scRNAseq analysis of Gstt2-KO mouse implanted with bladder tumors

Kane Toh*

2023-03-17

Contents

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

1 Summary

We label the seurat clusters using a revised list of cluster annotations, resulting in a total of 11 identified clusters consisting of 8811 cells. Next, using the *distinct* R package, we identified the differentially expressed genes between the KO and WT conditions for each cluster, and carried out an overrepresentation enrichment analysis on this gene set against the GO, KEGG and MSigDB databases.

2 Background

Bladder tumors were implanted orthotopically into 3 WT and 4 GSTT2-KO mice at 3-4 months of age and then treated with 4 instillations of M. bovis BCG, following which the bladders were harvested and isolated as single cells for scRNA-seq.

3 Methods

3.1 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.

3.2 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 (For more details, see QC). 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.

3.3 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. These cells expressed higher levels of heat shock protein transcripts (Hsp90aa and Hsp90ab), suggesting that these are stressed/dying cells.

We observed that samples intermix across sublibraries, sample conditions and mouse samples, indicating the absence of strong batch effects arising from these factors (Figure 1A-C). Several samples, such as the KO-BL, WT_TR and WT_TRBR, did not contribute to the CAFs and granulocyte/monocyte clusters, reflecting heterogeneity in cell type contribution across the samples (Figure 1D). In addition, we also corrected for the mouse sample batch effect with Seurat or Harmony and did not observe substantial differences in the marker gene list (not shown). Therefore, we proceeded with our analysis without batch effect correction.

