette score of 0.245, an ARI of 0.740, an AMI of 0.667, and an FMI of 0.83. Contrary to the tumour/non-tumour cell separation the other clustering methods also preformed relatively well with spectral clustering, showcased in Figures 29(g), 29(i), 29(h), performing second best with a silhouette score of 0.252, ARI equal to 0.629, AMI equal to 0.597 and an FMI of 0.774. And GMMs, Ward linkage showcased in the remaining sub-figures of Figure 29 produced similar results with silhouette scores of $0.206, 0.201$, ARI of $0.614, 0.576$, AMI of $0.508, 0.579$ and finally FMI equal to $0.751, 0.721$ respectively. These results validate the utilization of the NMF clustering of researchers in the reference paper.

To enhance the visualization and facilitate a more detailed comparison with the results of the reference paper, we present the immune cell sub-type gene-

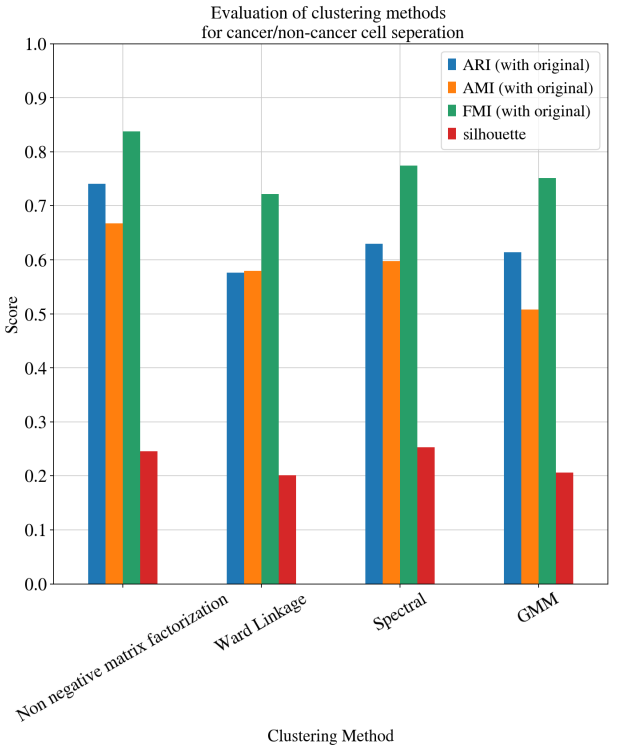


Figure 12: *Immune cell separation into B-cells, T-cells, Macrophages clustering evaluation*

set expression heat-map organized according to the identified with NMF clusters in Figure 30(b), along with the same heatmap annotated with the respective tumour groups 30(a). As showcased the majority of B-cells are primarily found in the Luminal B and TNBC Lymph node subtypes, whereas T-cells and macrophages are predominantly observed in TNBC tumors. These results are in line with the researchers' findings in the reference paper.

T-cell type separation The T-cells identified through NMF clustering were subsequently categorized based on the clustering of their gene expression patterns specific to various subtypes. This classification involved gene sets related to costimulatory, cytotoxic, exhausted, naive, and regulatory functions. Additionally, gene sets related to cell cycle phases G1/2 and G2/M were utilized for further analysis. Lacking the labels assigned in the reference paper for this stage the shilouete score results ranked average linkage as the best method with a score of 0.146, followed by complete linkage, spectral clustering ward linkage and single linkage with scores: $0.099, 0.059, 0.057, 0.041, 0.041$ respectively, as showcased in Figure 13. The clustering results in the 2-D PCA,UMAP and TSNE of all the method described above are showcased in Figure 32. Additionally, we generated a gene expression heat-map, which was annotated with tumor groups clustered based on the activation levels of T-cell signature pathways. This

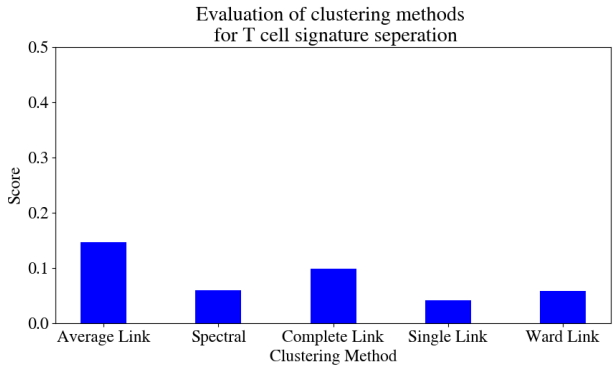


Figure 13: *T-cell signature pathways clustering scores per algorithm*

heat-map, shown in Figure 34, further validates our decision to use average linkage. Our findings align with those reported in the reference paper, where T-cells were extracted from four patients. Specifically, BC03 exhibited high Naive and Costimulatory pathway signatures, BC09 showed high Cytotoxic and Exhausted signatures, and BC07 displayed Naive and Regulatory pathway signatures.

Conclusions

Based on our comprehensive cluster analysis, it appears that the researchers most likely employed the Ward linkage method to separate cancer and non-cancer cells. Furthermore, they likely utilized average linkage for clustering T cell pathway scores. Finally we verified the usage of NMF as the best method for immune cell sub-type separation into T-cells, B-cells, and macrophages. Throughout the analysis, UMAP and t-SNE proved effective in revealing distinct clusters for different tumor groups, while PCA, limited to linear relationships, was unable to. Finally, we have replicated and verified the results of the paper, which demonstrate the possibility of separating and characterizing tumor and immune cells at the single-cell level, providing advantages for targeted therapies. However, it is important to note that the tumor groups are imbalanced and different in size, so caution for potential bias should be exercised. Thus To ensure the total validity of these findings, it is crucial for future research to be conducted using a more comprehensive and well-balanced single-cell data-set that pertains to breast cancer patients.

Figures

The remaining figures referenced in the methods & results sections of the report are showcased on the following pages: 11 – 22.

References

[1] Woosung Chung et al. “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer”. In: *Nature communications* 8.1 (2017), p. 15081.

[2] *Breast Cancer Information and Support*. URL: https://www.breastcancer.org/.

[3] Isaac Virshup et al. “anndata: Annotated data”. In: *BioRxiv* (2021), pp. 2021–12.

[4] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome biology* 19 (2018), pp. 1–5.

[5] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[6] Takuya Akiba et al. “Optuna: A next-generation hyperparameter optimization framework”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[7] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[8] *ESTIMATE*. 2016-09-26. URL: https://bioinformatics.mdanderson.org/public-software/estimate/.

[9] Deena M A Deena M A Gendoo et al. “Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer”. In: *PubMed* (2016).

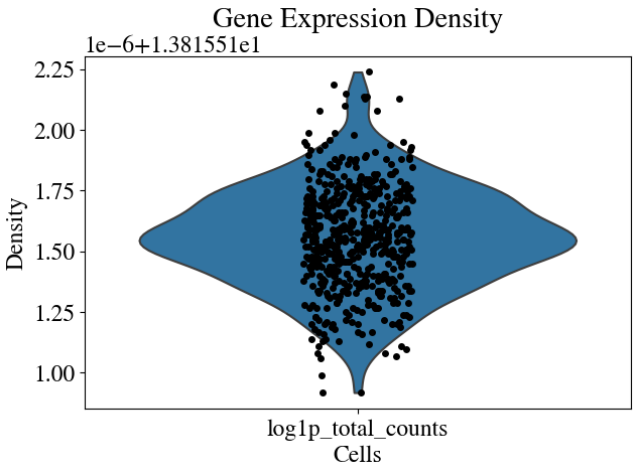
[10] URL: https://www.gsea-msigdb.org/gsea/index.jsp.

