¶ 02 Manual annotation of cell clusters ¶

Kane Toh*

¶ *2022-12-10* ¶

Contents

1 Summary (tl;dr)

We label the seurat clusters using a revised list of cluster annotations and remove two clusters, resulting in a total of 11 identified clusters consisting of 8811 cells.

Separately, we identified the differentially expressed genes between the KO and WT conditions. To assess which biological functions and pathways are over-represented in the KO condition, we identify genes that are upregulated in the KO condition (defined as having an average log2 fold-change > 0.5), and carried out enrichment analysis on this gene set against several popular curated databases.

We advise reviewing the list of enriched pathways to assess whether the results match up with a priori biological expectations. This list should be useful as a starting point for experimental validation.

*Genomics and Data Analytics Core (GeDaC), Cancer Science Institute of Singapore, National University of Singapore; kanetoh@nus.edu.sg

2 Methods

2.1 Alignment with the Parse Biosciences split-pipe program

Reads were aligned to the mouse genome (GRCm39) with the split-pipe (v0.9.6p) program from Parse Biosciences with the release 107 Ensembl GTF annotations. Both sublibraries were generated separately with the `--mode all` command. The filtered gene expression matrices across all samples, across both sublibraries, were then combined into a single Seurat (v4.3.0) object for downstream analysis.

2.2 Quality control and normalisation

Next, cells were filtered based on the following 3 criteria:

- Number of detected genes (nFeature_RNA) >300 &
- Number of UMI molecules (nCount_RNA) > 300 &
- Percentage mitochondrial RNA (percent.mt) < 10

This resulted in a total of `10,016` cells that were retained for further analysis. Cells were then log-normalised, followed by the identification of the top 2000 variable features in the dataset. Features in the dataset were then scaled and centered. To evaluate the effects of cell cycle heterogeneity, cells were assigned a cell cycle score based on the expression of G2/M and S phase markers for Mus Musculus that were obtained from the tinyatlas github repository shared by the Harvard Chan School Bioinformatics Core.

2.3 Dimensional reduction, clustering and marker gene identification

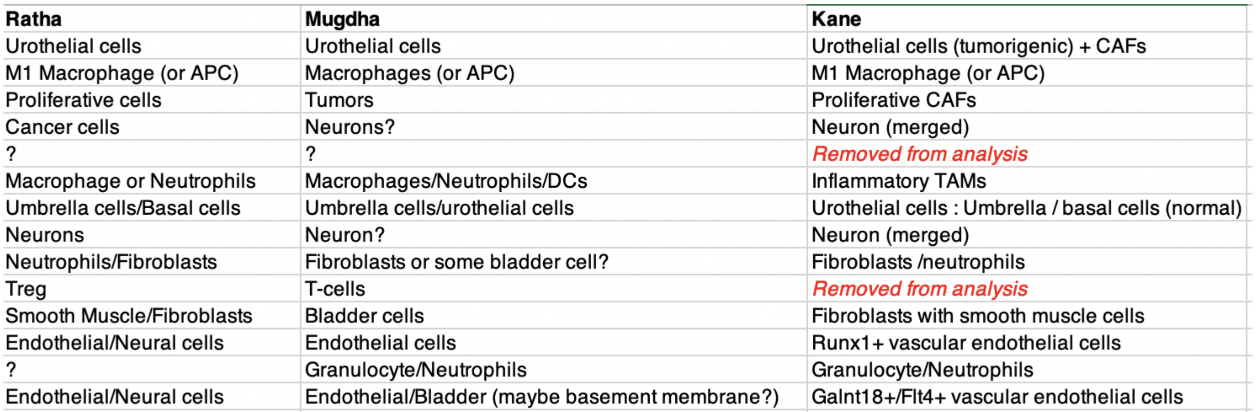
Graph-based clustering was implemented by first generating a shared Nearest-neighbor graph and then optimising the modularity function with the Louvain community detection algorithm. This led to the identification of 12 clusters. The Uniform Manifold Approximation and Projection (UMAP) dimensional reduction method was applied to embed cells in lower-dimensional space for ease of visualisation and for cell clustering. Cluster biomarkers were then identified via the `FindAllMarkers` function, using the Wilcoxon rank sum test for differential gene expression testing. Clusters were then manually annotated by curating gene signatures from literature. One of the clusters was removed as it did not correspond to known cell types. In addition, these cells expressed higher levels of heat shock protein transcripts (Hsp90aa and Hsp90ab), suggesting that these are stressed/dying cells.