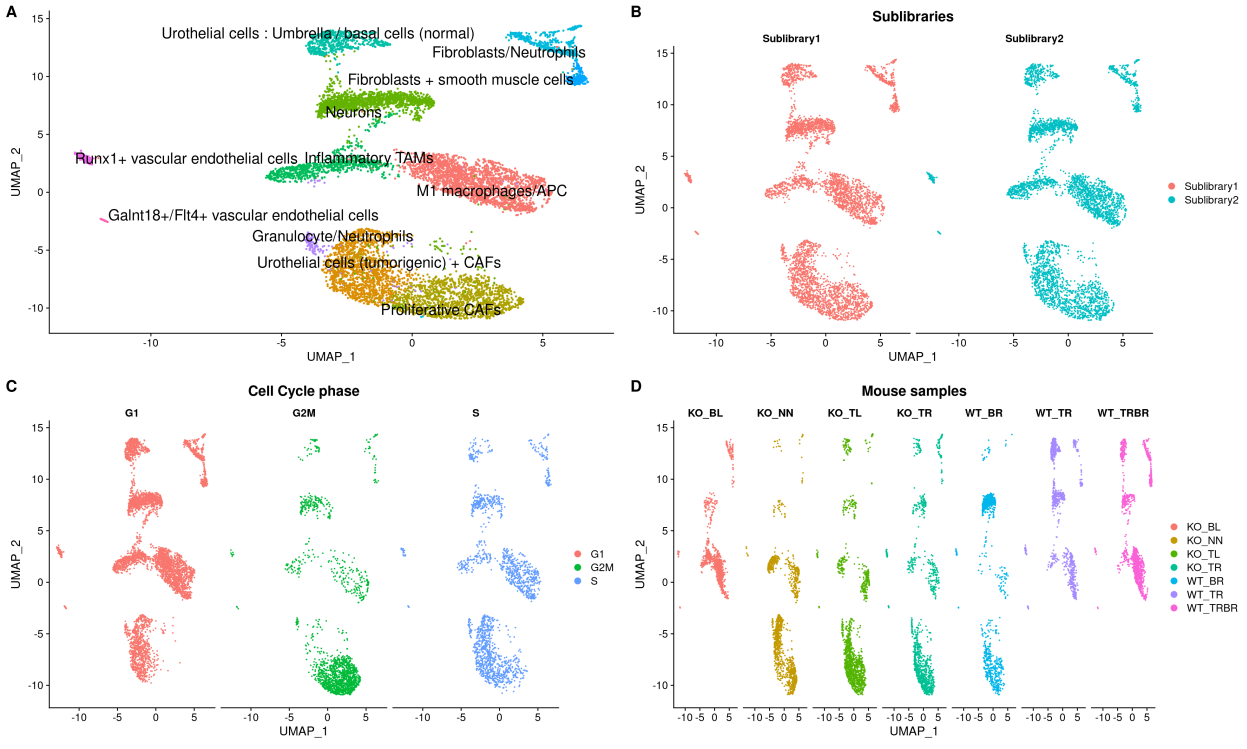

Figure 1: Batch effect assessment

3.4 Differential state analysis between the GSTT2-KO and WT samples

To investigate the differences in gene expression patterns between the GSTT2-KO and WT samples, we used the *distinct* approach for differential analysis (Tiberi et al., 2022: preprint). Distinct takes a full distribution approach to differential analysis, and thus can detect differences in gene expression such as differential variability and differential modality, on top of shifts in mean abundance between the two sample conditions. With the exception ofthe `P_4` parameter which we set to 20000 following the author's recommendation, all other parameters were set to the default, along with an adjusted p-value of 0.01.

Over-representation analysis of the DE's identified in *distinct* was carried out using the `clusterProfiler` package. These differentially expressed genes were tested against the KEGG pathway and module database, GO knowledge base, KEGG and MSigDB C5,C6,C7,C8 and hallmark datasets, 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.

4 QC

4.1 Phred quality scores (Q30) are similar between sublibraries (~85-90% high quality bases)

From Figure 2A, we find that both sublibraries have very similar fraction of reads with Phred quality scores of 30 or above for barcodes 1-3 (bc1_Q30, bc2_Q30 and bc3_Q30) and polyN bases (polyN_Q30) on read2, as well as the bases from the transcript sequence on read1 (DNA_Q30). A phred score of 30 for a base means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%, and therefore the greater the value of these fractions, the greater the fraction of high quality reads. From the data, we note that the range of values lies between around 0.85 - 0.90 for the reads from both sublibraries.

4.2 A substantial proportion of barcodes are invalid, likely due to lower quality scores

The `valid_barcode_fraction` for both sublibraries are around 60%, which implies that around 40% of the total reads from both sublibraries are discarded. The Phred quality scores for the barcode reads (bc1_Q30, bc2_Q30 and bc3_Q30) suggest that 40% of the barcodes did not pass the quality filter.

4.3 Alignment statistics are similar between sublibraries (~69% uniquely mapped reads)

Turning to the STAR read alignment statistics (Figure 2B), we observe that the proportion of reads that align uniquely to the genome is around 67-68% for both sublibraries. About less than 50% of the reads map to the transcriptome for both sublibraries. Both statistics tend towards the lower end as we would expect at least 70% or more uniquely mapped reads against a well-studied reference genome, and 50-60% of reads to map to the transcriptome. Fewer reads align to multiple regions of the genome, and even fewer reads are discarded as they map to too many loci. Thus, it appears that around 30% of the reads are unmapped for both sublibraries due to other unspecified reasons, presumably due to their lower base quality.

4.4 The KO samples, with the exception of KO_BL, have higher sequencing depth than the WT samples

With the exception of the KO_BL sample, the KO samples are sequenced to a greater depth (greater number of reads) than the WT samples (Figure 2C). In addition, these samples have a greater median number of