Code Availability

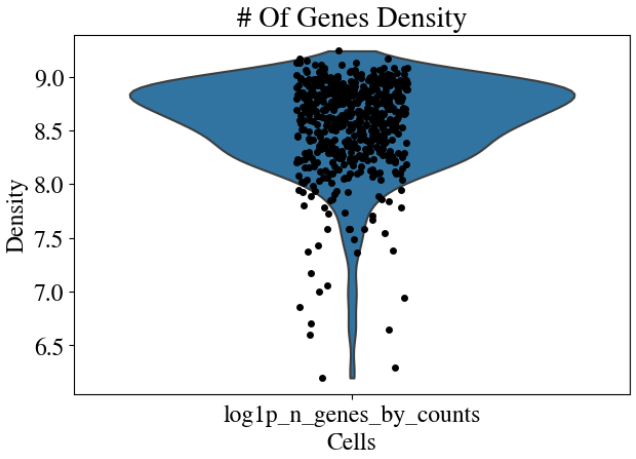
You can visit the GitHub repository of this project, to access all the code and results. However, please note that the data sets are not included in the repository due to their large file size.

Data Availability

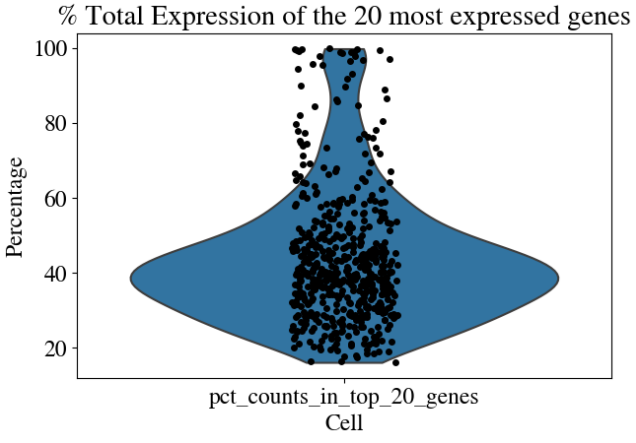
Both the reference paper datasets used in this project and the annotated Anndata objects generated during our analysis can be downloaded from this Google Drive folder.



(a) log-transformed overall count of all gene expression in the cells.

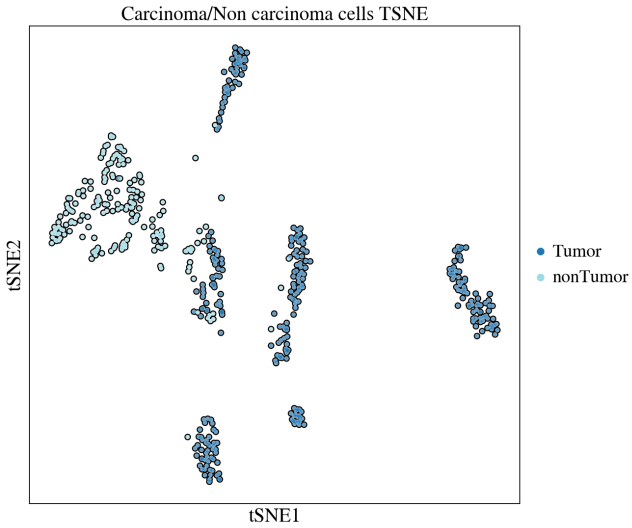


(b) log-transformed number of genes expressed in the cells.

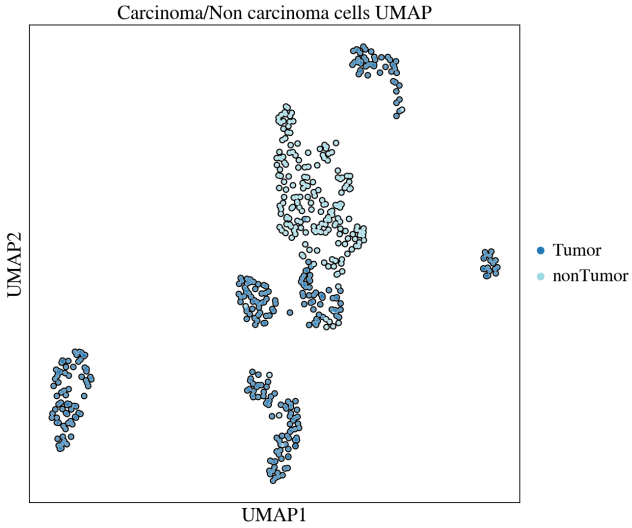


(c) Percentage of the total expression in the cells of the 20 most expressed genes.

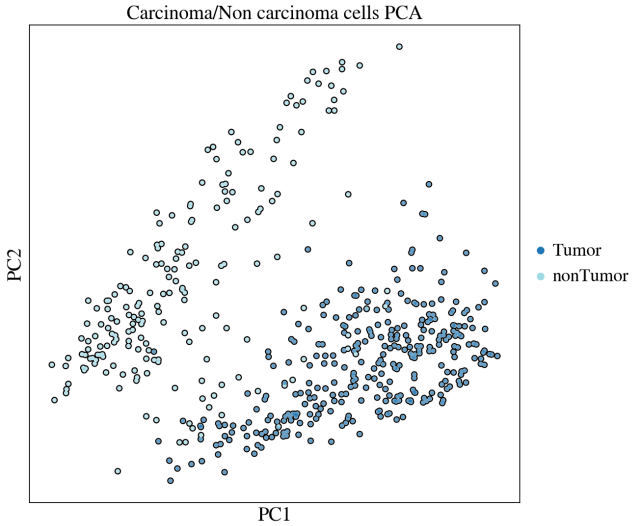
Figure 14: *Cell filtering prepossessing QC distributions*



(a) Reference paper, tumour/non-tumour cell ground labels, t-SNE.

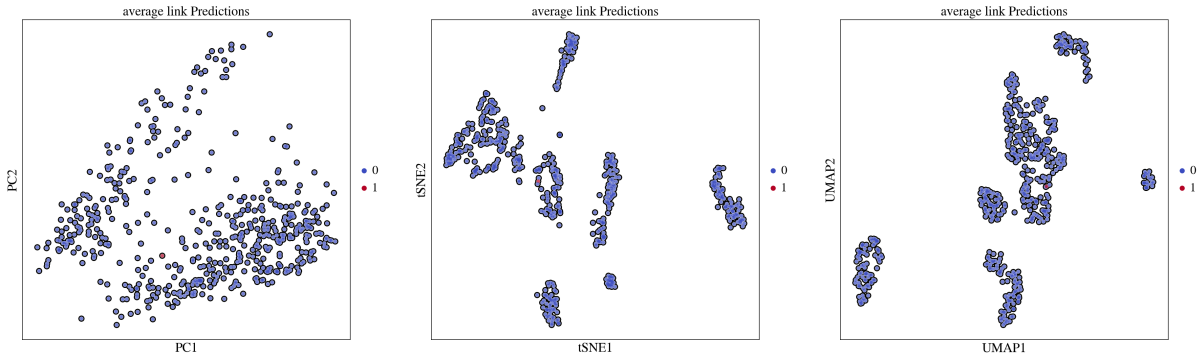


(b) Reference paper, tumour/non-tumour cell ground labels, UMAP.



(c) Reference paper, tumour/non-tumour cell ground labels, PCA.

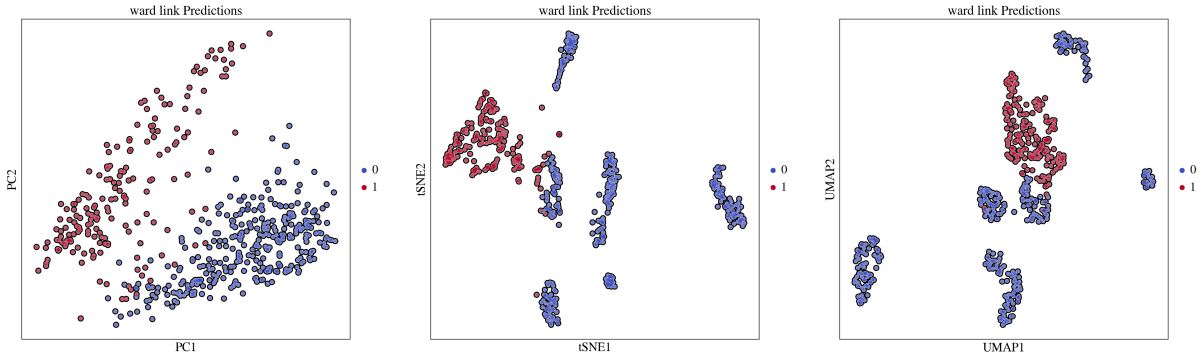
Figure 15: *Tumour& Non-tumour cell labels as identified after clustering in the reference paper*



(a) PCA, Average linkage tumour/non-tumour cell separation.