2.4 Over-representation analysis of differentially expressed genes

GSTT2-KO samples were compared against the WT samples using the function `FindMarkers`, with the first group labelled as pct.1 and the second group as pct.2. Genes upregulated in the GSTT2-KO samples were identified as those with an average log2 fold-change > 0.5, and over-representation analysis was carried out using the `clusterProfiler` package. These differentially expressed genes were tested against the KEGG pathway and module database, GO knowledge base, MSigDB C5,C6,C7,C8 and hallmark datasets and the Reactome dataset, with a Benjamini-Hochberg adjusted p-value cutoff of 0.05. The background gene list was defined to be the set of genes that were expressed in at least 1% of all cells that pass the quality control.

3 Cluster annotations with evidence for annotation

Current annotation state:



Ratha	Mugdha	Kane
Urothelial cells	Urothelial cells	Urothelial cells (tumorigenic) + CAFs
M1 Macrophage (or APC)	Macrophages (or APC)	M1 Macrophage (or APC)
Proliferative cells	Tumors	Proliferative CAFs
Cancer cells	Neurons?	Neuron (merged)
?	?	*Removed from analysis*
Macrophage or Neutrophils	Macrophages/Neutrophils/DCs	Inflammatory TAMs
Umbrella cells/Basal cells	Umbrella cells/urothelial cells	Urothelial cells : Umbrella / basal cells (normal)
Neurons	Neuron?	Neuron (merged)
Neutrophils/Fibroblasts	Fibroblasts or some bladder cell?	Fibroblasts /neutrophils
Treg	T-cells	*Removed from analysis*
Smooth Muscle/Fibroblasts	Bladder cells	Fibroblasts with smooth muscle cells
Endothelial/Neural cells	Endothelial cells	Runx1+ vascular endothelial cells
?	Granulocyte/Neutrophils	Granulocyte/Neutrophils
Endothelial/Neural cells	Endothelial/Bladder (maybe basement membrane?)	Galnt18+/Flt4+ vascular endothelial cells

Figure 1: Proposed updated cluster annotations

3.1 Urothelial cells

Cluster 1: `Urothelial cells (tumorigenic) + CAFs`

Proposed identity: potentially tumorigenic urothelial cells undergoing EMT along with CAFs.

Evidence:

1. Located adjacent to cluster 2 which contains proliferating CAFs
2. Distinguished from cluster 6 (Umbrella/Basal urothelial cells (normal)) by :

- Elevated HMGA1 and Vim: HMGA1 is a prognostic marker in bladder cancer Yang et al., 2011

Role in EMT together with Vimentin Ding et al., 2014

- Evidence for upregulation of oncogenic PD-L1: Upregulation of Cald1 Li et al., 2021

3. Presence of CAFs in the tumor microenvironment:

- Upregulation of Cald1 Du et al., 2021
- Presence of CAF markers: S100a4, Pdgfa, Vim, Ddr2 Chen et al., 2021

Cluster 7: `Urothelial cells : Umbrella / basal cells (normal)`

- Evidence:

1. High expression of urothelial cell marker genes such as Upk1a, Upk1b, Upk3a and Krt5.

3.2 Immune cells

3.2.1 Macrophages

Cluster 5: `Inflammatory TAMs with CAFs`

Evidence:

1. Distinguished from cluster 0 (M1 macrophages) by:

- Elevated expression of Cxcl2, Cxcl3, Il1rn, Il1b, G0s2 Table 1 of Ma et al., 2022
- Also by S100a9 (80% expressed), S100a8 (60% expressed)

Cluster 0: `M1 macrophages/APC`

Evidence:

1. High expression of M1 macrophage markers such as CD86, H2-Ab1, H2-Eb1 and Nos2

3.2.2 Granulocyte/Neutrophils

Cluster 10 - `Granulocyte/Neutrophils`

Evidence:

