Preliminary scRNA-seq analysis of mouse bladder cells

Kane Toh[*]

2022-09-08

Contents

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

1 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.

In our preliminary analysis of the scRNA-seq dataset, we address the following questions:

1. What are the cell types that are present in WT and KO samples?
   - Examine the cluster biomarkers / differentially expressed genes.
   - Automated cell-type annotation
2. Which cell types are amplified or exhausted in the KO vs WT condition (and vice versa)?
   - Detect the differentially abundant clusters between conditions
3. For each of the clusters identified, which genes show significantly different patterns of expression between both conditions?
   - Detect differentially expressed genes between conditions per cluster

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. Sublibraries were first generated separately and then combined into a single dataset using `split-pipe --mode combine` command, passing in the `comb_unfilt_dge True` argument to generate the unfiltered count matrix. Downstream analysis was carried out with Seurat (v4.0.6).

2.2 Quality control and normalisation

Cells were filtered based on the number of detected genes, total number of molecules, percentage of mitochondrial gene expression and the percentage of the highest expressed gene. This resulted in a total of `33,180` cells that were retained for further analysis. Normalisation of the scRNAseq count matrix was implemented using the `sctransform` method, which is based on the regularised negative binomial regression approach from Hafemeister and Satija, (2019). To mitigate 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 Dataset integration, dimensional reduction and clustering

Samples from the WT and GSTT2-KO conditions were integrated in Seurat via the identification of anchors (Stuart et al., 2019). The Uniform Manifold Approximation and Projection (UMAP) dimensional reduction technique was applied to embed cells in lower-dimensional space for ease of visualisation and for cell clustering. 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.

2.4 Identification of differentially expressed genes (cluster biomarkers)

Cluster biomarkers were then identified via the `FindAllMarkers` function, using the Wilcoxon rank sum test for differential gene expression testing.

2.5 Differential abundance (DA) analysis between clusters

DA analysis was conducted via the `EdgeR` (v3.38.4) package to test for significant changes in cluster abundance between WT and KO samples. Differences in cluster abundance between the KO and WT samples were tested with the `glmQLFTest` function.

2.6 Differential expression (DE) analysis between conditions

DE analysis between the WT and KO conditions were performed for each cluster using a pseudobulk approach. In brief, pseudobulk gene expression profiles were constructed by aggregating the counts for all cells with the same combination of sample label and sample condition. DE analysis was conducted in `DESeq2` (v1.36.0) and the "KO" level was contrasted to the "WT" level using the Wald test. The `lfcShrink` function was used to include the shrunken log2 fold changes and standard error output to the results table for further analysis.

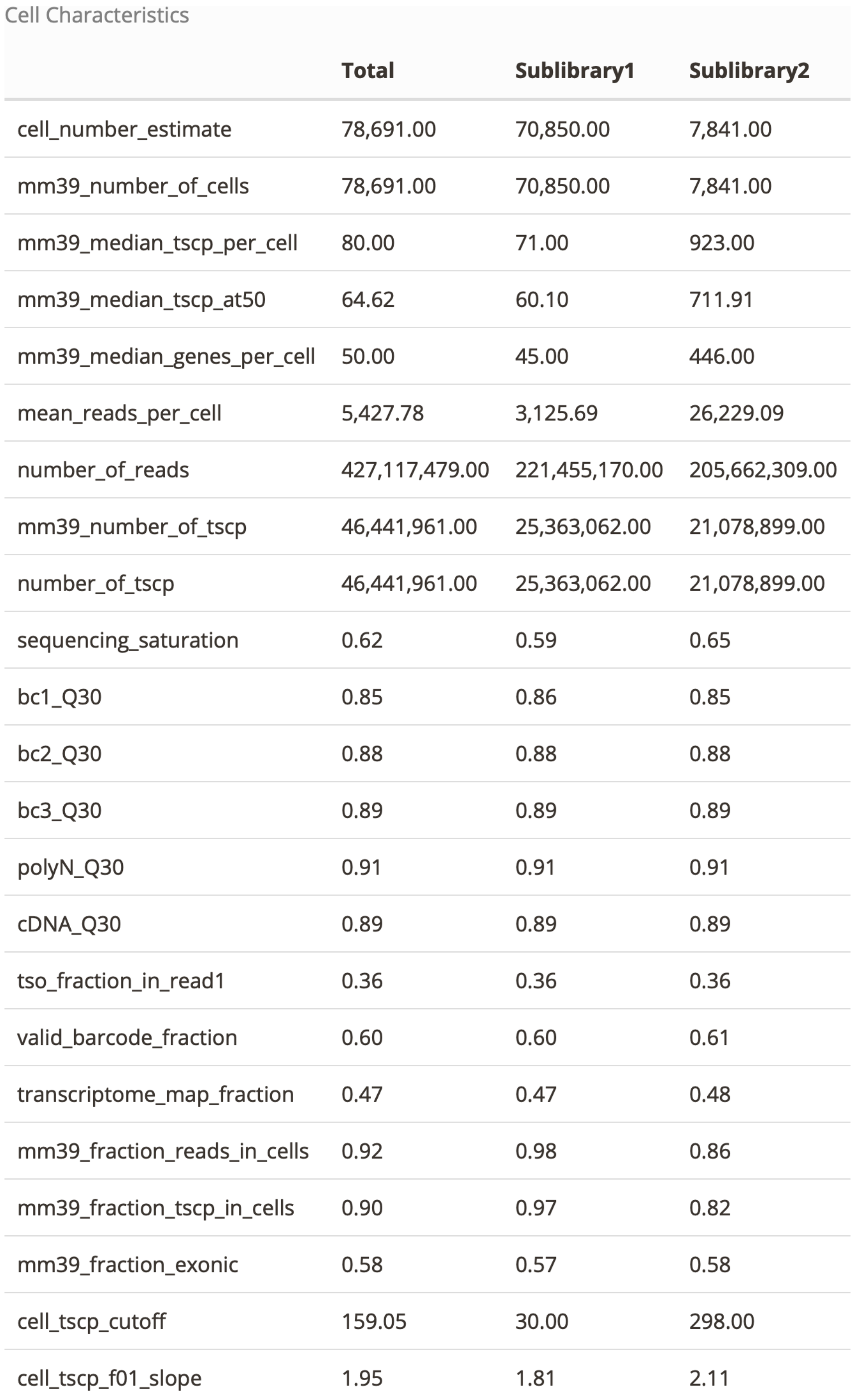
3 Alignment statistics

The summarised result outputs (`html` and `csv`) from the Parse Biosciences alignment can be found in the `results/sequencing` folder.