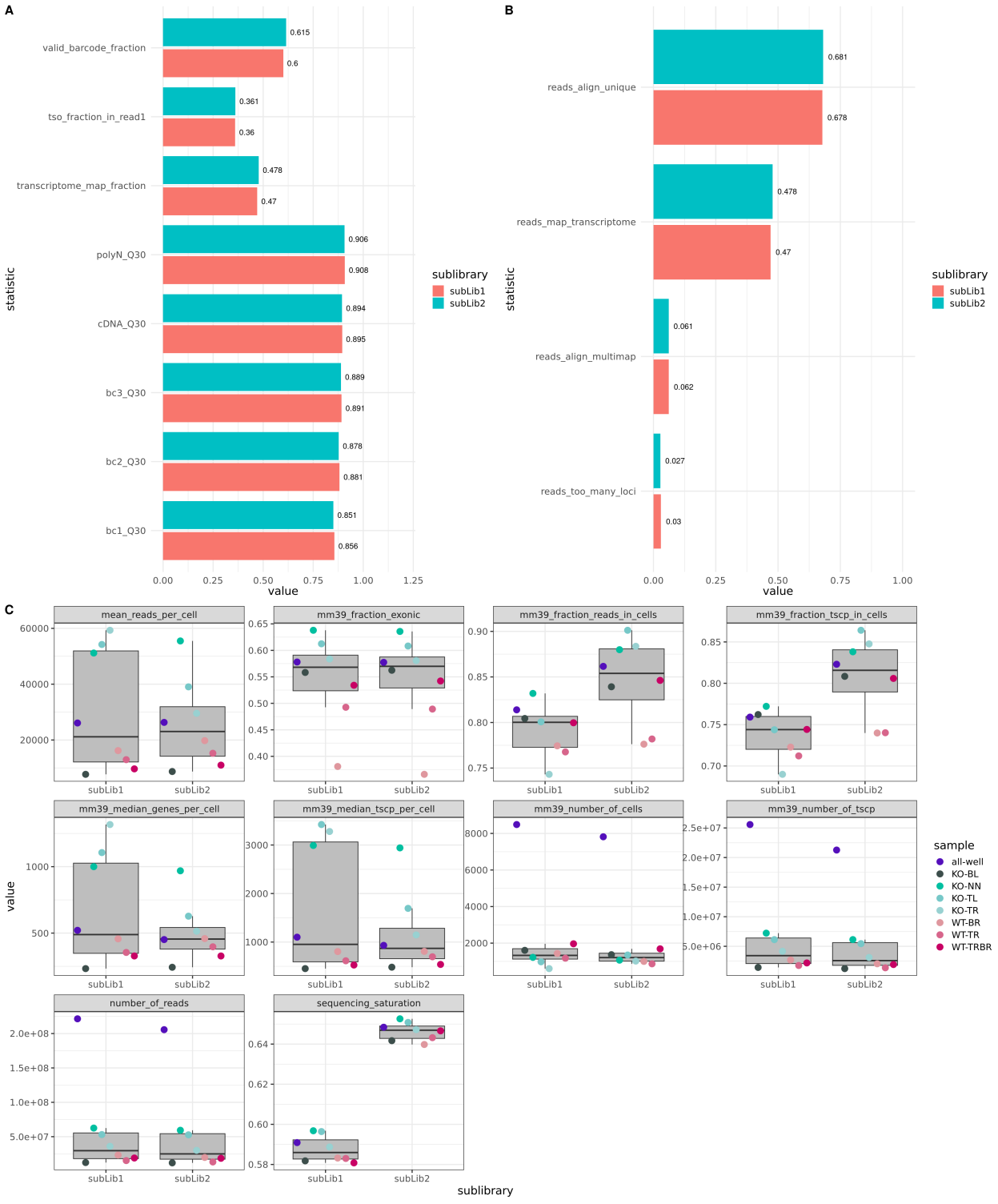

Figure 2: Quality control of raw output from the Parse Biosciences Evercode Whole Transcriptome Mini kit

transcripts and genes detected per cell (`mm39_median_tscp_per_cell`, `mm39_median_genes_per_cell`) and number of unique transcripts detected per sample (`number_of_tscp`).

In general, for both sublibraries, the median genes identified per cell is quite low (usually, we expect around 1-2k genes identified per cell but here it is around 500 per cell). In addition, for the WT samples in particular, the mean reads per cell is also low (we expect at least 20-50k reads per cell on average, but it is less than 20k for the WT samples and KO_BL). Taken together, these features mean that we may not have enough power to properly cluster and separate cells as an insufficient number of transcripts are identified to distinguish between the cell populations.

However, upon examining the sequencing saturation levels for the samples in both sublibraries, we find that they are around 58-65%, which is reasonably high. At this sequencing saturation level, we will need around 2.5 more reads on average for the discovery of an additional unique transcript. Despite the relatively high sequencing saturation levels, considering that the number of detected genes are low for all samples, it will be highly beneficial to resequence the samples (with emphasis on the WT samples) to greater depth if both time and cost permit.