1. High expression of canonical neutrophil markers S100a9 (60%), Csf3r (33%).

3.3 Neurons

Clusters 4 and 6 (combine) : `Neurons`

Evidence:

1. High expression of neuronal markers such as Tenm2, Cntnap2,Csmd1 etc.

3.4 Fibroblasts

Cluster 2: `proliferative CAFs`

Evidence:

1. High expression of mitotic markers (also see cell cycle imputation diagram) e.g. Top2a, MKi67, Cenpf, Cenpe and Incenp
2. Expression of fibroblast gene S100a41 in about 67% of cells

Cluster 8: `fibroblasts/neutrophils`: Evidence:

1. Expression of Col3a1, Col1a1, Cola2, Gpx3, Clec3b

Cluster 9: `fibroblasts with smooth muscle cells`

Evidence: 1. Expression of des, TPM2, Gpx3

3.5 Endothelial cells

Clusters 11 and 12 distinguished by expression of several marker genes.

Cluster 11: `Runx1+ vascular endothelial cells`

Cluster 12: `Galnt18+/Flt4+ vascular endothelial cells`

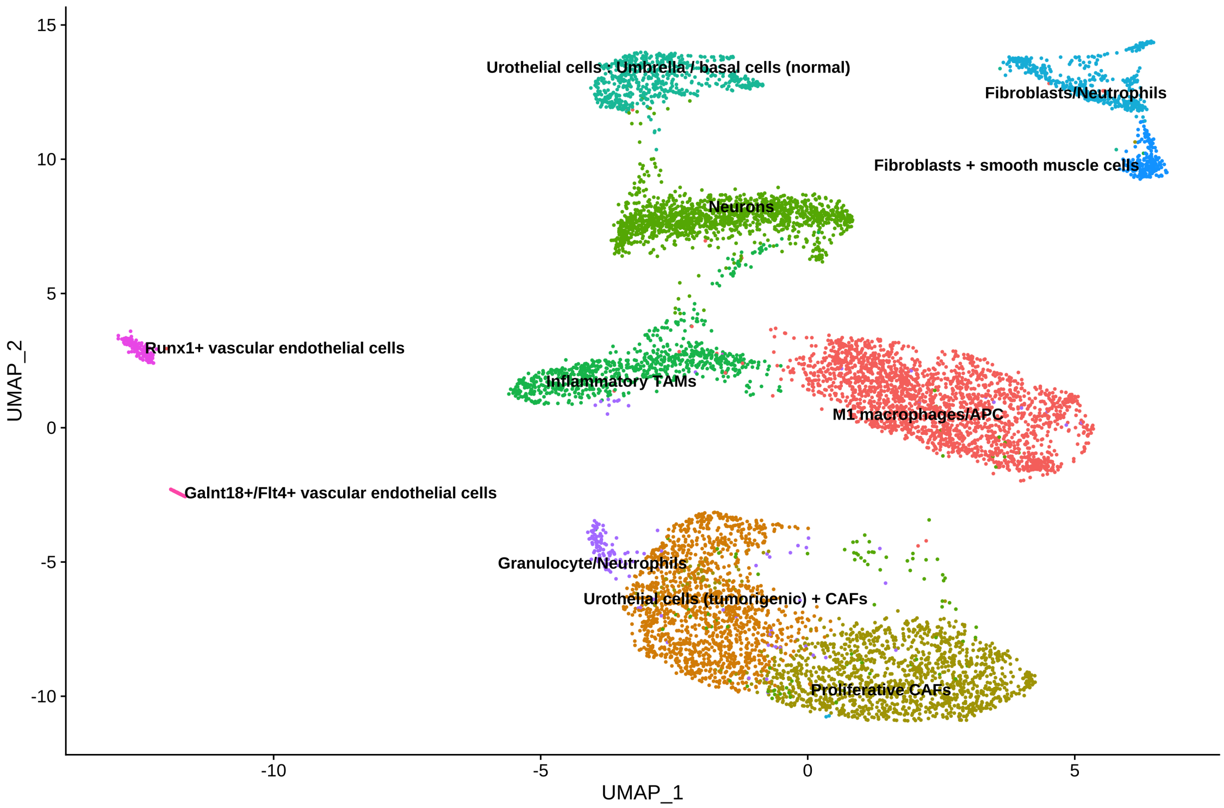


Figure 2: UMAP plot after manual annotation of clusters

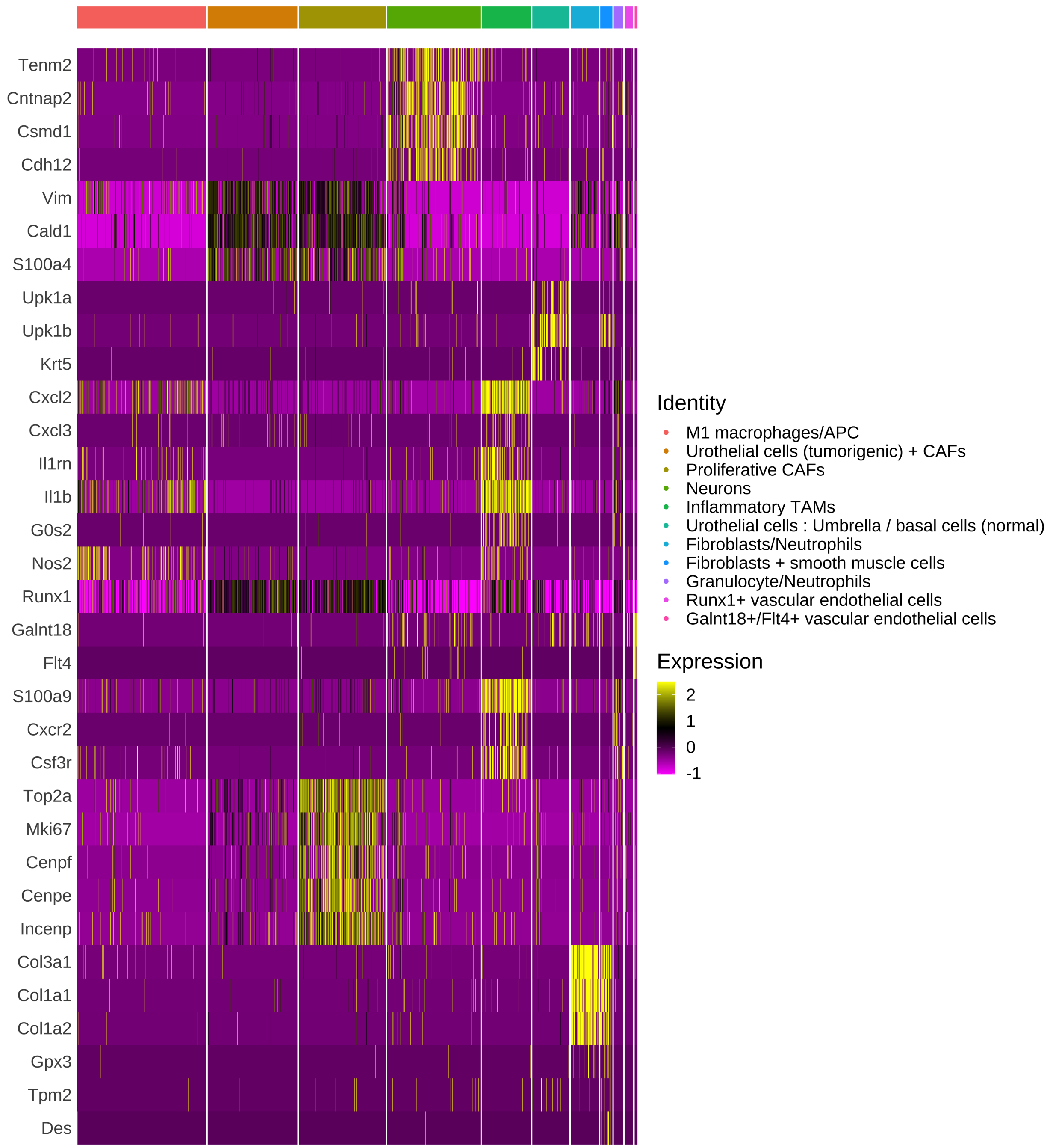


Figure 3: Heatmap of manually selected marker genes to highlight the different cell types identified. The heatmap distills the information from the various violin plots

