Identifying related gene sets in colorectal cancer

Single Cell Sequencing Analytics Summer School 2022

zweiguyz

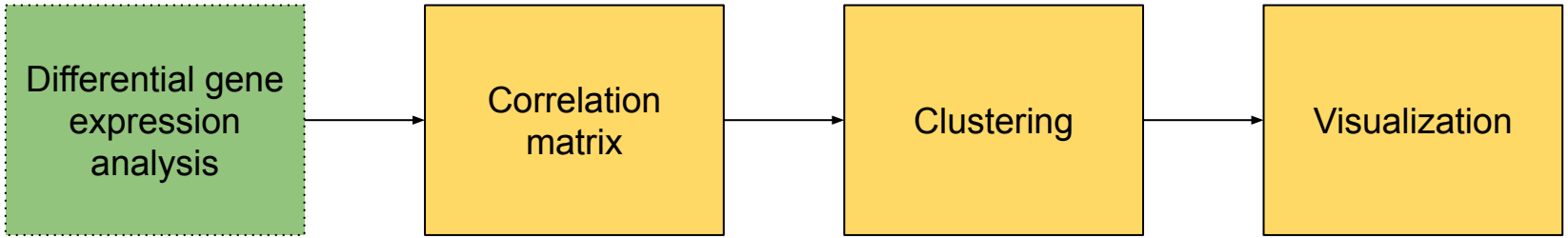
Workflow pipeline

Colorectal cancer data set: cells from patients with and without tumor.

Question: Do the correlation patterns of genes in certain cell subtypes change in healthy vs. cancer tissue?

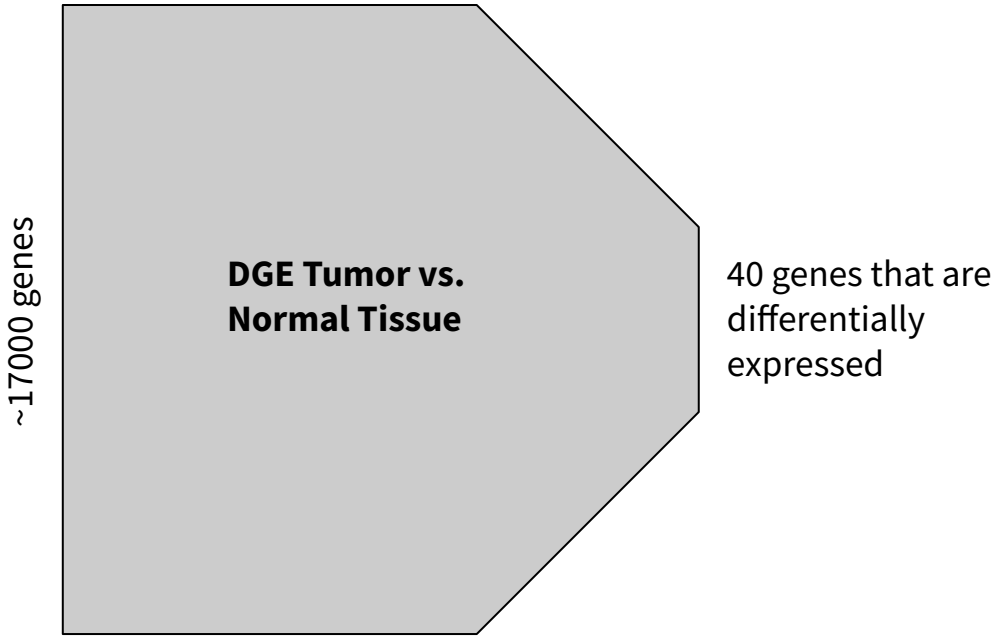
Workflow pipeline

Colorectal cancer data set: cells from patients with and without tumor.



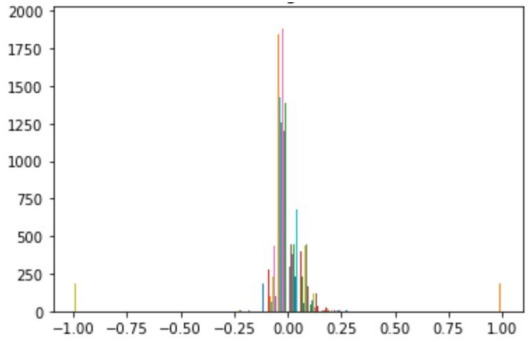
Question: Do the correlation patterns of genes in certain cell subtypes change in healthy vs. cancer tissue?