3.1 Evaluation of sequencing statistics

From the `all-well_analysis_summary.html` html file, we observe that the number of transcripts in each well is quite low, with median transcripts/cell = 80 and median genes/cell = 50 (Figure 1). This is related to the imbalanced number of cells identified in the sequencing sublibraries (Figure 1), with around 70,000 cells identified in sequencing sublibrary 1 vs 7800 cells in sublibrary2 (See also the two `ana_summary.csv` files). As both sublibraries are sequenced to similar depths of around 200 million reads (Figure 2), this suggests that each cell in sublibrary1 is assigned a fewer number of reads. Indeed, this would account for the reduction in sequencing saturation as well as lower median number of transcripts and reads identified per cell in sublibrary1 vs sublibrary2 (Figure 1). Despite this, the sequencing saturation levels for both sublibraries (0.59 and 0.65) are reasonably high (>0.5), implying that we have hit a point of diminishing returns and additional sequencing will be less productive in terms of detecting additional transcripts and genes per cell. Therefore, it appears that the sequencing depths for both sublibraries are appropriate to capture the transcriptomic complexity of the sample cell populations.

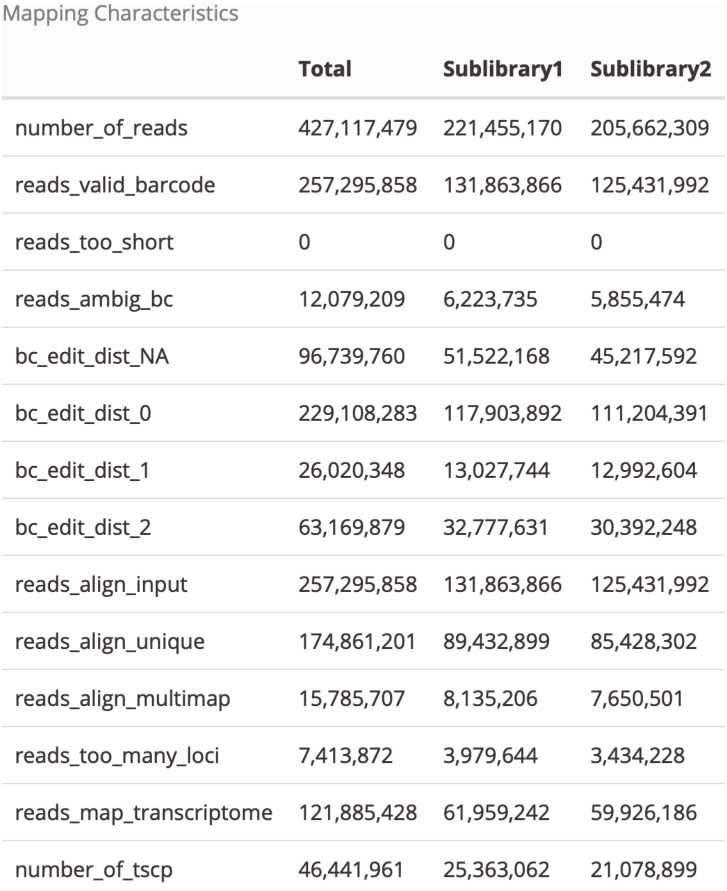
In addition, we observe that the barcode-rank plot lacks a distinct "drop" in the curve. This can occur when there is an excess number of dead/damaged cells in the data set, creating an intermediate "smoothness" on the curve where a drop would be expected. It is likely therefore that the number of cells estimated with the default parameters in the Parse Biosciences pipeline has led to an overestimation of the actual cell number. Therefore, as recommended by the Parse Biosciences team, we reran the pipeline in combine mode to produce the unfiltered matrix, and then read in the unfiltered (DGE_unfiltered) matrix into Seurat to enable manual adjustment of the cell cutoff.



Cell Characteristics

	Total	Sublibrary1	Sublibrary2
cell_number_estimate	78,691.00	70,850.00	7,841.00
mm39_number_of_cells	78,691.00	70,850.00	7,841.00
mm39_median_tscp_per_cell	80.00	71.00	923.00
mm39_median_tscp_at50	64.62	60.10	711.91
mm39_median_genes_per_cell	50.00	45.00	446.00
mean_reads_per_cell	5,427.78	3,125.69	26,229.09
number_of_reads	427,117,479.00	221,455,170.00	205,662,309.00
mm39_number_of_tscp	46,441,961.00	25,363,062.00	21,078,899.00
number_of_tscp	46,441,961.00	25,363,062.00	21,078,899.00
sequencing_saturation	0.62	0.59	0.65
bc1_Q30	0.85	0.86	0.85
bc2_Q30	0.88	0.88	0.88
bc3_Q30	0.89	0.89	0.89
polyN_Q30	0.91	0.91	0.91
cDNA_Q30	0.89	0.89	0.89
tso_fraction_in_read1	0.36	0.36	0.36
valid_barcode_fraction	0.60	0.60	0.61
transcriptome_map_fraction	0.47	0.47	0.48
mm39_fraction_reads_in_cells	0.92	0.98	0.86
mm39_fraction_tscp_in_cells	0.90	0.97	0.82
mm39_fraction_exonic	0.58	0.57	0.58
cell_tscp_cutoff	159.05	30.00	298.00
cell_tscp_f01_slope	1.95	1.81	2.11

Figure 1: All well stats.

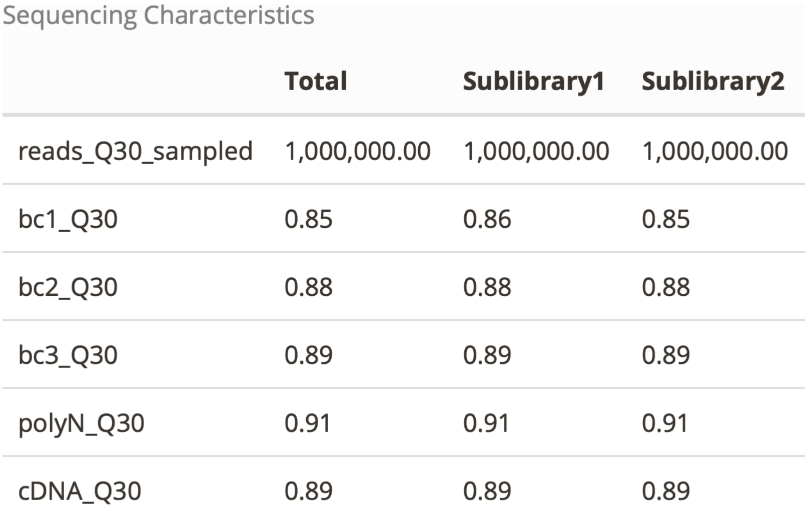


Mapping Characteristics

	Total	Sublibrary1	Sublibrary2
number_of_reads	427,117,479	221,455,170	205,662,309
reads_valid_barcode	257,295,858	131,863,866	125,431,992
reads_too_short	0	0	0
reads_ambig_bc	12,079,209	6,223,735	5,855,474
bc_edit_dist_NA	96,739,760	51,522,168	45,217,592
bc_edit_dist_0	229,108,283	117,903,892	111,204,391
bc_edit_dist_1	26,020,348	13,027,744	12,992,604
bc_edit_dist_2	63,169,879	32,777,631	30,392,248
reads_align_input	257,295,858	131,863,866	125,431,992
reads_align_unique	174,861,201	89,432,899	85,428,302
reads_align_multimap	15,785,707	8,135,206	7,650,501
reads_too_many_loci	7,413,872	3,979,644	3,434,228
reads_map_transcriptome	121,885,428	61,959,242	59,926,186
number_of_tscp	46,441,961	25,363,062	21,078,899

Figure 2: mapping stats.