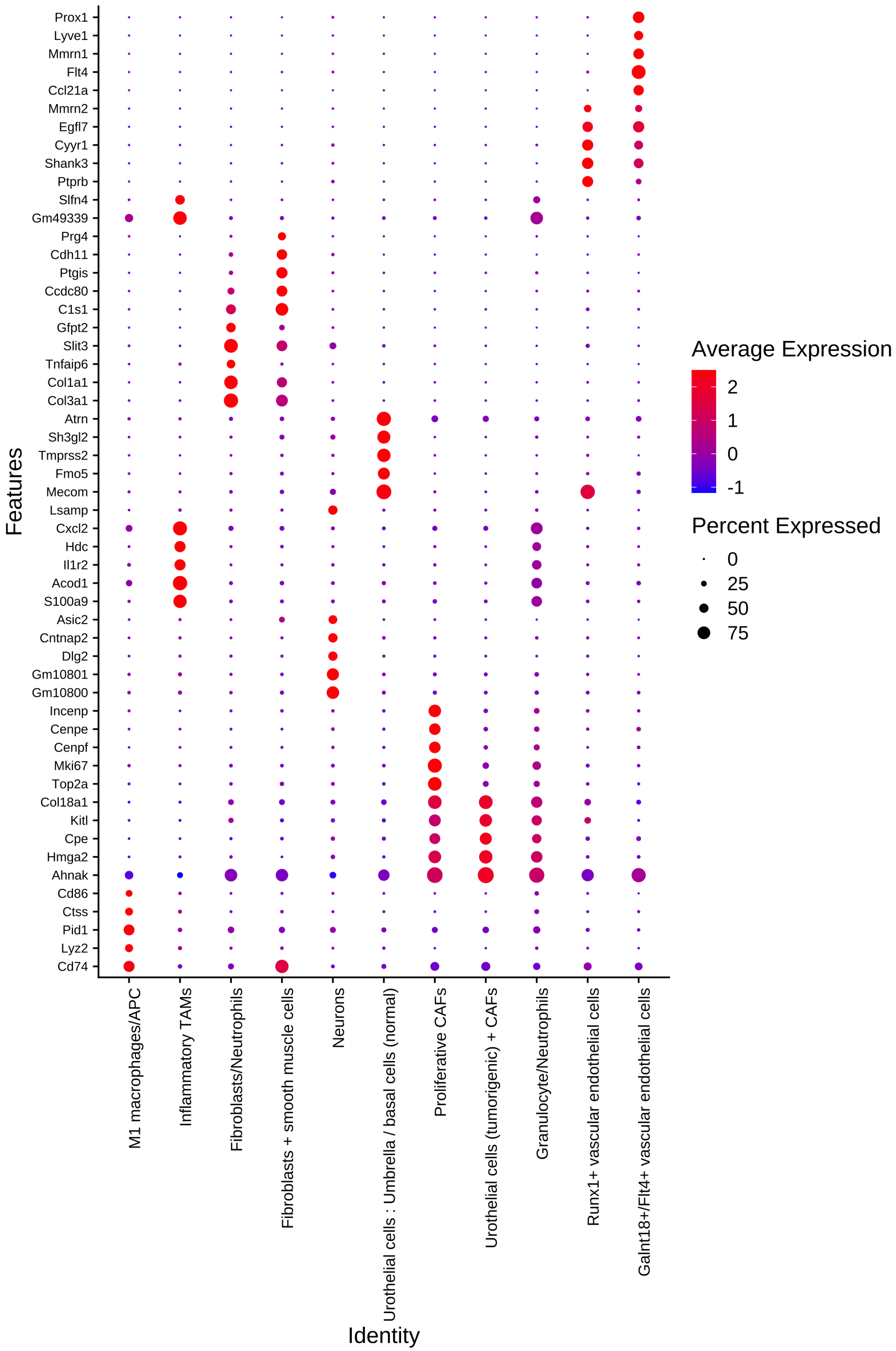


Figure 4: Dotplot of the top 5 positively expressed marker genes for each of the 11 identified clusters