4.5 QC summary

- Phred quality scores appears lower than optimal, which may explain why around 40% of the barcode reads from both sublibraries were discarded. This may also explain the lower proportion of uniquely mapped reads (around 68%).
- There appears to be an imbalance in the sequencing depth for the KO vs WT samples, with the WT samples generally having lower median number of genes and transcripts detected, as well as lower mean number of reads and reads per cell. In addition, the median number of genes detected for all samples is less than 2000, which may result in difficulties in stable clustering and resolving the different cell types between the samples. It may therefore be useful to sequence these samples to greater depth if the downstream analysis does not show clear separation of cell types.
- In conclusion, QC of the parse biosciences output suggest that the *data generated is suboptimal*, and downstream *results should be interpreted with caution.*
- We share the merged seurat R object prior to cell selection at `results_170323/seuratOriginal_16112022.RDS`.

5 Cell selection

As described in Methods, 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

4 quality control criteria: the number of UMI molecules, number of detected genes, percentages of ribosomal and mitochondrial genes, before and after quality control, are shown in Figure 3A-B. As the WT and KO_BL samples are of lower quality than the remaining 3 KO samples (Refer to Figure 2), more cells were filtered away for these 4 samples (Figure 3C). Cells derived from samples across both KO and WT conditions were removed in this process (Figure 3D).