(b) TSNE, Average linkage tumour/non-tumour cell separation.

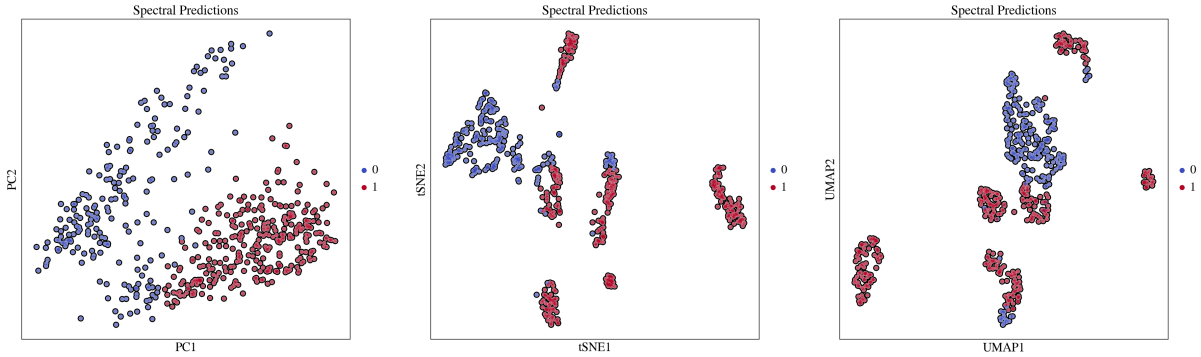
(c) UMAP, Average linkage tumour/non-tumour cell separation.



(d) PCA, Ward linkage tumour/non-tumour cell separation.

(e) TSNE, Ward linkage tumour/non-tumour cell separation.

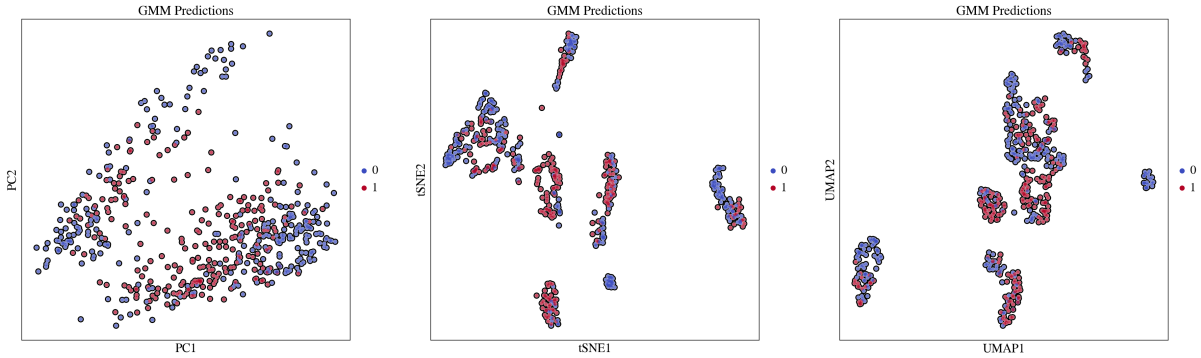
(f) UMAP, Ward linkage tumour/non-tumour cell separation.



(g) PCA, Spectral clustering tumour/non-tumour cell separation.

(h) TSNE, Spectral clustering tumour/non-tumour cell separation.

(i) UMAP, Spectral clustering tumour/non-tumour cell separation



(j) PCA, GMM clustering tumour/non-tumour cell separation.

(k) TSNE, GMM clustering tumour/non-tumour cell separation.

(l) UMAP, GMM clustering tumour/non-tumour cell separation.

Figure 16: *Tumour& Non-tumour cell labels as identified with our own clustering methodologies*

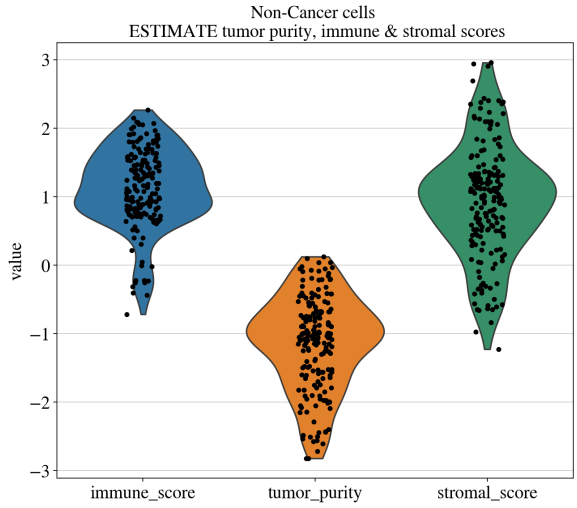


Figure 17: *Predictions of Estimate package Immune scores, Stromal scores and Tumor purity of non-Carcinoma cells. Upon examination, we noted that non-tumor cells exhibited low tumor purity scores, indicating a minimal presence of tumor cells within these samples. However, both the immune scores and stromal scores were observed to be significantly high in these cells. This stark contrast in results is in direct opposition to the patterns observed in cancer cells. Furthermore, we identified outlier cells within the immune scores violin plot, and upon analysis, these outliers were subsequently classified as stromal cells. This finding further supports the reverse relationship between immune and stromal scores in non-tumor cells, highlighting the distinct characteristics and composition of these cell types.*

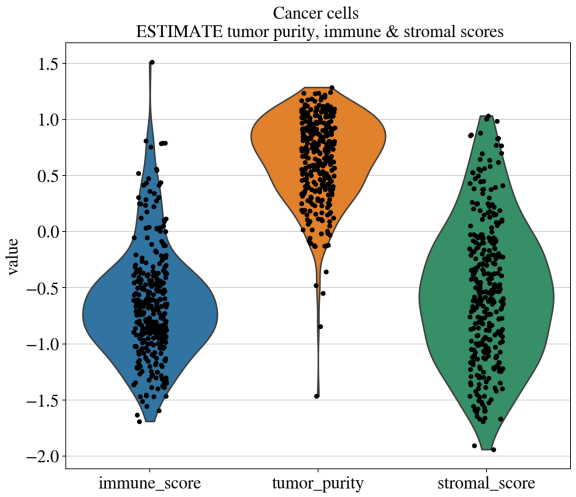


Figure 18: *Predictions of Estimate package Immune scores, Stromal scores and Tumor purity of Carcinoma cells. Upon analyzing the results, we observed that tumor cells displayed high tumor purity scores, indicating a strong presence of tumor cells within these samples. However, the immune scores and stromal scores for these cells were observed to be significantly low.Furthermore, in the violin plot, we identified four outliers that stood out from the clustering results. These outliers corresponded to misclassified tumor cells that were erroneously categorized as non-tumor cells. This misclassification contributed to the observed discrepancies in the immune scores and stromal scores within these particular cells.*

Figure 19: *Predicted scores of Estimate Package for Carcinoma and non-Carcinoma cells*

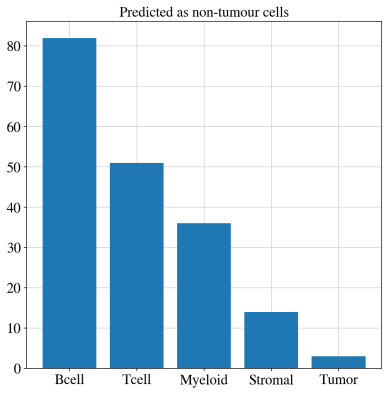


Figure 20: *This barplot provides an overview of the initial predictions for non-carcinoma cells. Notably, the majority of the cells are classified as stromal cells, indicating a significant presence of stromal components within the non-tumor samples. However, upon closer examination, we identified three cells that were misclassified as tumor cells.*