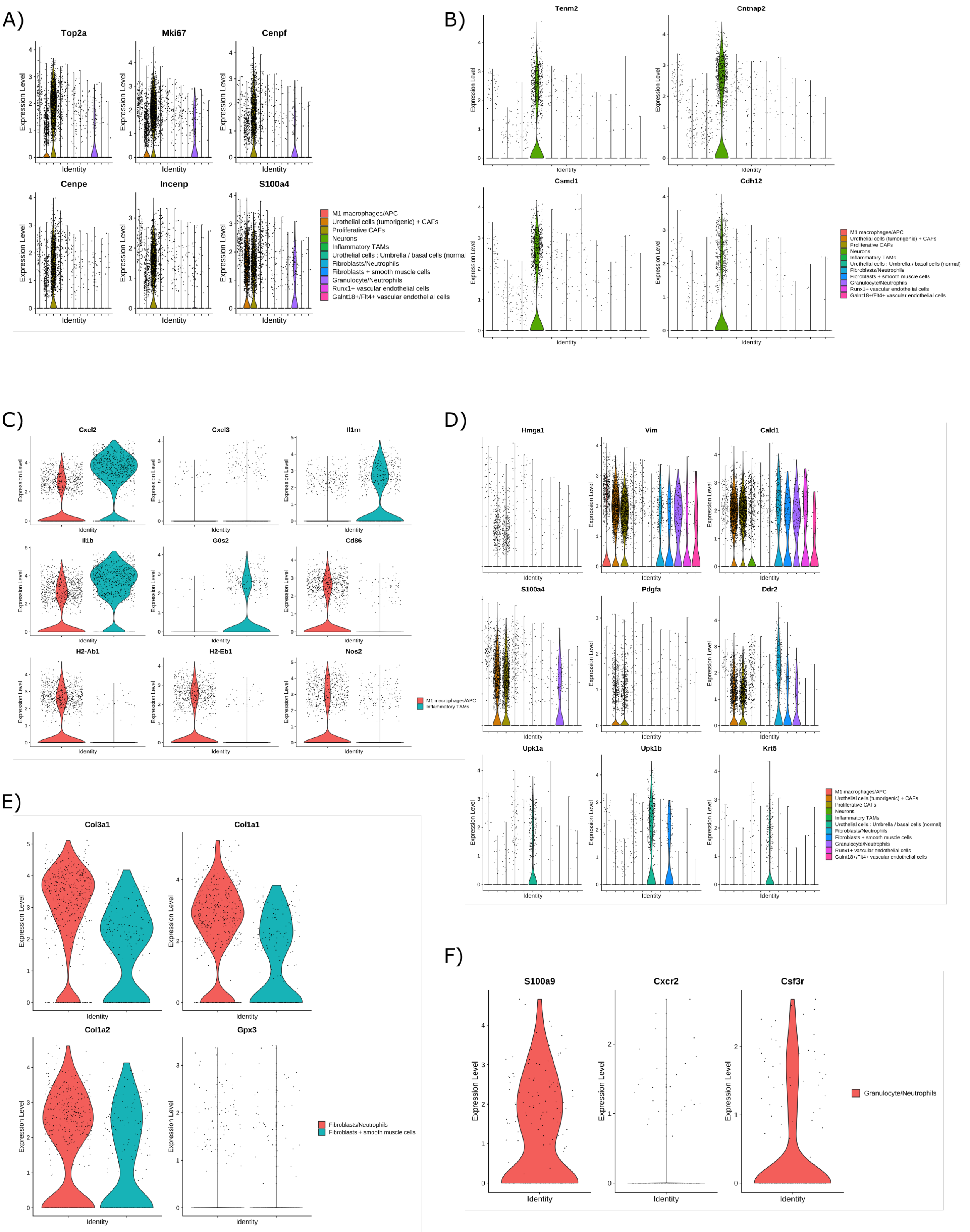


Figure 5: Selection of violin plots to highlight expression of marker genes for each cluster. A) Proliferative CAFs B) Neurons C) Distinguishing between M1 macrophages and inflammatory TAMs D) Urothelial cells and CAFs E) Fibroblasts F) Granulocyte/Monocyte

