Delineating spatially significant intra- and inter-tumoral cancer biomarkers using graph attention on local subgraphs of tumor microenvironments

Single-cell spatially resolved transcriptomics (scST), which achieves cellular or sub-cellular resolution, enables the measurement of gene expression in individual cells and their specific locations within a tissue. These spatial gene expression (SGE) patterns can potentially delineate tissue structures and pathology phenotypes. Traditionally, differential expression analysis (DEA) is used to study markers of differentially expressed genes (DEGs) across cells under various conditions, such as different pathology types. DEA seeks to identify unique genes that are up- or down-regulated, irrespective of their spatial distribution. However, DEA often fails to distinguish genes with specific spatial patterns or genes expressed in only a few spatially segregated cells within the tumor microenvironment (TME). To address these challenges, we propose a novel graph-based learning paradigm to identify and elucidate biomarkers of spatial significance. Our approach begins with the creation of a graph representation of the local TME from the spatial transcriptomic dataset, with nodes representing cells and node features representing gene expressions. We generate these local TME subgraphs using a density-aware farthest-point sampling algorithm. Subsequently, we employ a Graph Attention Network (GAT) for effective modeling of the spatial data, defining a classification task to predict node-level labels for each condition (such as different tumor subtypes or pathology types). After successful training, the GAT model identifies unique SGE patterns that characterize these conditions. The model's learned representations are then queried through iterative gene masking to generate spatially significant biomarker scores for each gene in each cell, aiding in the precise classification of the condition.

We applied our method to five non-small cell lung cancer (NSCLC) scST samples, which were annotated for complex acinar (CA), micropapillary, and solid tumor subtypes by a board-certified pathologist. After achieving a balanced accuracy of 95% using the GAT model, we clustered the gene-level biomarker significance scores generated for the tumor cells and identified unique intra-tumoral microenvironments within the CA subtypes, as well as distinct inter-tumoral microenvironments in the micropapillary and solid subtypes. A careful evaluation of these clusters, particularly in the CA subtype, revealed that the clusters not only capture the geometric location of the cells but also their unique gene expression signatures. These signatures delineate an intra-tumoral landscape defined by fibroblasts (cluster 1) at the tumor-stroma boundary, neutrophils (cluster 2) in the tumor core, and T cells (cluster 3) in spatially segregated groups.

Furthermore, we found that spatially significant cancer biomarkers (SSCBs) are more indicative of survival than DEGs. Among the common markers MIF, MZT2A, RPL22, and HSP90AB1 identified across the intra-tumoral CA clusters, we discovered significant and unique SSCBs: IFITM3 and EIF5A in cluster 1 (CA tumor boundary), NDRG1 in cluster 2 (CA tumor core), and MALAT1 and SOX4 in cluster 3 (CA near high T-cell areas). We ranked the SSCBs by their significance scores and conducted Gene Set Enrichment Analysis (GSEA) which revealed richer functions in cluster 1 compared to DEGs, including apoptosis, angiogenesis, and Myc targets V1 suggesting a proliferative TME at the tumor boundary. Additionally, clusters 2 and 3 displayed previously unidentified functions related to hypoxic and inflamed TME, actively recruiting neutrophils and lymphocytes respectively, thereby highlighting intra-tumoral variations that DEA failed to capture. The advantages of delineating SSCBs over DEGs can be attributed to the focused analysis of local expressions of subgraphs and the gene-level masking used to identify SSCBs.