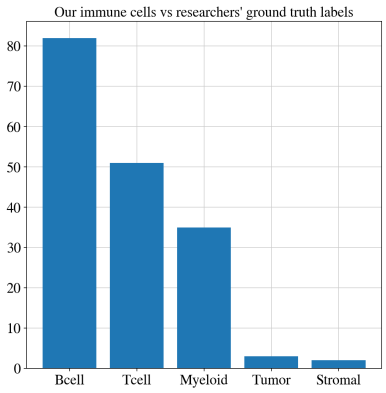


Figure 21: *This barplot showcases the classification of cells as immune cells based on the initial predictions. While there are still three misclassified tumor cells present, an interesting observation is the reduction in the number of stromal cells, which now stands at two. This reduction was achieved by reclassifying the outliers identified in the immune scores violin plot for the non-carcinoma cells as stromal cells.*

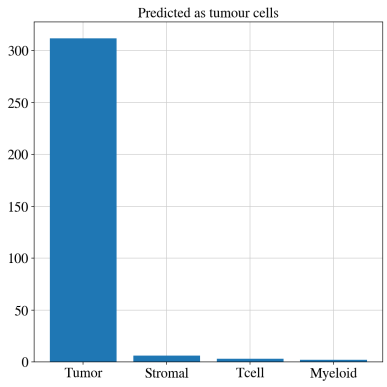


Figure 22: *This barplot presents the predicted tumor cells resulting from the clustering analysis. It is evident that the number of misclassified non-tumor cells as tumor cells is relatively small. Specifically, there are two stromal cells and two immune cells that were incorrectly classified as tumor cells.*

Figure 23: *Predictions of tumor and non-tumor cells and predictions of stromal an immune cells.*

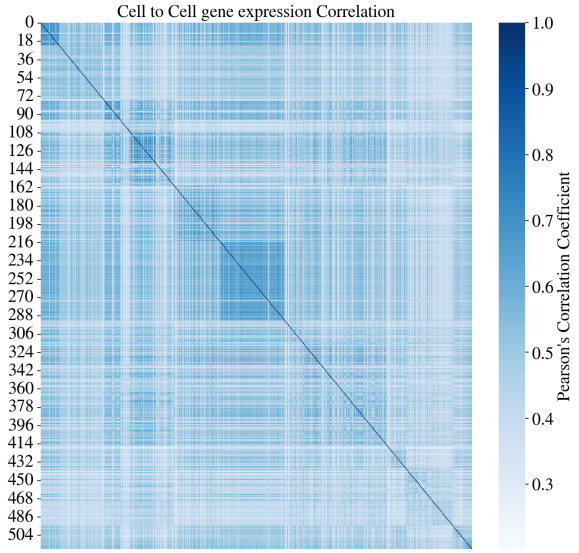


Figure 24

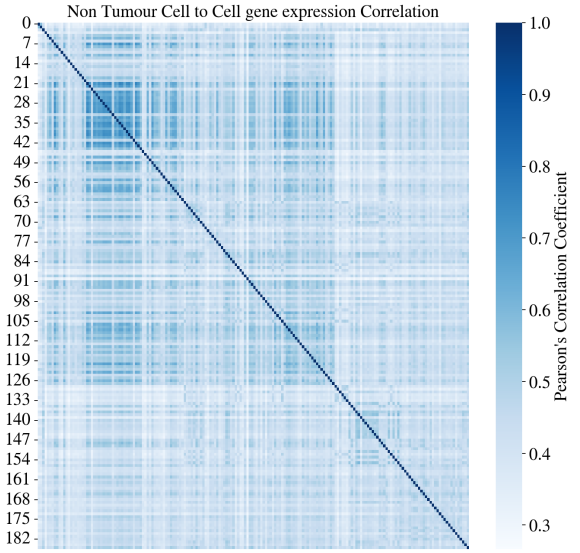


Figure 25

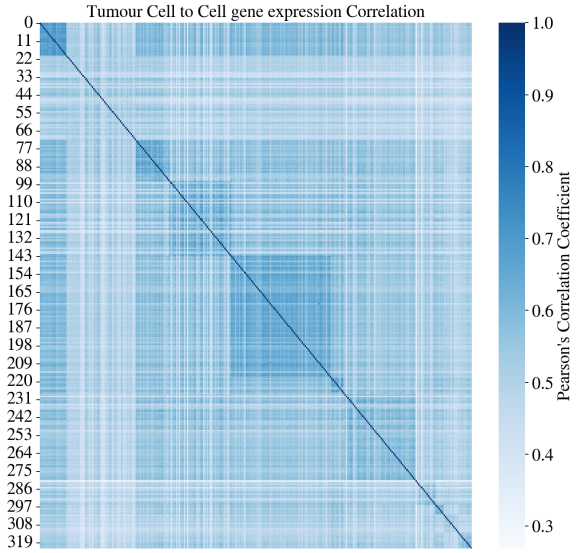
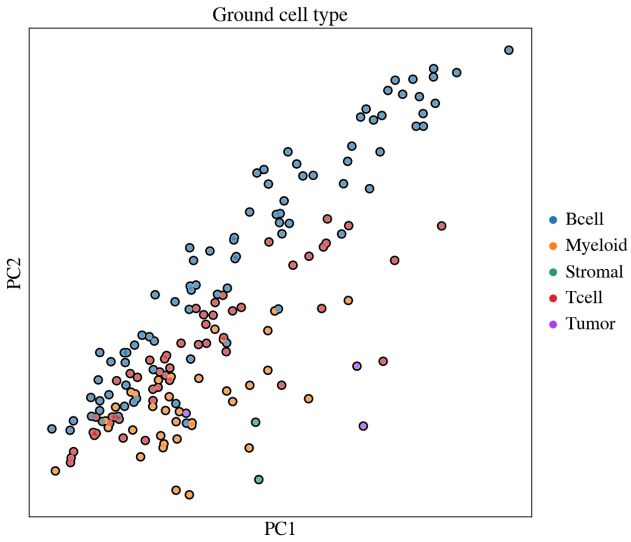
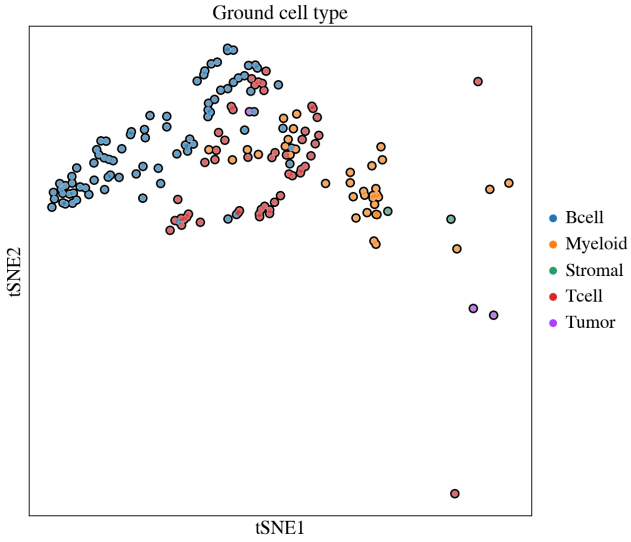


Figure 26

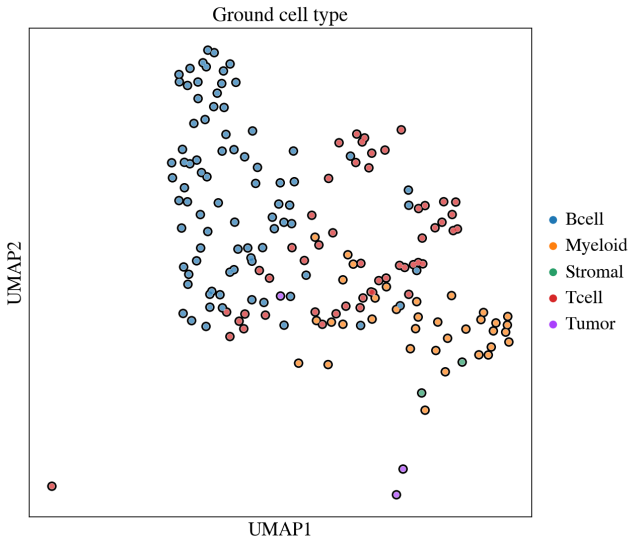
Figure 27: *Three correlation plots were generated using Pearson's correlation coefficient. Figure 24 illustrates the correlation between all cell types in the dataset, providing a comprehensive overview of the intercellular relationships. In Figure 25, the correlation analysis was focused exclusively on non-tumor cells. Figure 26, on the other hand, displays the correlation patterns restricted to tumor cells, allowing for a more detailed exploration of the relationships within this subset.*



(a) Reference paper immune cell ground labels, PCA.