Table 1: DEGs between KO and WT

X	avg_log2FC
Peak1	1.723444
Rpph1	1.668899
Gm19951	1.632383
Gphn	1.522817
Plec	1.277845
Ahnak	1.241102
S100a9	1.198081
Mki67	1.098485
Slfn4	1.098240
Cmss1	1.081260

4 Differentially expressed genes between KO and WT condition

We then identified a list of genes that were upregulated in the GSTT2-KO condition vs the WT condition (`KO_vs_WT_marker_genes.csv`) and performed over-representation analysis (ORA) on the genes that were upregulated in the KO condition (average logFC $>$ 0.5) to assess whether known biological functions or processes are over-represented. Results in the form of csv files and dotplots can be found in the `enrichment_analysis` folder.

An example dotplot for the ORA analysis conducted against the MSigDB hallmark gene sets is shown in Figure 6, highlighting the overrepresentation of genes in the G2M checkpoint, MYC targets, IL2 STAT5 signalling gene sets etc.

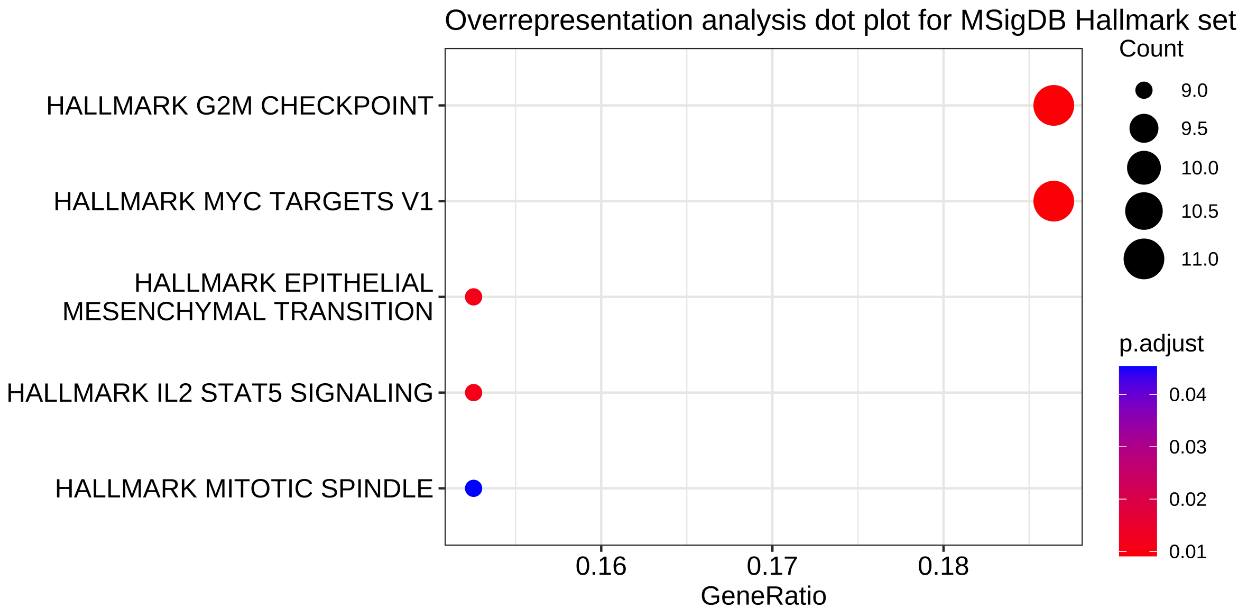


Figure 6: Dotplot showing the MSigDB hallmark gene sets that are over-represented in the set of genes enriched in the GSTT2-KO samples relative to the WT samples.

5 Next steps

1. Identification of tumor cells

Initial attempts at identifying the tumor cells via expression of the human prostate-specific antigen gene (KLK3) was unsuccessful. We will attempt these two methods other recently published methods to identify tumor cells:

- Dohmen et al., 2022
- Gasper et al., 2022

2. Cell-cell communication (CellChat)
3. Pseudotime analysis (monocle3)
4. RNA-velocity (dynamo)