3.2 Sequencing quality of sublibraries



Sequencing Characteristics

	Total	Sublibrary1	Sublibrary2
reads_Q30_sampled	1,000,000.00	1,000,000.00	1,000,000.00
bc1_Q30	0.85	0.86	0.85
bc2_Q30	0.88	0.88	0.88
bc3_Q30	0.89	0.89	0.89
polyN_Q30	0.91	0.91	0.91
cDNA_Q30	0.89	0.89	0.89

Figure 3: sequencing stats.

From Figure 3, the %Q30 score, which represents the percentage of bases with a quality score of 30 or higher, is high (above 80%). This reflects that the sequencing quality of the libraries is high.

4 Quality Control

We considered the distributions of 4 different metrics during cell filtering (Figure 4):

1. nCount_RNA: the total number of reads (specifically, UMIs) in the dataset
2. nFeature_RNA: the total number of observed genes
3. percent.mt: percentage of mitochondrial gene expression
4. percent.Largest.Gene: percentage of the data comes from the single most observed gene

Currently, we retain cells that pass the following thresholds (Figure 5) :

- nFeature_RNA > 10 and < 1000
- nCount_RNA > 100 and < 3000
- percent.mt < 15
- percent.Largest.Gene < 50

This retains a total of `33,180` cells from the initial pool of 192,758 cells, or equivalently, around `17%` of the cells from the initial unfiltered output were retained for further analysis. Amongst these cells, `22,227` cells are derived from KO samples and the remaining `10,953` from WT samples.

Optimisation note: Our selection of cells is rather permissive as this still retains many cells with a low number of features and UMI counts. The stringency of cell filtering can be tuned iteratively, on the basis of the downstream results.