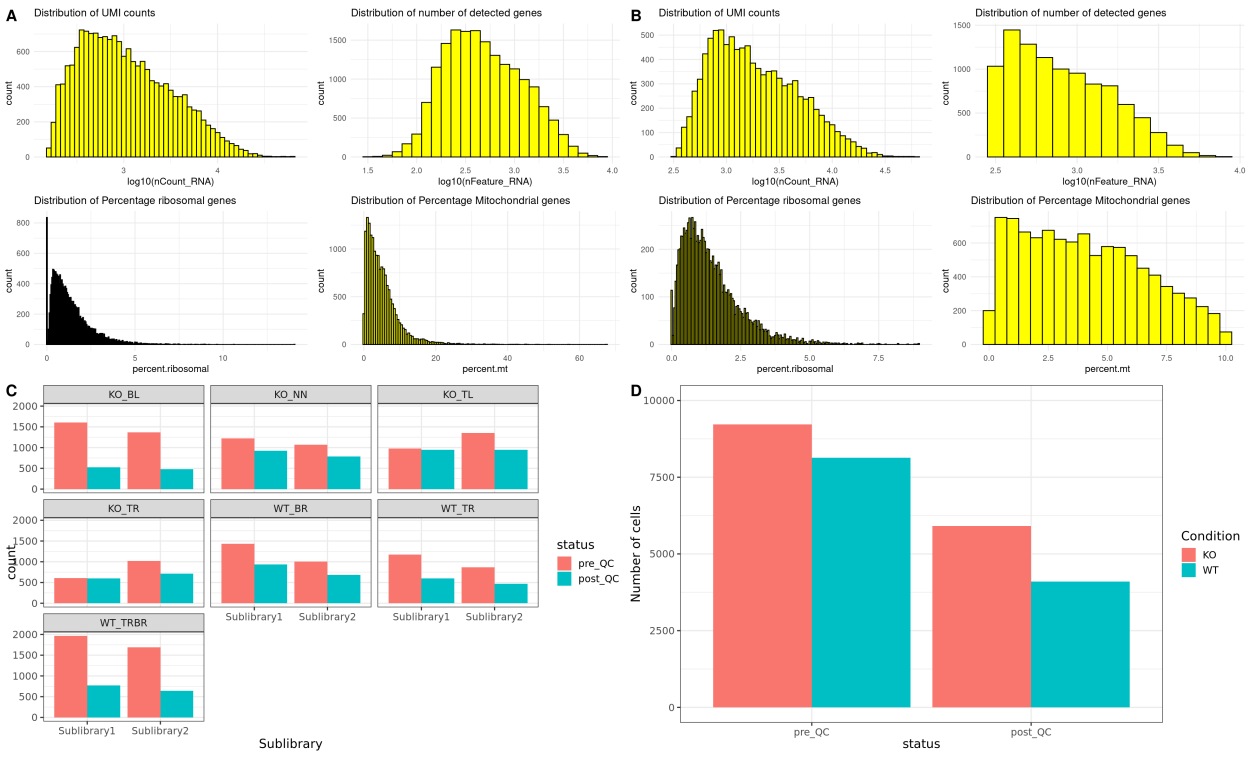

Figure 3: Downstream filtering to remove low quality cells

6 Cluster annotations

The list of positive markers for each of the 11 clusters below, set at an adjusted p value threshold of 0.01, can be found in `results_170322/pos_markers_11_clusters.csv`.

We share the merged seurat R object post-QC and with cluster annotations at

`results_170323/mbladder.unintegrated.labelled.RDS`.

6.1 Urothelial cells (tumorigenic) + CAFs

Potentially tumorigenic urothelial cells undergoing EMT along with CAFs.

Distinguished from the `Umbrella/Basal urothelial cells (normal)` cluster 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