Differential gene expression analysis

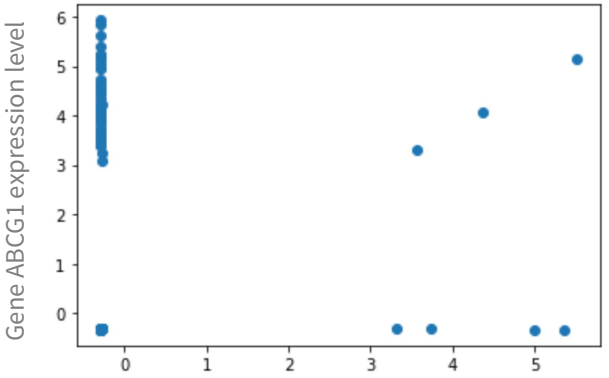


Pearson correlation

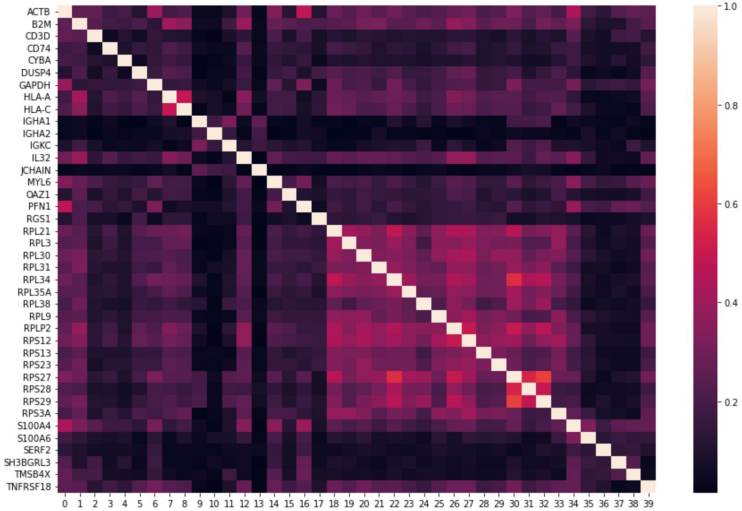
Heatmap of Pearson correlation matrix



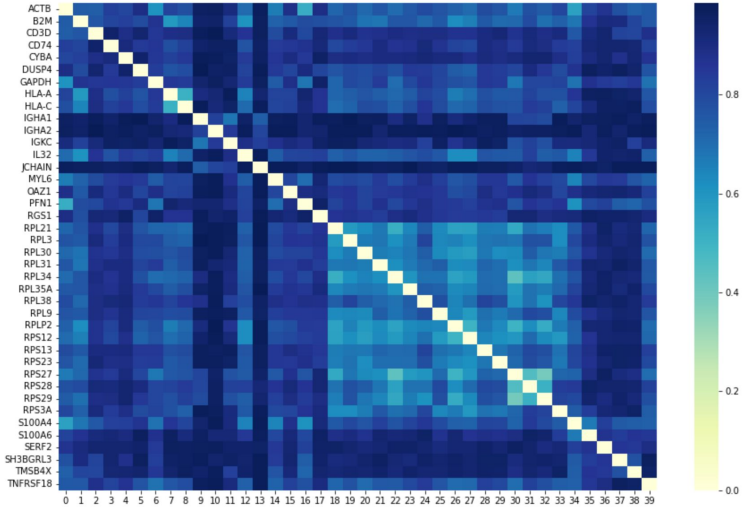
Histogram of the correlations



Distance correlation (Székely et al., 2007)