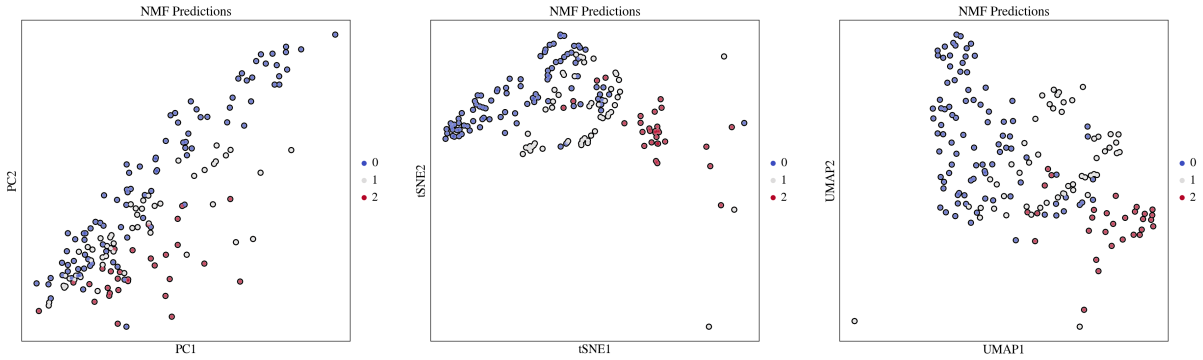


(b) Reference paper immune cell ground labels, TSNE.



(c) Reference paper immune cell ground labels, UMAP.

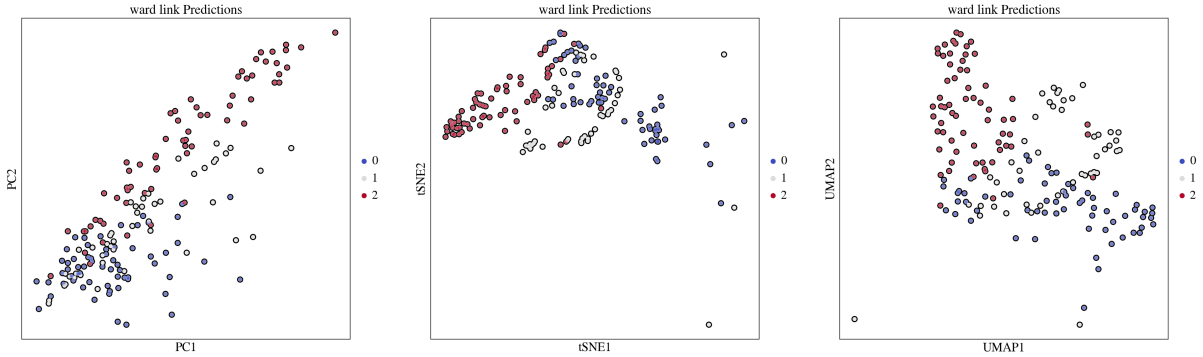
Figure 28: *Immune cells labeled with the labels identified in the reference paper*



(a) PCA, NMF immune cell sub-type separation.

(b) TSNE, NMF immune cell sub-type separation.

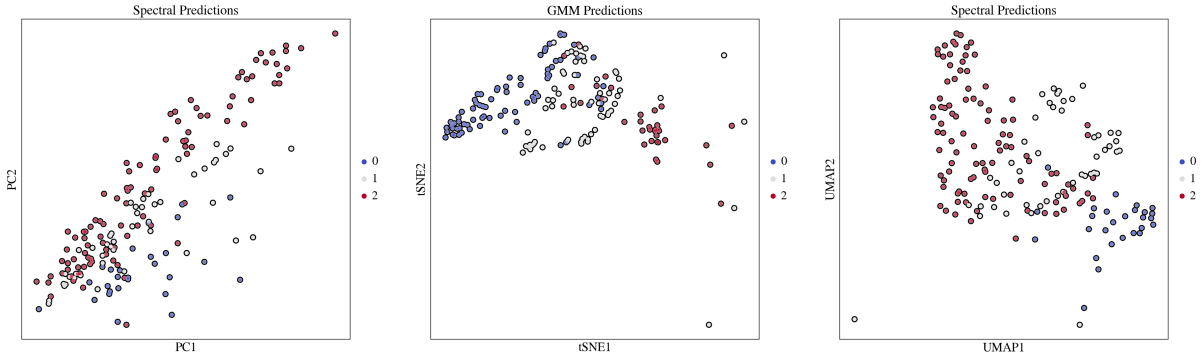
(c) UMAP, NMF immune cell sub-type separation.



(d) PCA, Ward linkage immune cell sub-type separation.

(e) TSNE, Ward linkageimmune cell sub-type separation.

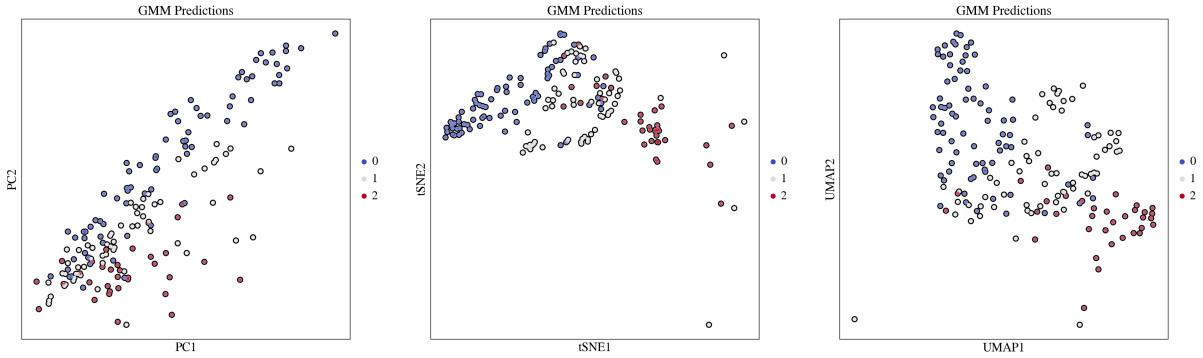
(f) UMAP, Ward linkage immune cell sub-type separation.



(g) PCA, Spectral clustering immune cell sub-type separation.

(h) TSNE, Spectral clustering immune cell sub-type separation.

(i) UMAP, Spectral clustering immune cell sub-type separation.