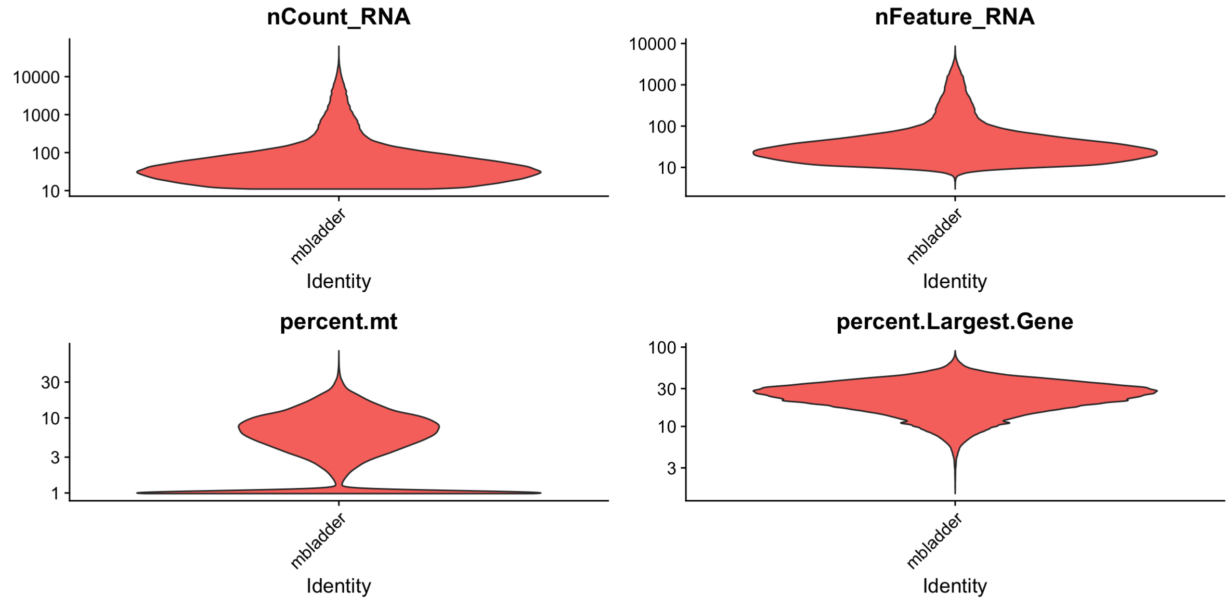


Figure 4: Violin plots of 4 metrics used for quality control

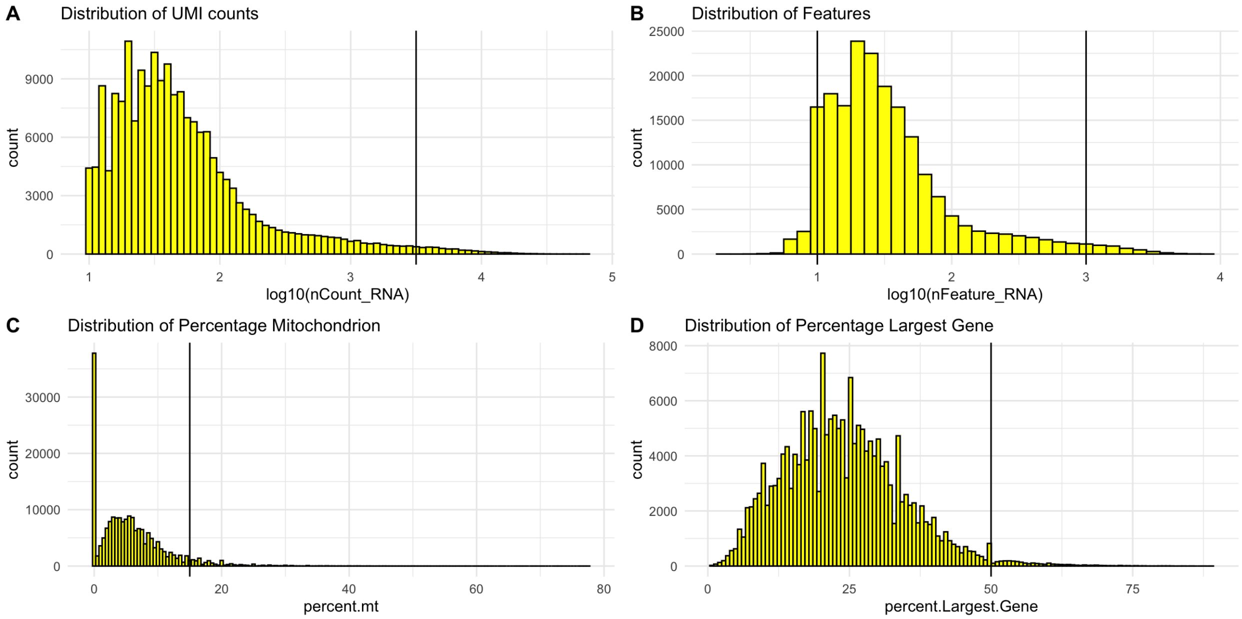


Figure 5: Thresholds used in the 4 chosen metrics for filtering cells

5 Preliminary analysis of seurat object

As our investigation is focused on the differences in transcriptomic profiles between samples in the WT and GSTT2-KO conditions, we first assessed whether there are technical and biological sources of variation that can potentially confound our analysis. The PCA plots suggest that variation in the cell-cycle is unlikely to be an interesting source of variation in the dataset as cells from all 3 phases (G1, G2/M and S) have a high overlap with one another (Figure 6A). Similar considerations lead us to conclude that there is no detectable batch-effect arising from sublibrary preparation or sample assignment (BL/BR/NN etc.) (Figure 6B,C). In contrast, cells can be separated in PCA space, along PC2 in particular, by their KO and WT status (Figure 6D). This therefore motivated us to consider integrating the KO and WT datasets together so that shared cell populations/clusters can be identified and compared between these conditions.

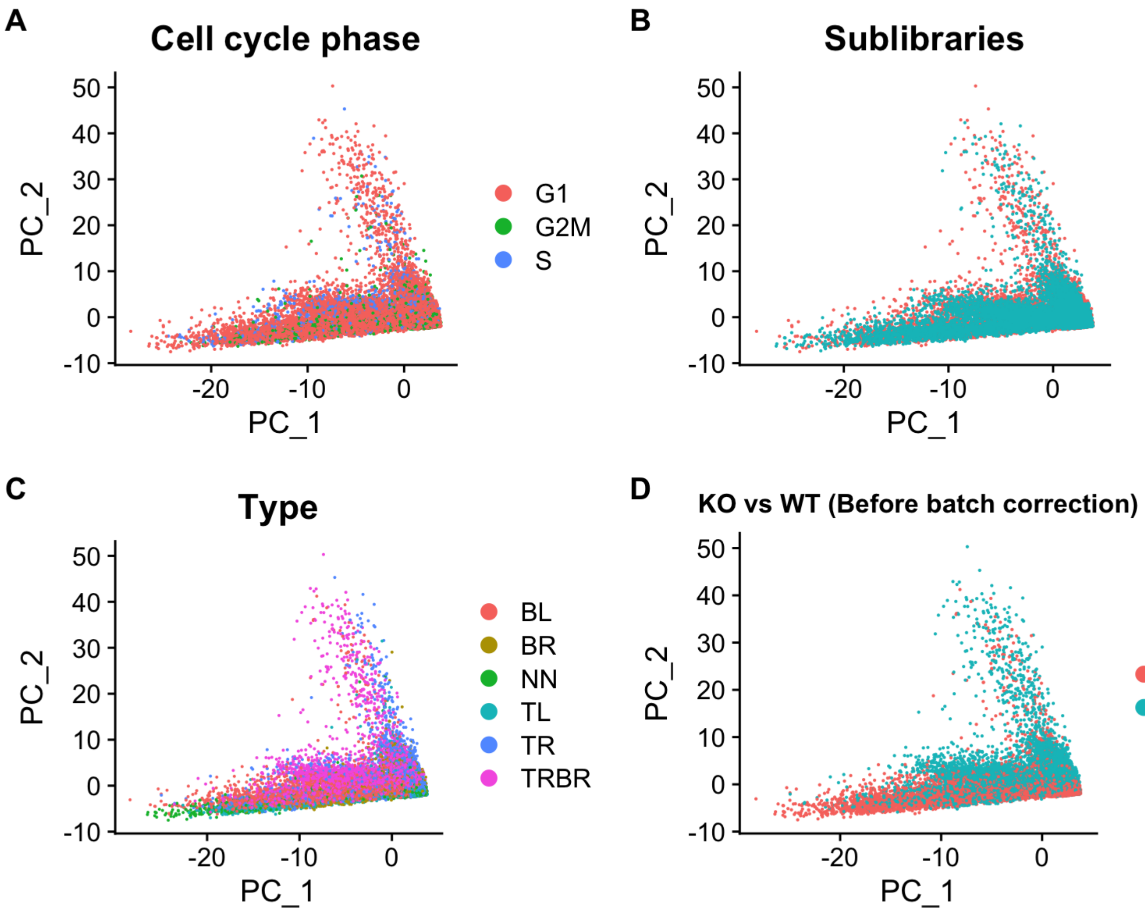


Figure 6: Contribution of the cell-cycle, sublibrary number, sample assignment and sample condition (KO vs WT) on cell-cell variation.

6 Integration of KO and WT samples

Integration of the data by the samples' KO and WT condition results in an improved overlap of cells in the PCA dimensions (Figure 7B) when compared against the non-integrated dataset (Figure 7A), visually highlighting the impact and utility of the integration.

Next, UMAP was used to embed cells of the integrated dataset in reduced dimensions, which is then followed by louvain clustering of the dataset. From this, we observe that KO and WT cells are distributed between all the resulting 13 clusters (Figure 8).