- 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

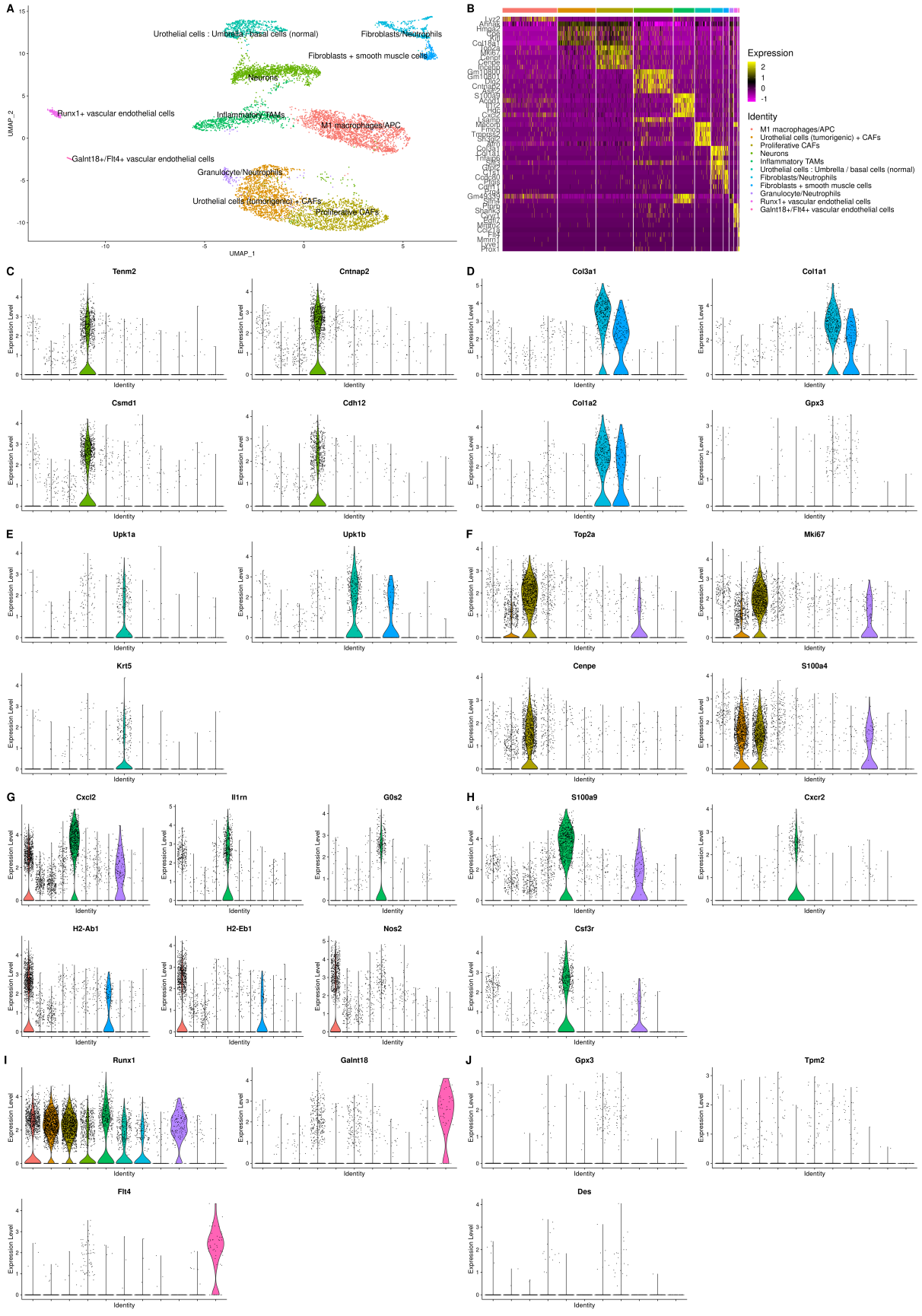

Figure 4: Cluster annotations with selected marker genes

6.2 Urothelial cells : Umbrella / basal cells (normal)

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

6.3 Inflammatory TAMs with CAFs

Evidence:

Distinguished from `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)

6.4 M1 macrophages/APC

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

6.5 Granulocyte/Neutrophils

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

6.6 Neurons

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

6.7 Proliferative CAFs

High expression of mitotic markers (also see cell cycle imputation diagram) e.g. Top2a, MKi67, Cenpf, Cenpe and Incenp. May consist of largely tumor MB49-PSA cells.

Expression of fibroblast gene S100a41 in about 67% of cells.

6.8 Fibroblasts/neutrophils

Expression of Col3a1, Col1a1, Cola2, Gpx3, Clec3b

6.9 Fibroblasts with smooth muscle cells

Expression of des, TPM2, Gpx3

6.10 Runx1+ vascular endothelial cells

6.11 Galnt18+/Flt4+ vascular endothelial cells

7 Differentially expressed genes between KO and WT condition

The list of differentially expressed genes upregulated in the KO vs WT (WT vs KO) condition, can be found in `results_170322/distinct_log2fc_topresults_log2fcsorted_KOup.csv` (`results_170322/distinct_log2fc_topresults_log2fcsorted_WTup.csv`).

Figure 5 captures the output of distinct. Specifically, panels A-C are violin plots showing the log expression levels of several DE genes between the KO and WT conditions in the `Granulocyte/Neutrophils` (A), `M1 macrophages/APC` (B) and `Inflammatory TAMs` (C) clusters. Panel D shows the density plots of 3 genes in the `inflammatory TAMs` cluster. In E left, we show an upset plot on the left which highlights the concordance of DE genes across clusters. For example, 477 genes were DE between KO and WT in the neuronal cluster only, whereas 15 genes are identified as DE in 3 clusters: Umbrella/Basal urothelial cells (normal), proliferative CAFs and neurons clusters. In E right, we display a heatmap showing the expression of the top few DE genes for each cluster (rows) across the mouse samples (columns). Note that whilst the heatmap provides a sample-level view of the scaled expression level of each gene, it does not show display the expression of the gene across clusters for each sample. Hence, it is an incomplete representation as distinct identifies genes that have a different distribution between conditions in each cluster, and not just across samples as a whole.

We also conducted an over-representation analysis of the DE's identified in *distinct* that were enriched in the KO vs WT samples, against the GO, KEGG and MSigDB databases. The list of enriched pathways, as well as dotplots for the top list of enriched pathways can be found in `results_170322/enrichment_analysis`.