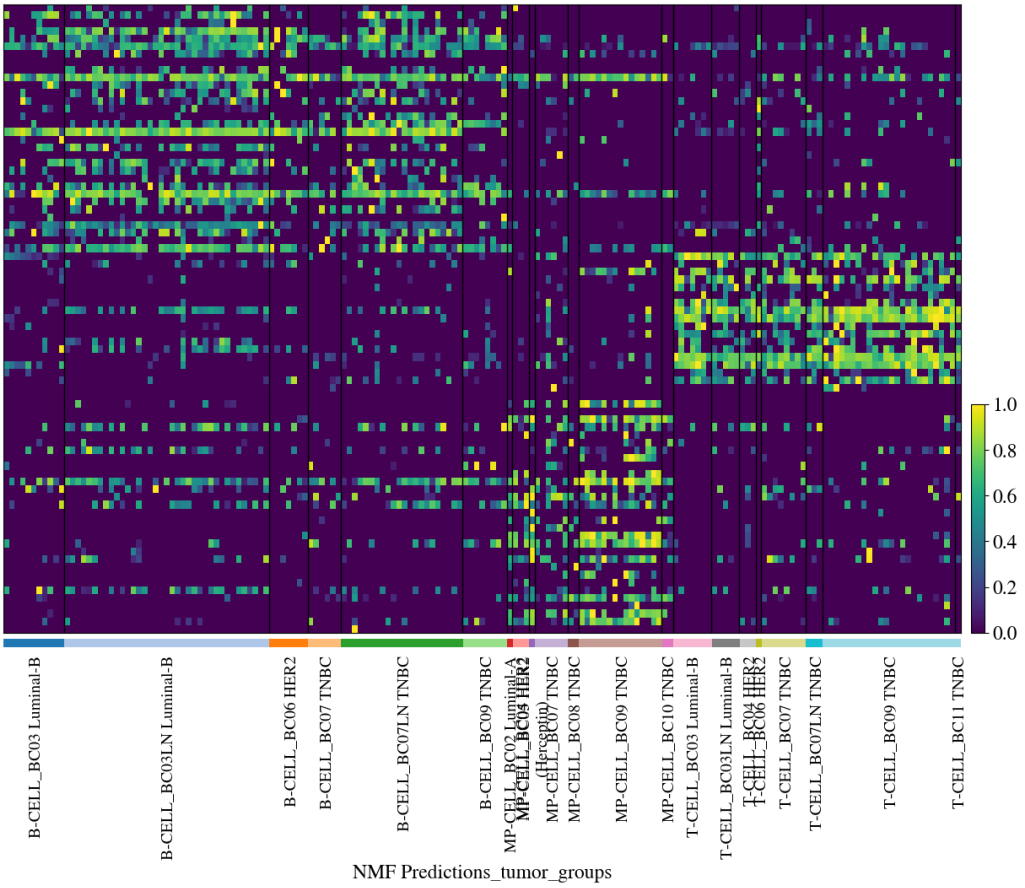


(j) PCA, GMM clustering immune cell sub-type separation.

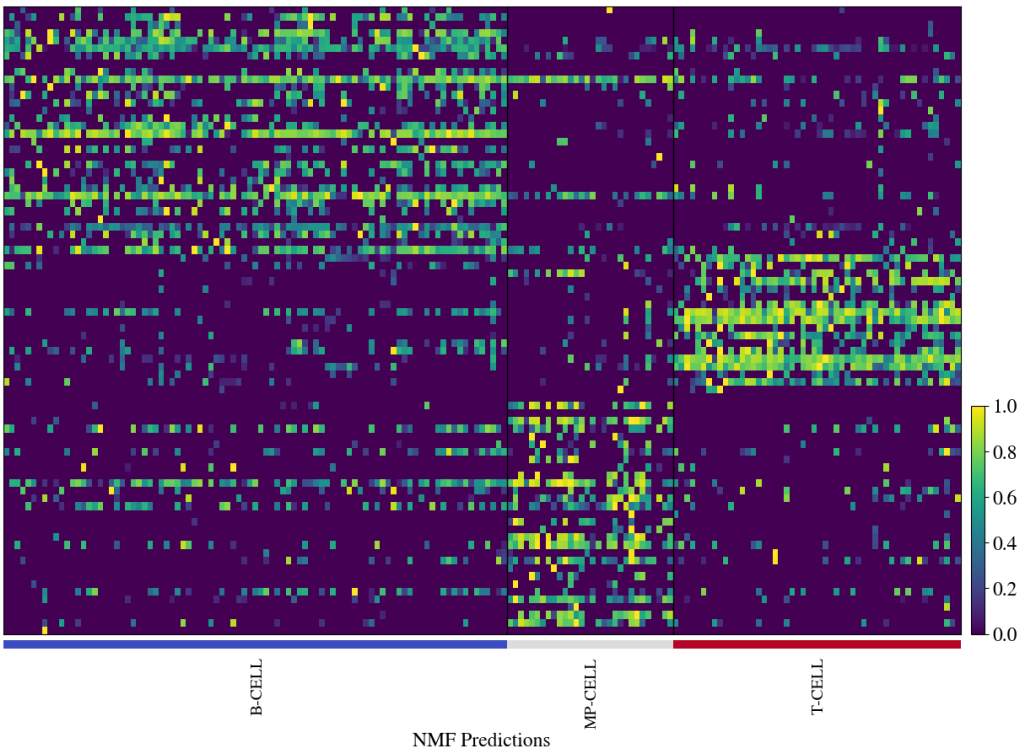
(k) TSNE, GMM clustering immune cell sub-type separation.

(l) UMAP, GMM clustering immune cell sub-type separation.

Figure 29: *immune cell sub-type separation labels as identified with our own clustering methodologies*



(a) Immune cell sub type gene-set expression for clusters identified by NMF



(b) Immune cell sub type gene-set expression for clusters identified by NMF & annotated with tumour groups

Figure 30: *Gene-set expression heat-maps for the NMF predicted clusters*

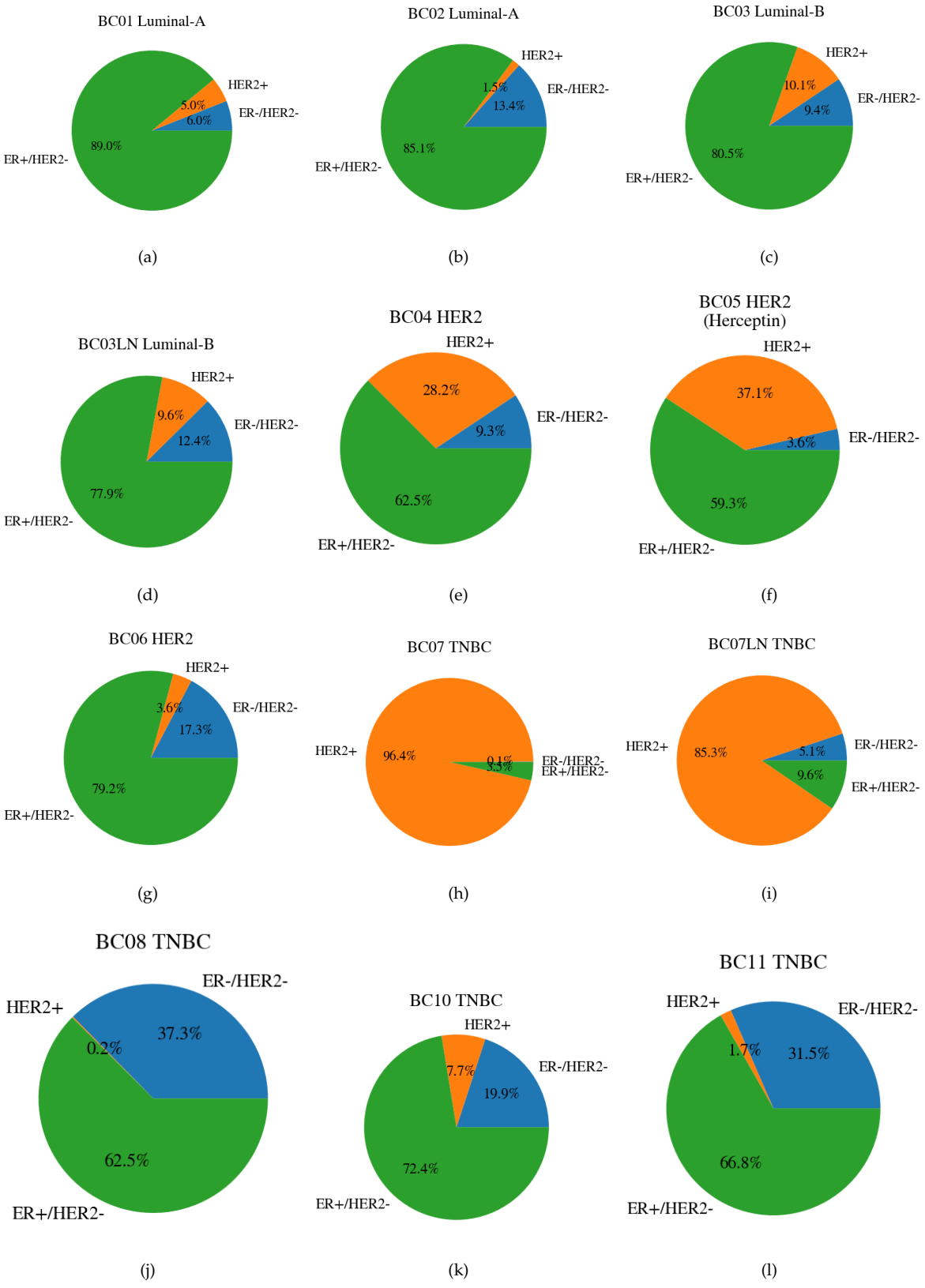
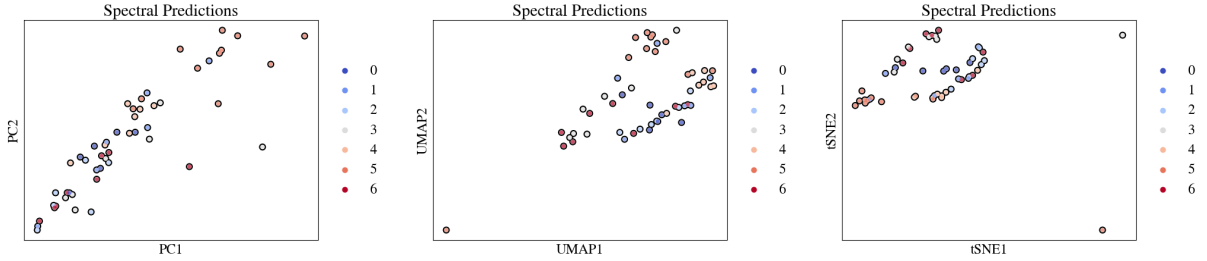