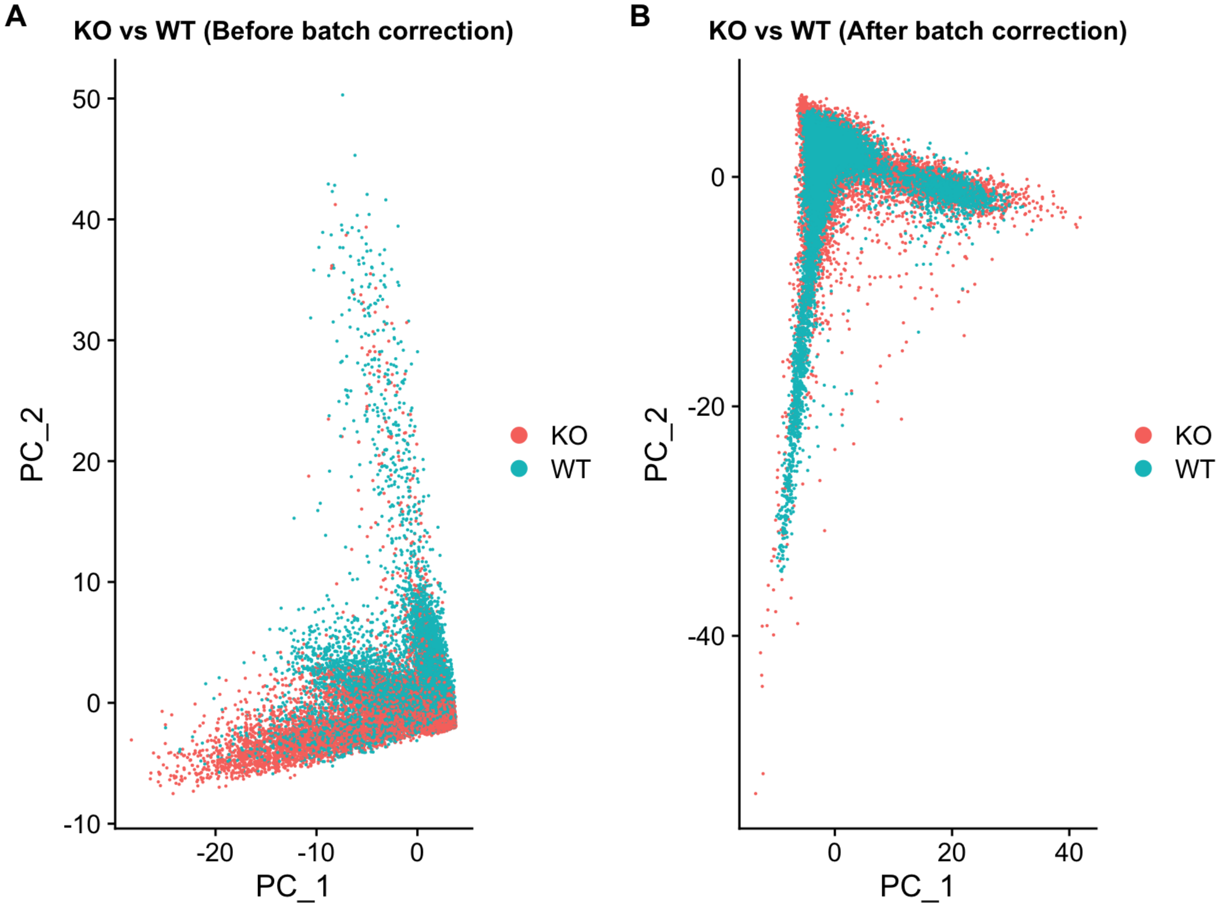


Figure 7: PCA plot showing the cell embedding before and after data integration

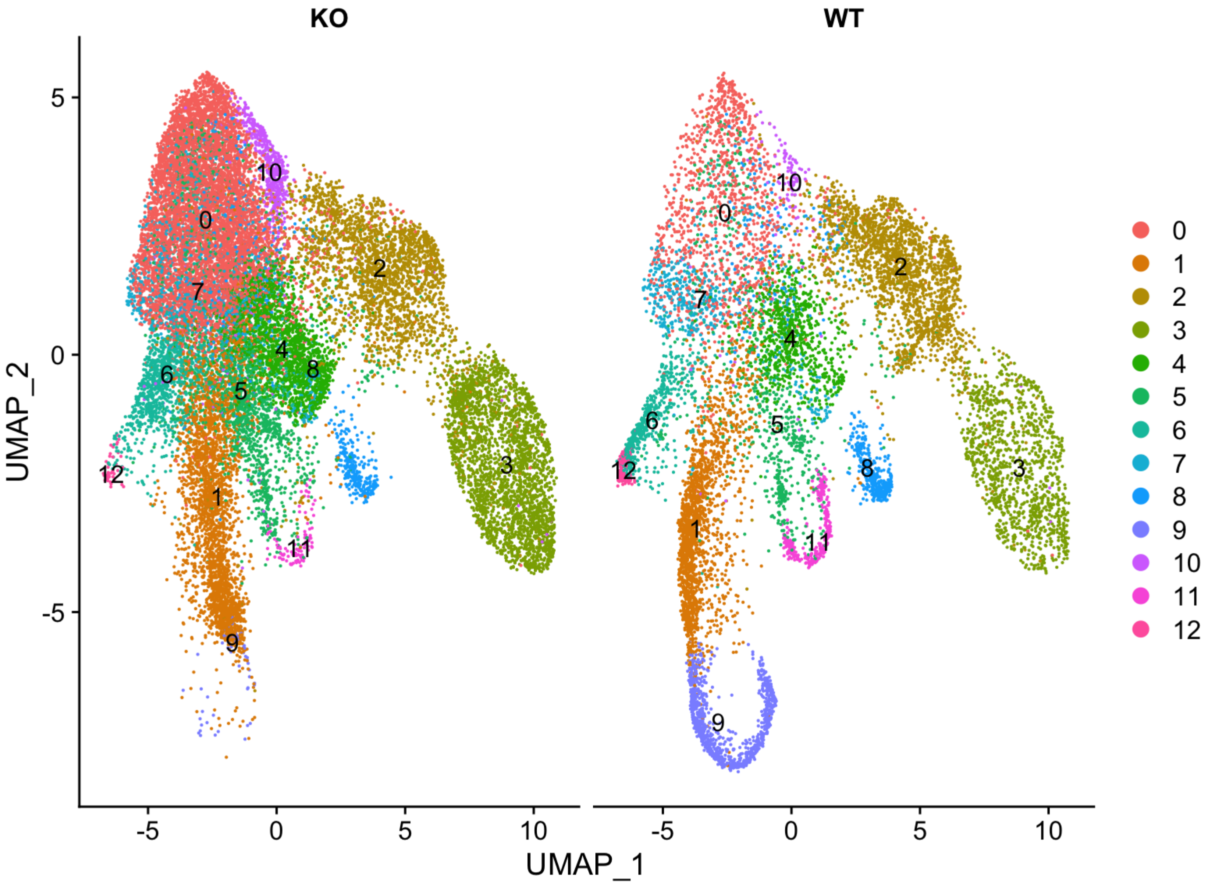


Figure 8: UMAP plots of cells derived from the KO and WT conditions, highlighting the clusters identified by Seurat

7 Examine cluster biomarkers

Marker genes for each cluster are saved in `results/clusterDEGs/all_markers.csv`, and positive markers for each cluster are saved in `results/clusterDEGs/all_markers_pos_only.csv`. Expression of the top 5 markers for each cluster was displayed in the form of a dotplot (Figure 9).

Optimisation note: Clustering parameters can be modified to increase/decrease the clustering resolution. We can investigate particular clusters further and subcluster cells within that cluster. The identified DEGs will change based on these tweaks.