8 References

- Seurat: Hao Y, Hao S, Andersen-Nissen E, III WMM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021). "Integrated analysis of multimodal single-cell data." Cell. doi:10.1016/j.cell.2021.04.048, https://doi.org/10.1016/j.cell.2021.04.048.
- Distinct: Simone Tiberi, Helena L Crowell, Lukas M Weber, Pantelis Samartsidis, Mark D Robinson bioRxiv 2020.11.24.394213; doi: https://doi.org/10.1101/2020.11.24.394213

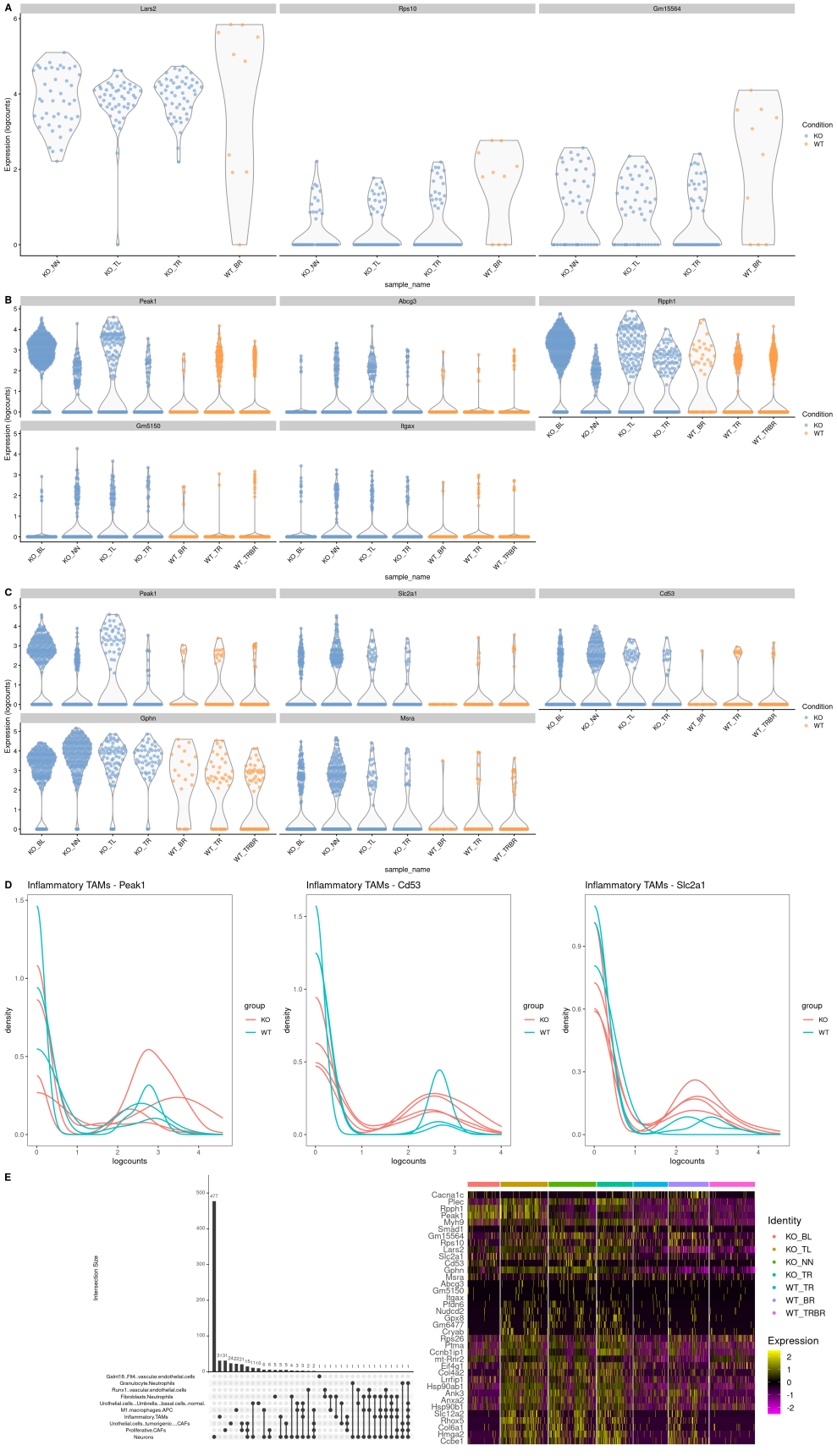

Figure 5: Differential distribution analysis with distinct