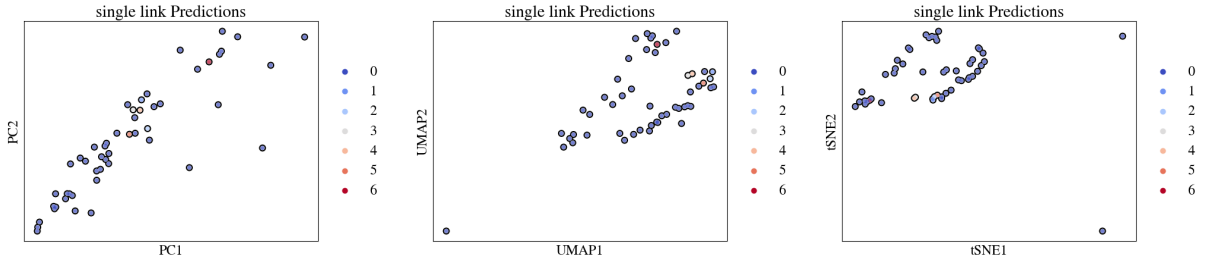
Figure 31: *Results from Genefu package for classification of carcinoma cells*



(a) PCA, Spectral T-cell signature pathway separation.

(b) UMAP, Spectral T-cell signature pathway separation.

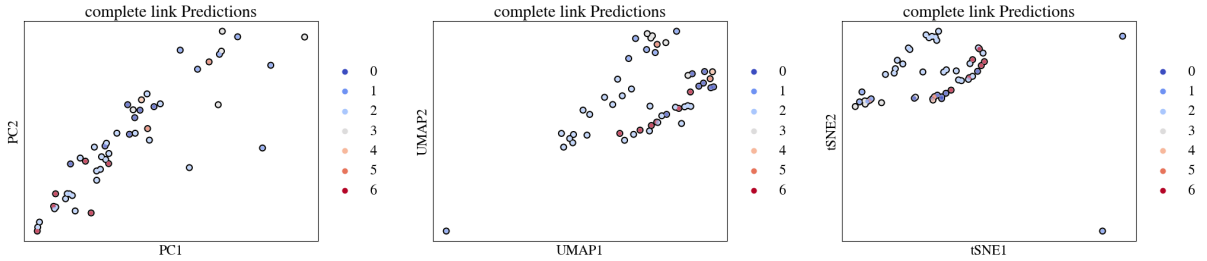
(c) TSNE, Spectral T-cell signature pathway separation.



(d) PCA, Single linkage T-cell signature pathway separation.

(e) UMAP, Single linkage T-cell signature pathway separation.

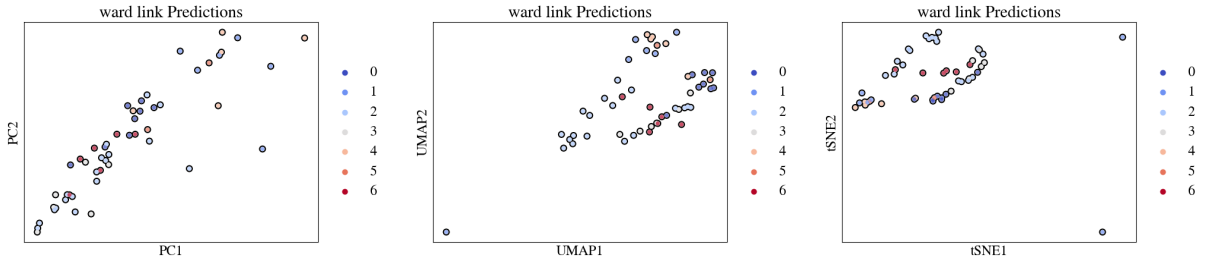
(f) TSNE, Single linkage immune cell sub-type separation.



(g) PCA, Complete linkage immune cell sub-type separation.

(h) UMAP, Complete linkage immune cell sub-type separation.

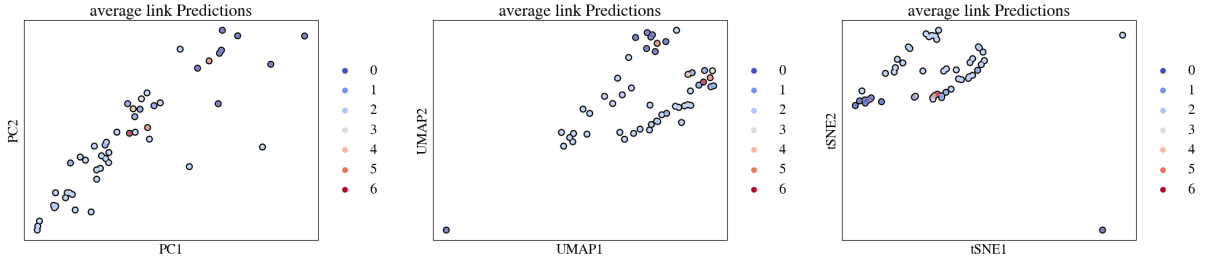
(i) TSNE, Complete linkage immune cell sub-type separation.



(j) PCA, Ward linkage immune cell sub-type separation.

(k) UMAP, Ward linkage immune cell sub-type separation.

(l) TSNE, Ward linkage immune cell sub-type separation.