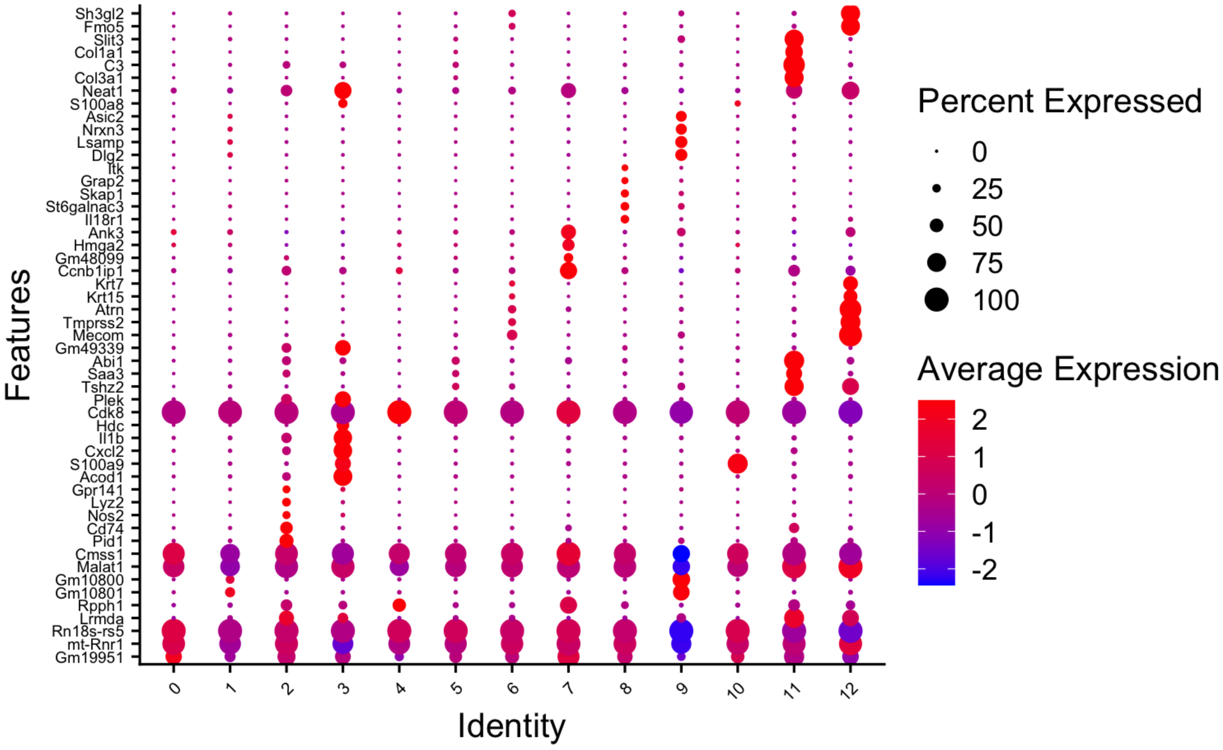


Figure 9: Dotplot showing marker gene expression across the 13 clusters

7.1 Interpretation of DEG csv files

We provide 2 csv files containing information on the cluster biomarkers.

The `all_markers.csv` file, as the name suggests, consists of all markers identified for each cluster that has an adjusted p value < 0.05. Aside from the gene name, the 2 other columns that are important are the `avg_log2FC` and `p_val_adj` columns.

- The `avg_log2FC` represents the log2 fold-change of the average expression of the gene between 2 groups. In this case, the first group is the cluster of interest (cluster 0 in the table below) and the second group corresponds to all the remaining clusters. For example, we see that the `Gm19951` lncRNA transcript is, on average, expressed at 2.2237583 times higher in cluster 0 relative to all other clusters. In contrast, the `Rpph1` gene is expressed at 3.6477327 times lower in cluster 0 relative to all other clusters. For the `all_markers_pos_only.csv` file, all the `avg_log2FC` values are positive, which reflects the fact that the list consists of only positively expressed marker genes.

- The `p_val_adj` value represents the adjusted p-value, based on using bonferroni correction on the total number of genes in the dataset. The table is sorted in increasing levels of the adjusted p-value, so genes at the top of the list of each cluster have a higher degree of confidence. Nevertheless, in general, we should be cautious in interpreting these values due to selection-bias: genes used for clustering are the same genes tested for differential expression (Zhang et al., 2019).

In addition, the `pct.1` value is the percentage of cells in the cluster where the gene is detected while `pct.2` is the percentage of cells on average in all the other remaining clusters where the gene is detected. Ideally, a positive marker would be expressed exclusively in the cluster of interest (pct.1 = 1) and completely silenced in all other clusters (pct.2 = 0).

```
##                   p_val avg_log2FC pct.1 pct.2     p_val_adj cluster      gene
## Gm19951    0.000000e+00  1.1534442 0.630 0.534  0.000000e+00       0   Gm19951
## mt-Rnr1    0.000000e+00  0.5419385 0.935 0.907  0.000000e+00       0   mt-Rnr1
## Rn18s-rs5  0.000000e+00  0.3957353 1.000 1.000  0.000000e+00       0 Rn18s-rs5
## Lrmda      0.000000e+00 -1.5806799 0.043 0.242  0.000000e+00       0     Lrmda
## Rpph1      0.000000e+00 -1.8672749 0.056 0.275  0.000000e+00       0     Rpph1
## Fth1      1.885707e-307 -1.3293280 0.089 0.312 1.070101e-302       0      Fth1
```

```
##               p_val avg_log2FC pct.1 pct.2    p_val_adj cluster   gene
## Sf3b12 7.142052e-09  0.2618206 0.187 0.074 0.0004052972      12  Sf3b1
## Gab11  1.367921e-08  0.3163192 0.121 0.040 0.0007762681      12   Gab1
## Cyb5r3 8.849245e-08  0.2827261 0.101 0.032 0.0050217697      12 Cyb5r3
## Fam13b 1.486881e-07  0.2770553 0.111 0.038 0.0084377542      12 Fam13b
## Pak2   2.369129e-07  0.2778288 0.116 0.041 0.0134443333      12   Pak2
## Herc1  5.135286e-07  0.2809353 0.106 0.037 0.0291417226      12  Herc1
```

7.2 Automated cell-type annotation

Automated, reference-based cell-type annotation was carried out with scmap (v1.18.0), using the 3-month old bladder samples from the Tabula Muris Senis project. Both the Smart-seq2 and 10x datasets were obtained and then integrated in Seurat, and then used as the scmap reference for `scmap-cluster`. Whilst the majority of cell were not annotated, we hope this would provide the investigators with some assistance in the manual annotation of the clusters.