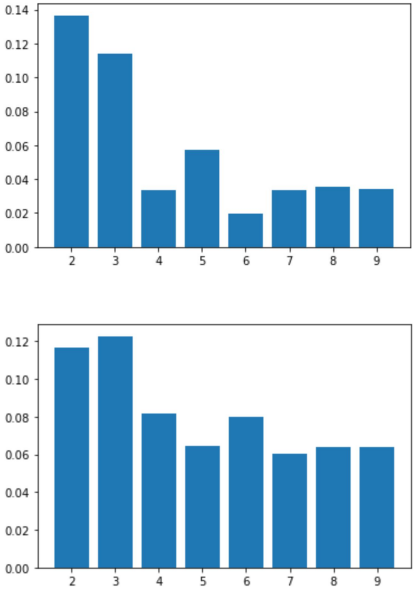
$$d(x,y) = 1 - |\rho(x,y)|$$



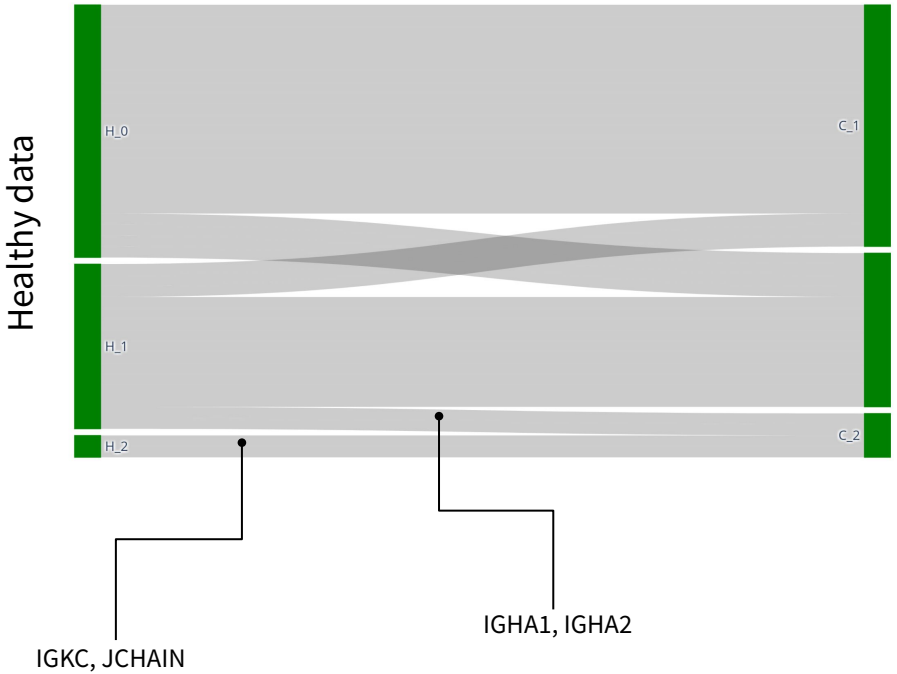
DC matrix

Distance matrix

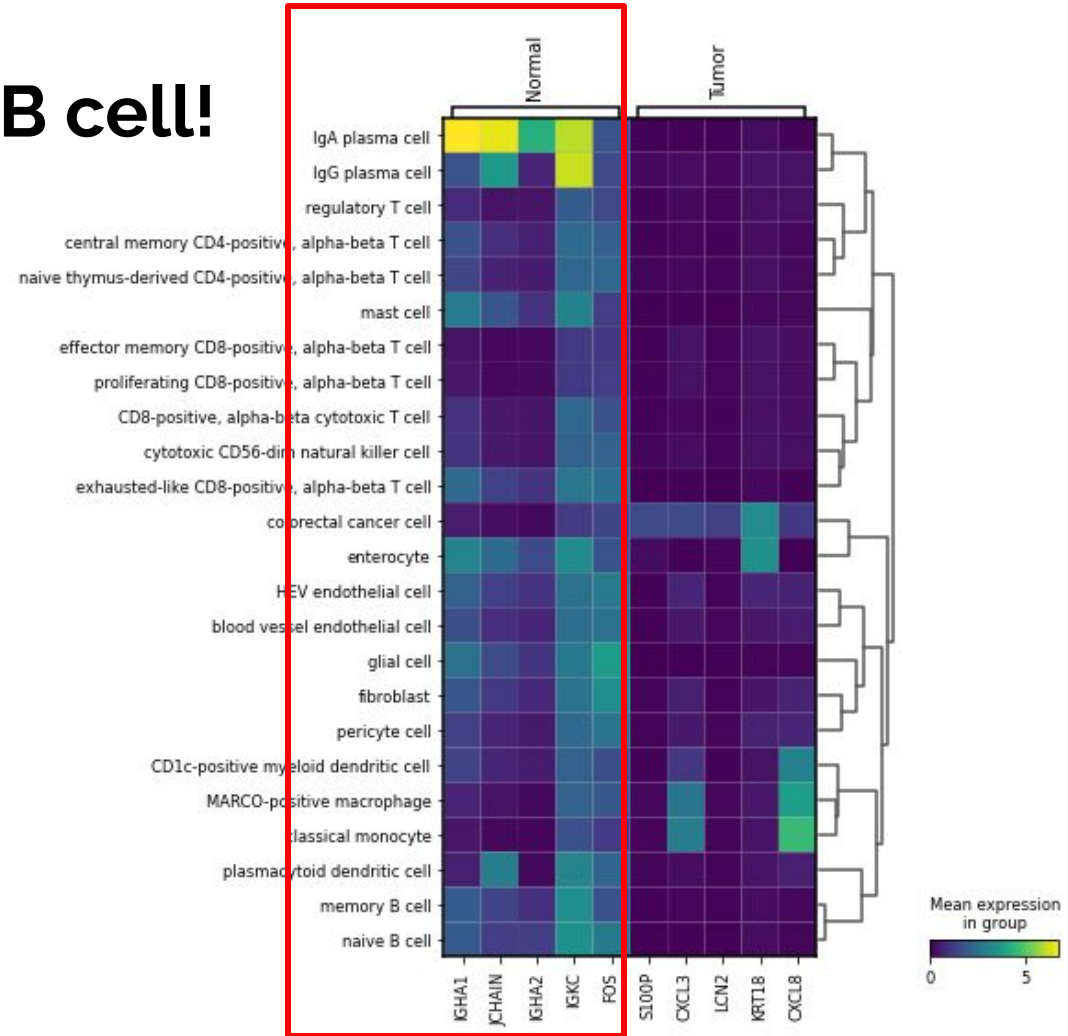
Comparing Clusters



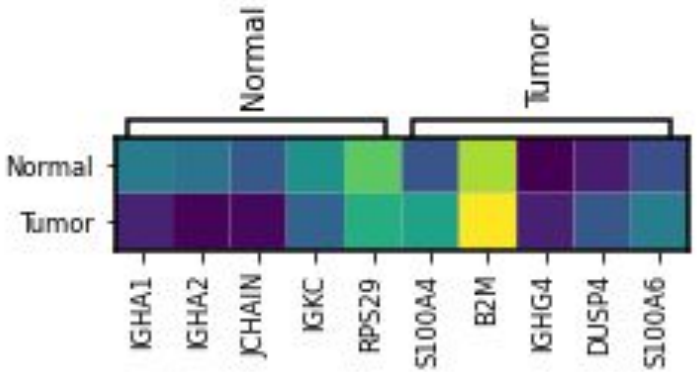
Average silhouette scores for healthy (top) and cancer (bottom) data