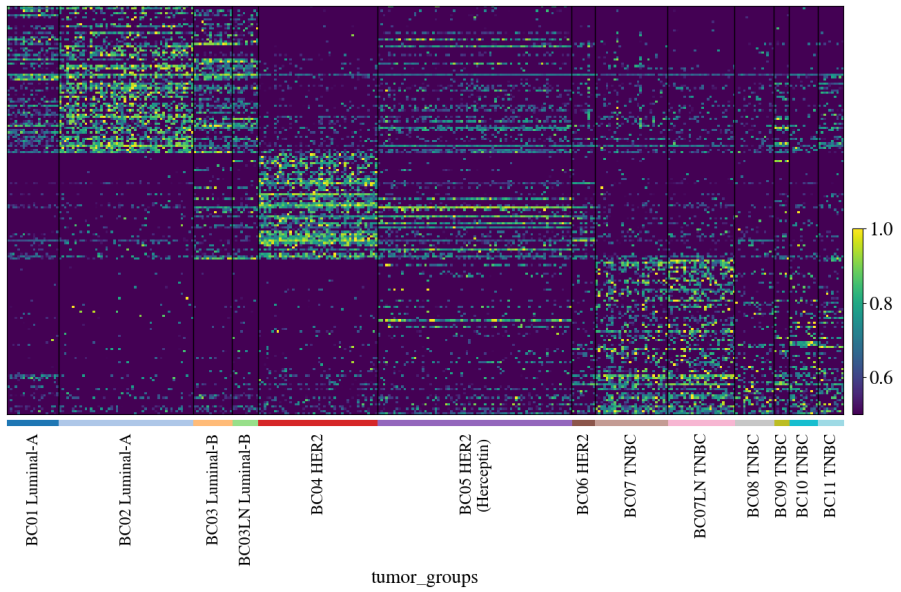


(m) PCA, Average linkage immune cell sub-type separation.

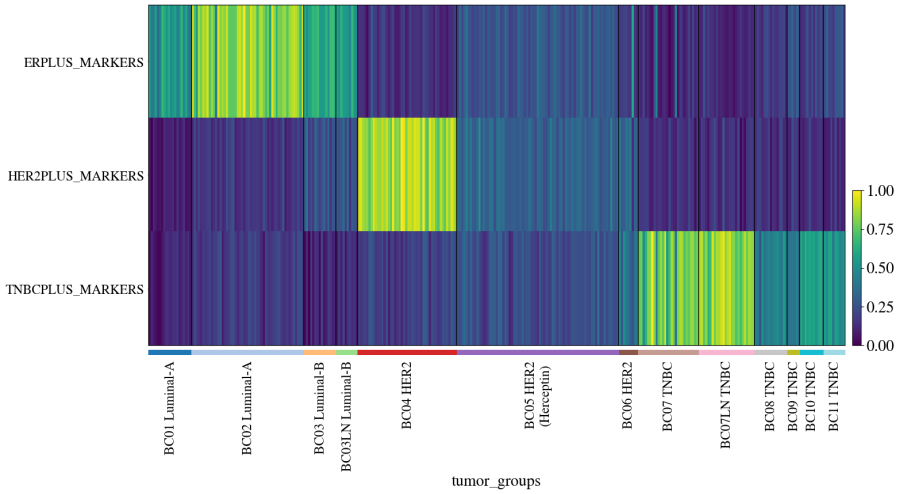
(n) UMAP, Average linkage immune cell sub-type separation.

(o) TSNE, Average linkage immune cell sub-type separation.

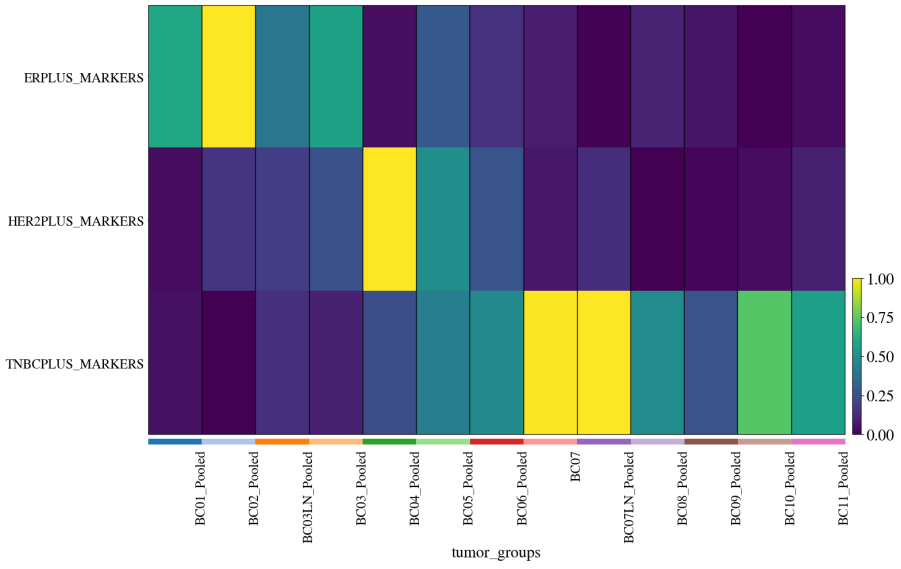
Figure 32: *T-cell separation via clustering the expression of signature pathways*



(a)



(b)



(c)

Figure 33: *Results for gene expression analysis for ER+, HER2+, and TNBC marker genes for tumor cells and bulk tumors.*

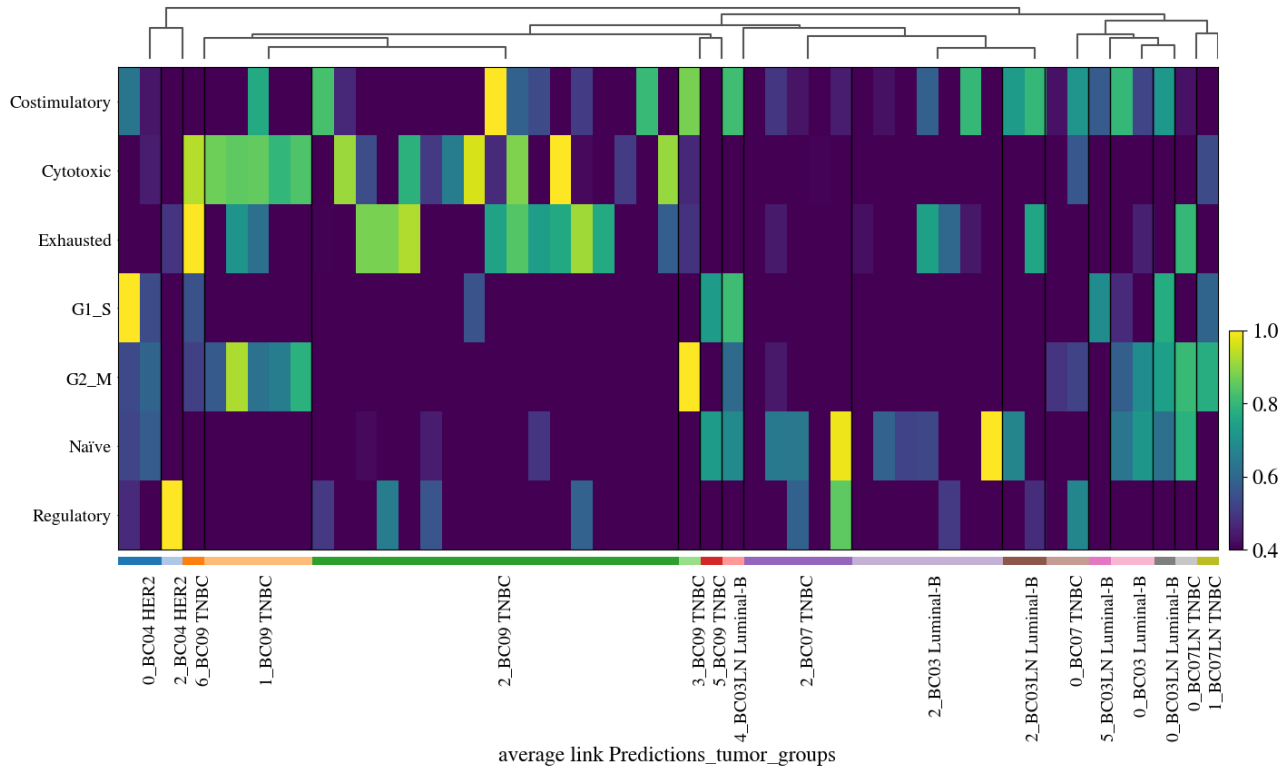


Figure 34: *Gene expression heat-map annotated with the tumor groups clustered via activation levels of T-cell signature pathways*