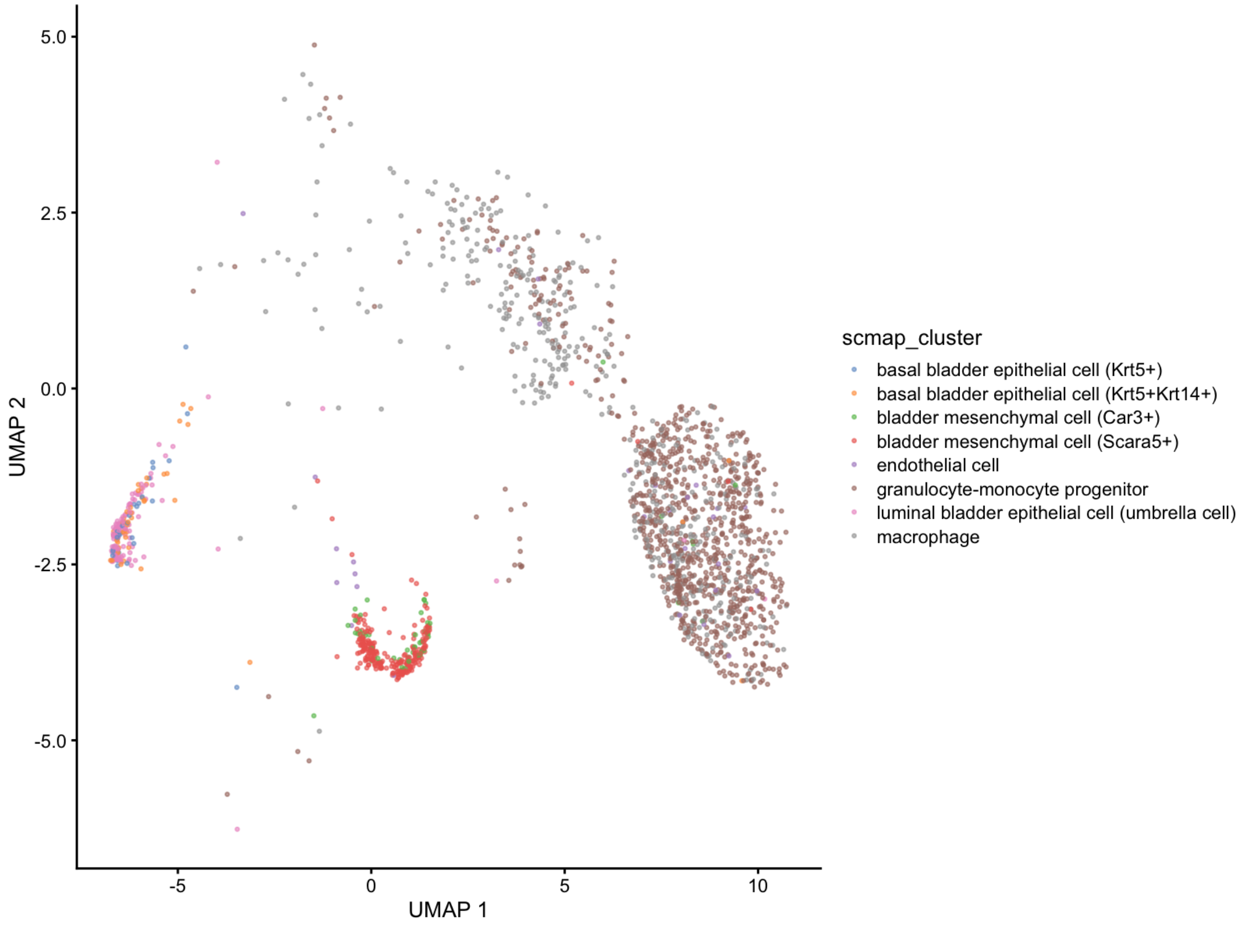


Figure 10: Automated annotation of cells

7.3 Identifying differentially abundant clusters

From the DA analysis, we identify cluster 9 as the only cluster (at an FDR < 0.05) that is strongly enriched in the WT condition relative to the KO condition.

```
## # A tibble: 10 x 6
##    `Seurat Cluster numbers`  logFC logCPM      F   PValue     FDR
##                       <dbl>  <dbl>  <dbl>  <dbl>    <dbl>   <dbl>
##  1                        9  4.49     15.2 29.1   0.000628 0.00816
##  2                       12  3.17     13.1  3.18  0.113    0.498
##  3                       10 -1.38     14.1  2.61  0.144    0.498
##  4                        8  1.04     15.2  2.24  0.173    0.498
##  5                       11  2.24     14.0  2.03  0.193    0.498
##  6                        0 -1.29     17.6  1.69  0.230    0.498
##  7                        3 -0.414    16.9  0.288 0.606    0.825
##  8                        2  0.517    17.3  0.237 0.639    0.825
##  9                        5 -0.359    16.2  0.218 0.653    0.825
## 10                        1  0.279    17.2  0.214 0.656    0.825
```

7.4 Identifying differentially expressed genes between KO and WT

We examined the DEGs between KO and WT conditions for each cluster. With the exception of clusters 11 and 12, we found DEGs for all clusters and have included the results in `results/DEConditions/cluster{0-12}/`. For instance, for cluster 9, we observe that samples are well-separated based on their KO/WT conditions along the first PC, implying that we should expect to find DEGs that drive the difference between samples (Figure 11). In Figure 12, we observe that genes encoding heat shock proteins are enriched in the KO samples, perhaps indicating that these cells are highly stressed in the KO samples.