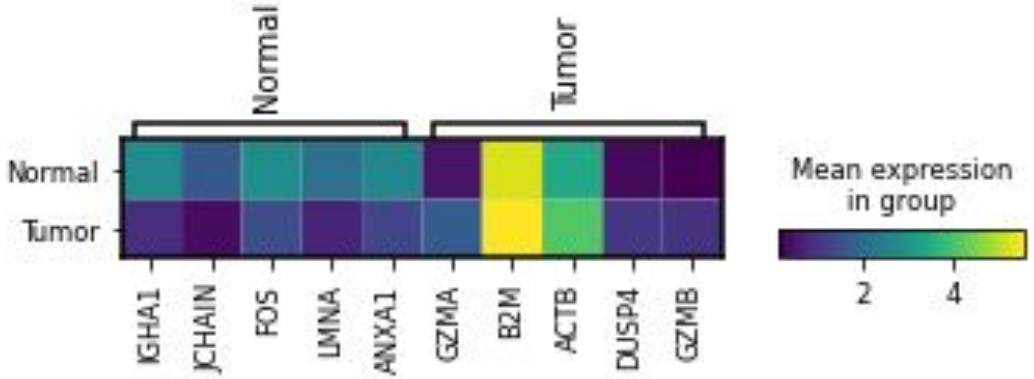
DGE - Everything is a B cell!