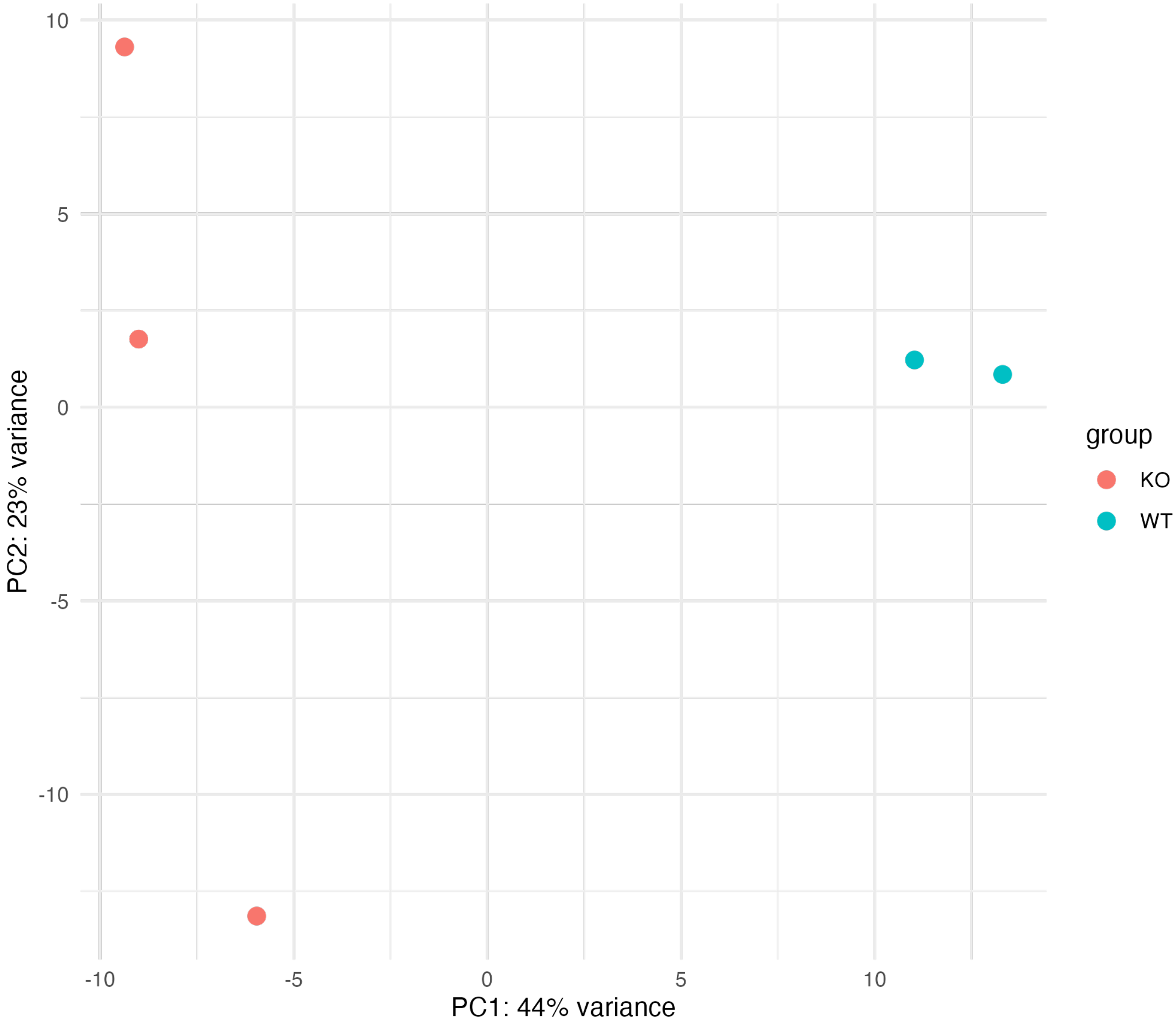


Figure 11: PCA plot showing separation of samples within cluster 9 based on sample condition

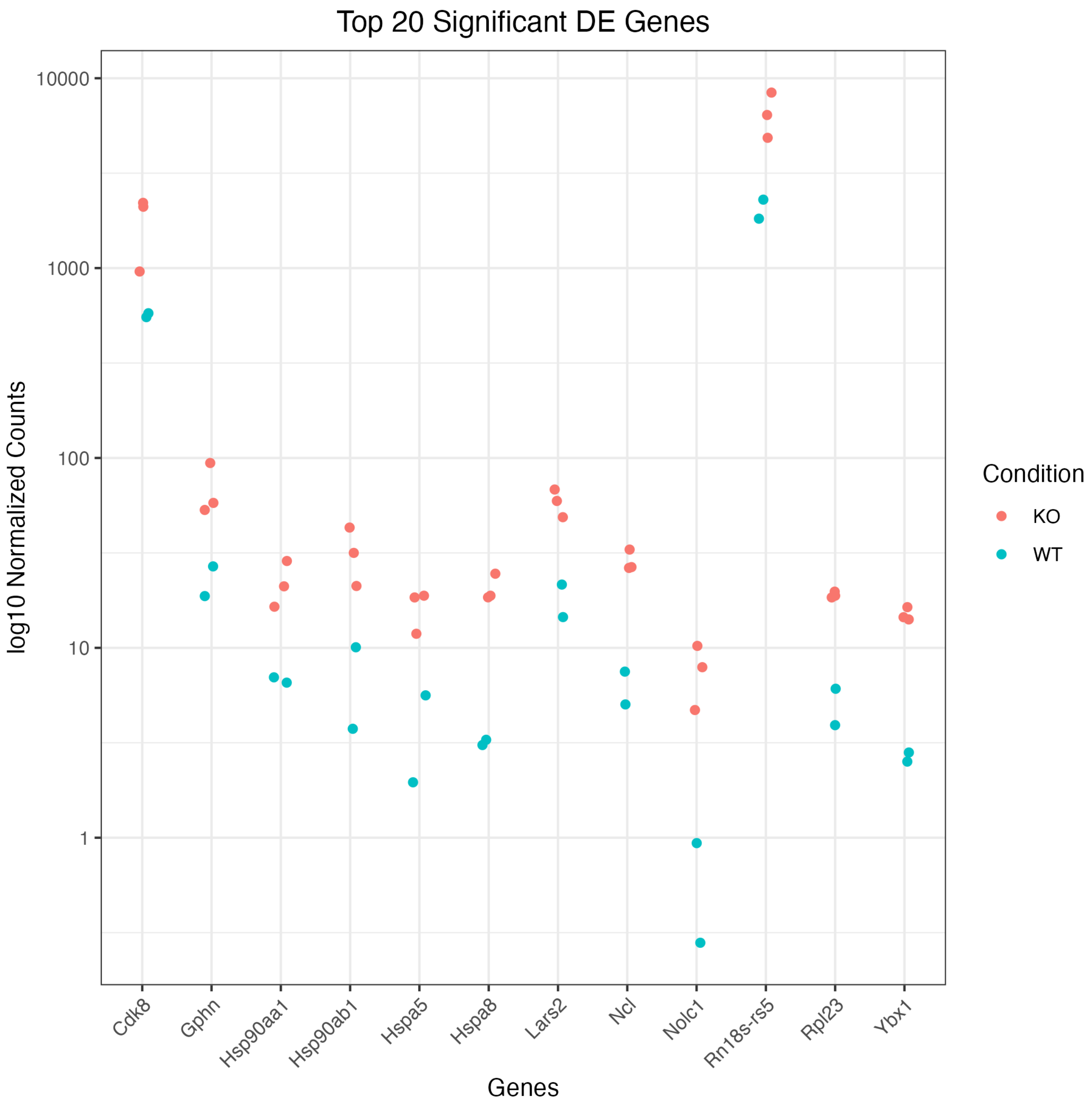


Figure 12: Top 20 DEGs that separate samples in cluster 9, based on their KO and WT status

8 Proposed next steps

8.1 For investigators

- Assess sensibility of results, based on identified cluster biomarkers (`results/clusterDEGs/*.csv`) as discussed in the Interpret DEG csv files section. Provide suggestions if clusters should be split or merged. Investigators can reference the automated annotations in (Figure 10) if helpful.

8.2 Computational analyses

- Optimise upstream steps (alignment, cell-filtering, clustering parameters) based on feedback from investigators
- Which cells are most likely to be tumor cells?
  - Manual marker gene annotations by investigators
  - Inferring CNV status, and identifying tumor cells as cells with CNV
- How do gene expression profiles change for cells in various clusters, based on their KO or WT status?
  - Pseudotime analysis with RNA velocity
  - Clustering trajectories based on expression patterns
- Which pathways are enriched/depleted in the KO vs WT conditions?
  - GSEA analysis / functional enrichment of genes