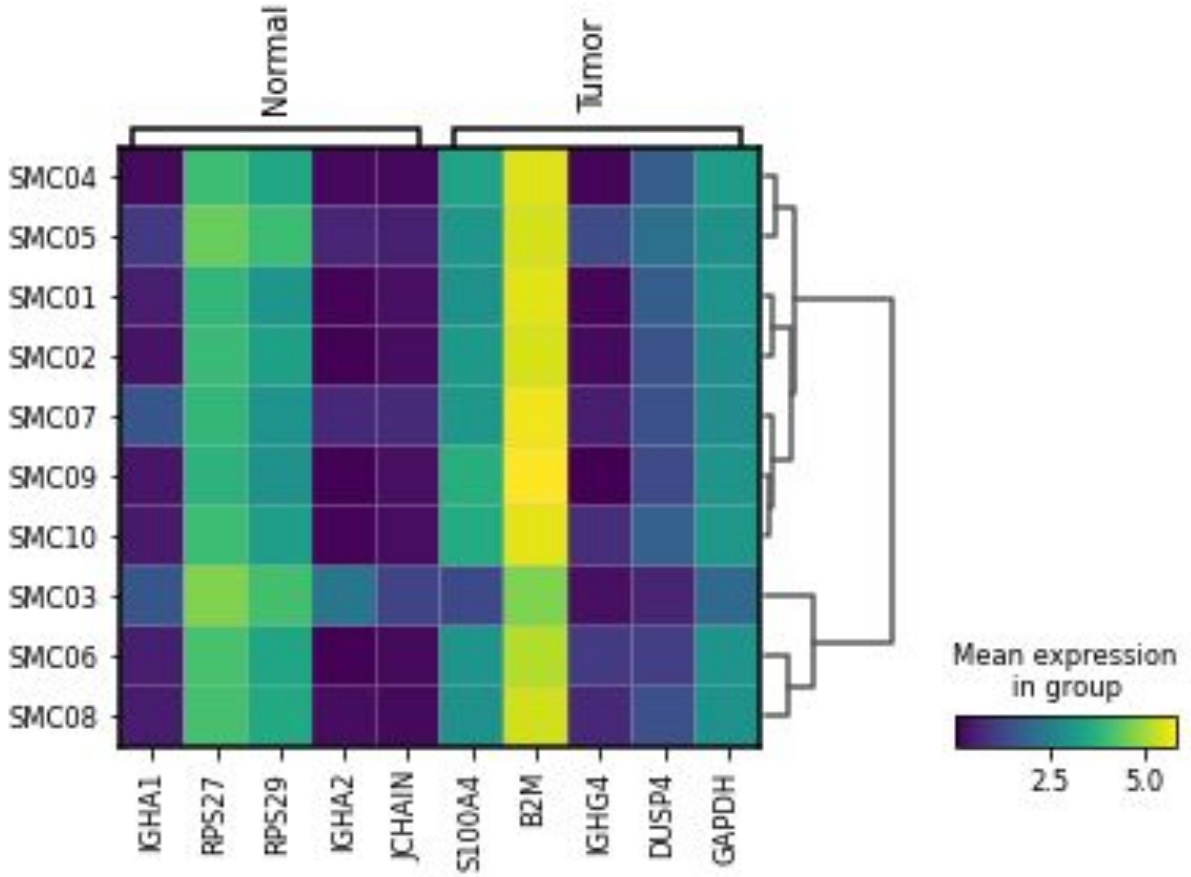
T-Reg DGE



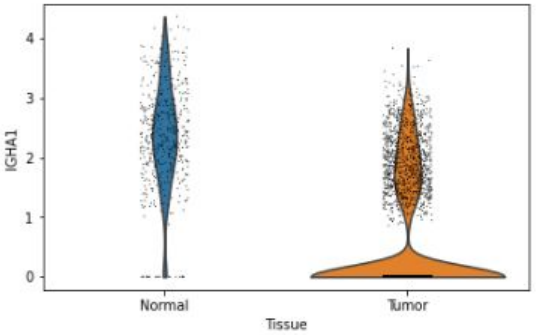
CD4 αβ+ DGE

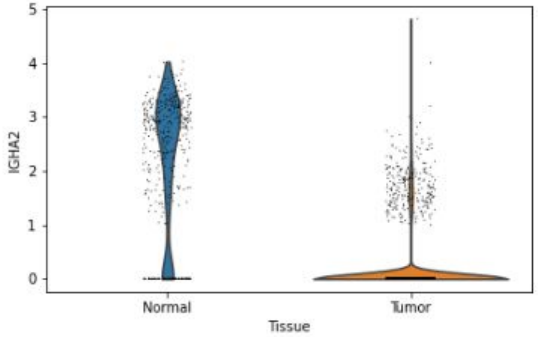


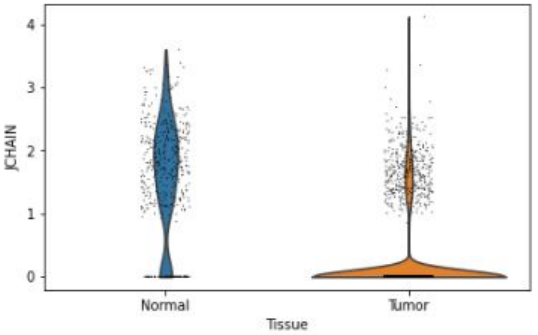
T-Reg DGE Per Patient



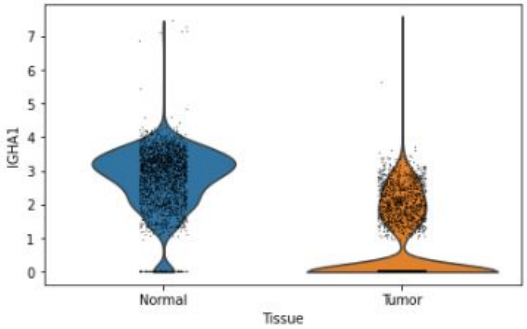
B cell gene expression among T-Regs

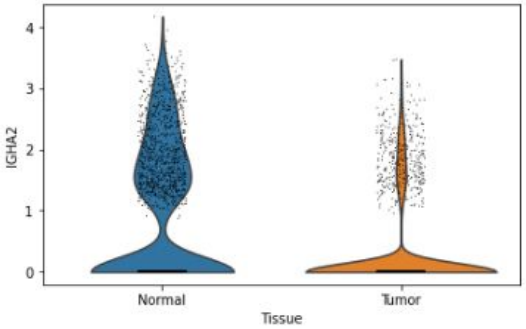


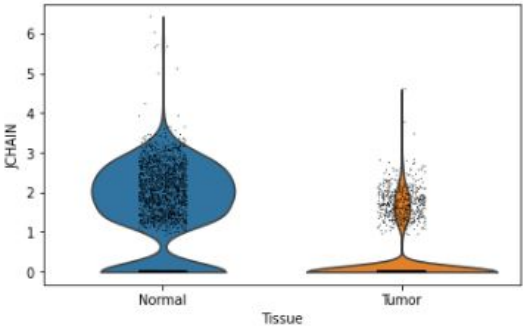




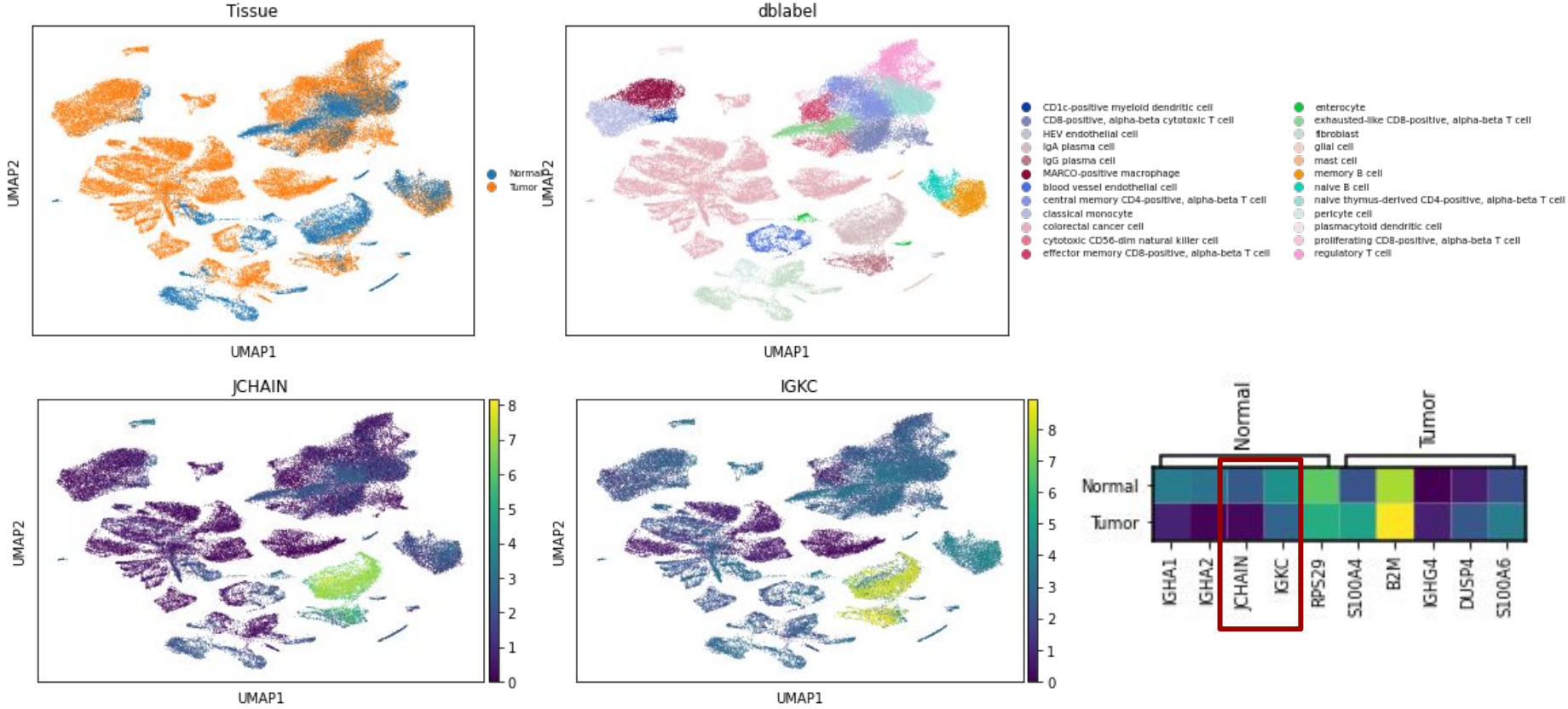
CD4+ αβ



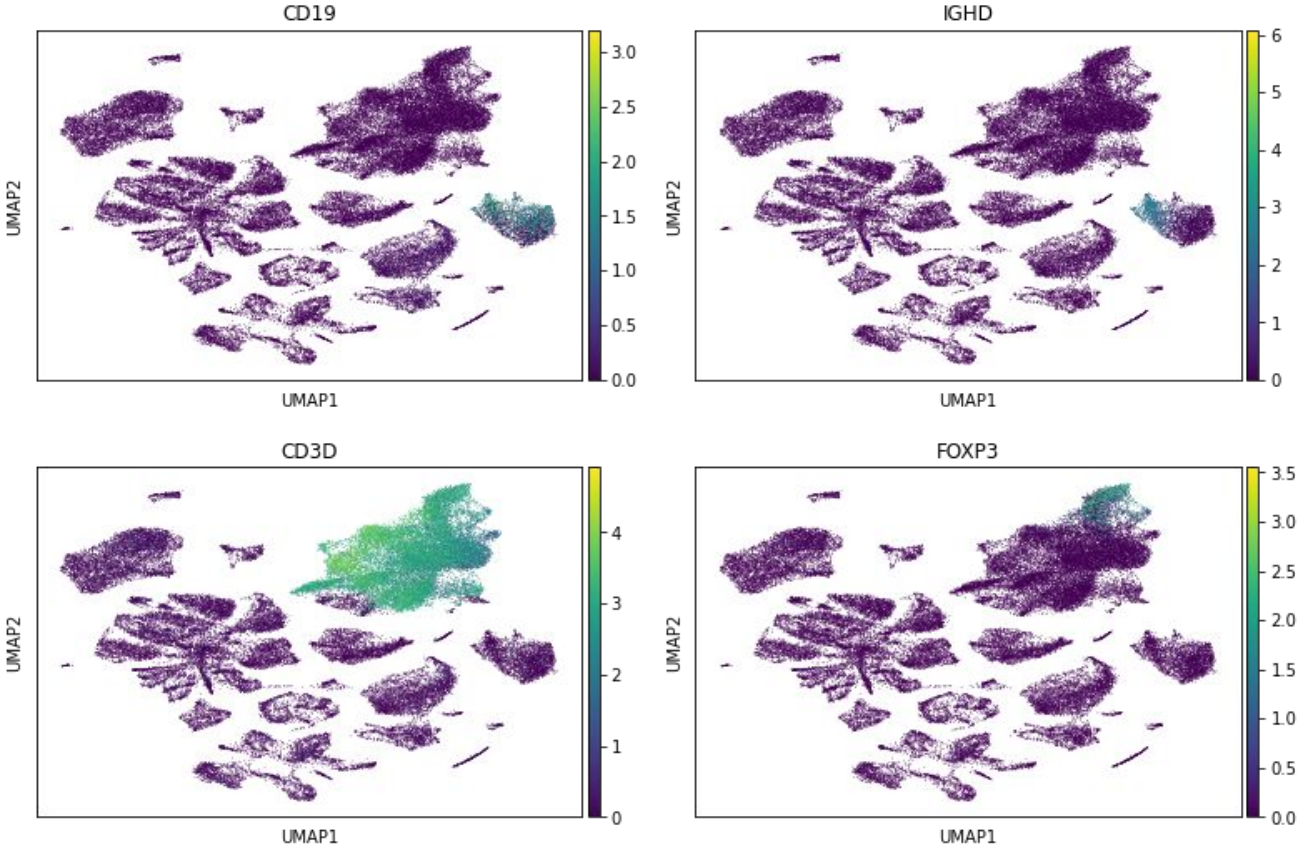




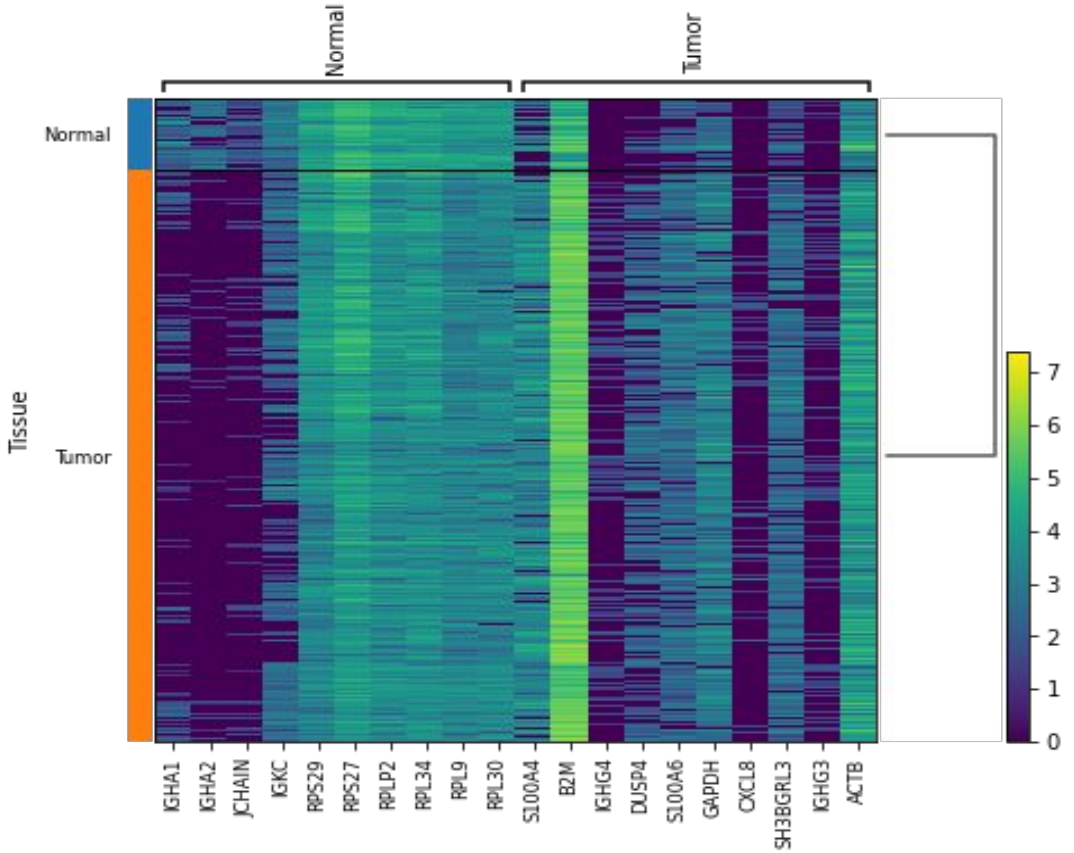
B cell gene expression



Not data is erroneous

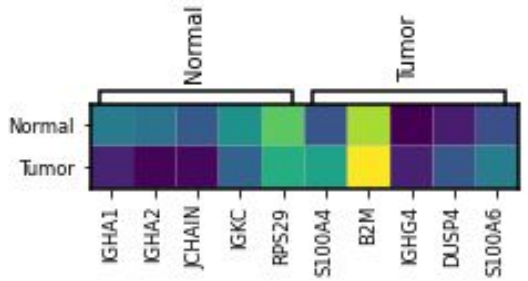


Unbalanced datasets



Number of Tregs in Tissue

Normal	Tumor
463	3716



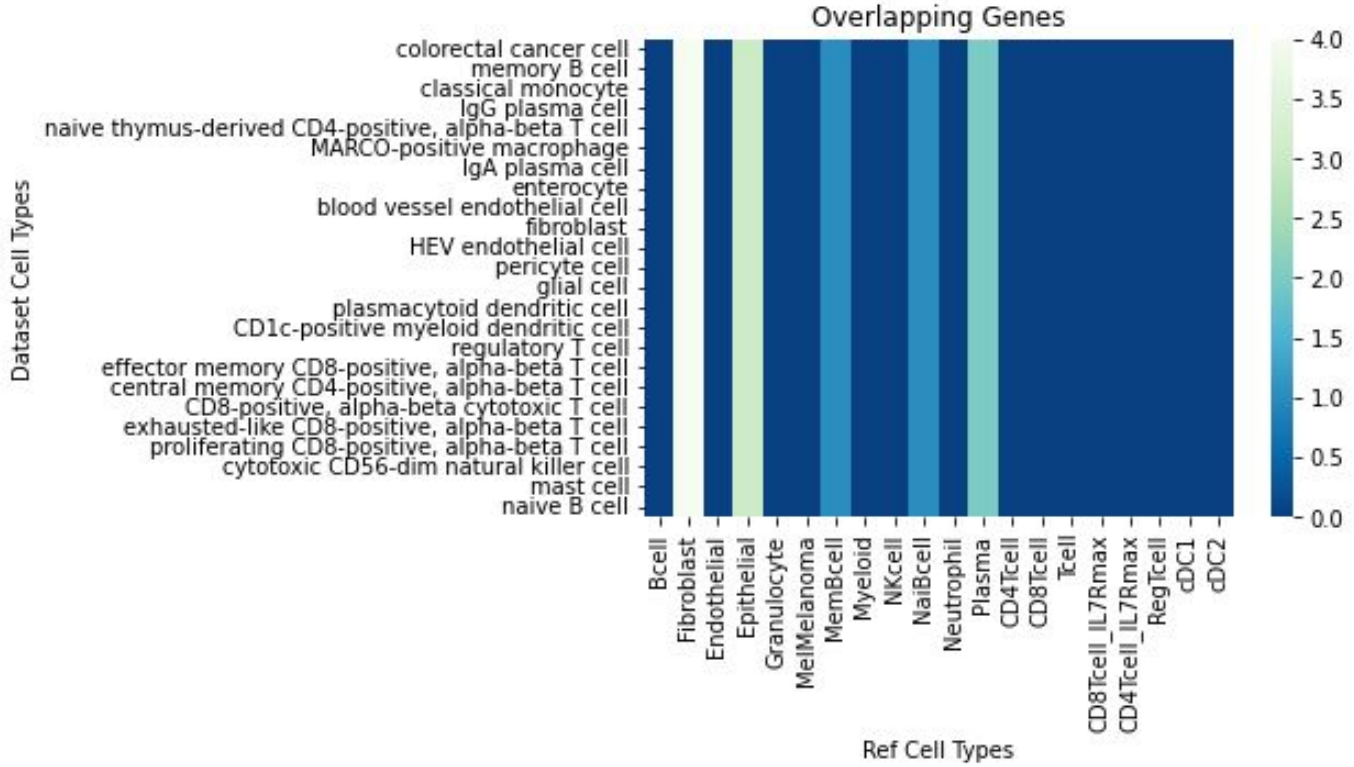
Cell Type Label QC

Algorithm 1: Cell Type Label QC From Reference Marker Genes

input : scRNA-seq AnnData obj. $\mathcal{A}$, dictionary $\mathcal{M}$ of reference cell type markers from GMT file $\mathcal{G}$
Hyperparameters `flag_threshold`
output: Dictionary $\mathcal{D}_{flag}$ of Flagged Cell Type Labels
Heatmap $\mathcal{H}$ of Number of Genes

Initialize dictionary Dictionary $\mathcal{D}_{\texttt{flag}}$ to store QC flagged cell types
Initialize list `unique_celltypes` of unique cell types in $\mathcal{A}$
Execute Differential Gene Expression analysis on $\mathcal{A}$
for $i \leftarrow 0$ **to** *length(`unique_celltypes`)* **do**
    for *all key value pairs (k, v) in* $\mathcal{M}$ **do**
        Initialize list `ref_genes` $= v$
        Initialize list `genes` of differentially expressed genes from `unique_celltypes`$[i]$ in $\mathcal{A}$
        $\mathcal{S} =$ (`genes` $\cap$ `ref_genes`)
        if *length* $\mathcal{S}$ `flag_threshold` **then**
            Append $\mathcal{S}$ and ref celltype k to $\mathcal{D}_{flag}$[`unique_celltypes`$[i]$]
        end
    end
end
Create Heatmap, $\mathcal{H}$ with y axis as cell type in $\mathcal{D}$, x axis as ref cell type in $\mathcal{M}$
and cell colors as number of overlapping genes
return $\mathcal{D}_{flag}$, $\mathcal{H}$

Ex Output - Algorithm 1



Summary

- Differential gene expression analysis as Feature selection
- Association measure variations: Pearson, Spearman, distance correlation, ...
- Clustering algorithms variations: k-means, hierarchical clustering, ...
- Visualization variations: heatmap, Sankey diagram (river flow diagram), ...

Outlook:

- Other association measure variations
- Clustering: mixture modeling, ...
- Further visualizations
- Workflow adaptability for other datasets

Considerations and limitations

Quality control

- B cell genes expressed in all cells?
- unbalanced datasets (Number of patients, number of cells, …)

Feature selection

- DGE testing – confounding factors,
- different approaches instead of DGE testing

Biological interpretation

- correlation vs. causation vs. coincidence