01_preliminary_analysis_parse

Kane Toh*

2022-10-21

Contents

1 Background **1**

2 Quality control and normalisation **2**

3 Batch effect assessment and sample integration **5**

3.1 Effects of integration 5

4 Differential gene expression analysis **5**

5 Summary **5**

6 Next steps **11**

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 uncover the cell types that can be discovered in the WT and KO scRNA-seq samples. Our analysis follows these 3 steps:

1. Performing QC on the filtered sample matrices individually.
2. Correcting for batch effects in the scRNAseq dataset
3. Extracting the list of cluster biomarkers / differentially expressed genes.

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

2 Quality control and normalisation

For each of the 14 samples derived from the 2 sublibraries, we loaded the corresponding filtered matrix into a Seurat object and examined the distribution of UMI counts, number of detected genes, percentage of ribosomal and mitochondrial genes.

From our initial QC analysis (`00_qc.pdf`, Figure 4), we noted that the median number of reads and detected transcripts sequenced from the *KO_NN* library is greater than the *KO_BL* samples. This pattern is reflected in Figure 1 and Figure 2 in panels A and B. Comparing panels A, we find that most cells from the *KO_BL* library have around 316 to 1780 UMI counts whereas many cells from the *KO_NN* library have UMI counts that are 3160 and greater. Similarly, by comparing panels B, we observe that the *KO_BL* distribution is left-shifted compared to the *KO_NN* distribution.

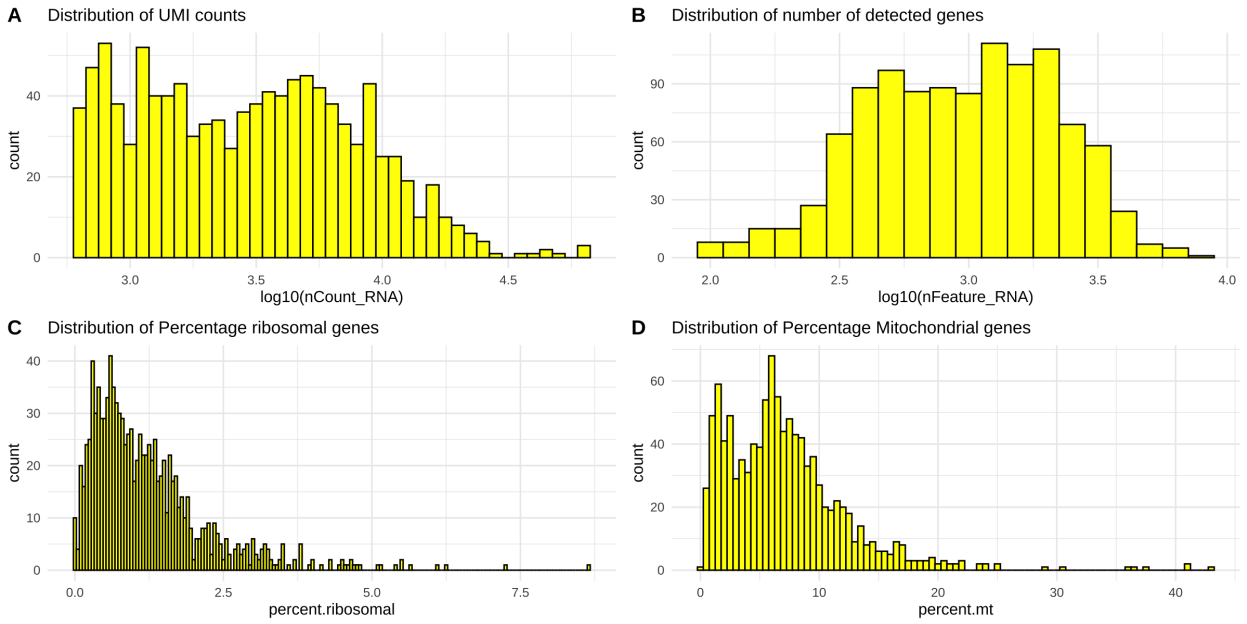

Figure 1: QC metrics for sublibrary2 KO-NN sample

From examining the QC distributions across all samples, we elected to retain cells that pass the following 3 QC thresholds:

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

The result of the QC step led to a reduction in the number of cells retained for downstream analysis in both sublibraries. We find that our selected thresholds remove cells from both sublibraries (Figure 3) and from the KO and WT conditions (Figure 4) without dramatic bias. In addition, the procedure removes more cells from the KO_BL and WT_TRBR samples than others (Figure 3), simply because these samples have fewer median number of reads and detected genes per cell. The result of the above thresholding on the distribution of various QC metrics is displayed in (Figure 5).

We now have a combined (all 14 samples across 2 sublibraries) seurat object of 56748 features and 9651 cells for downstream analysis.

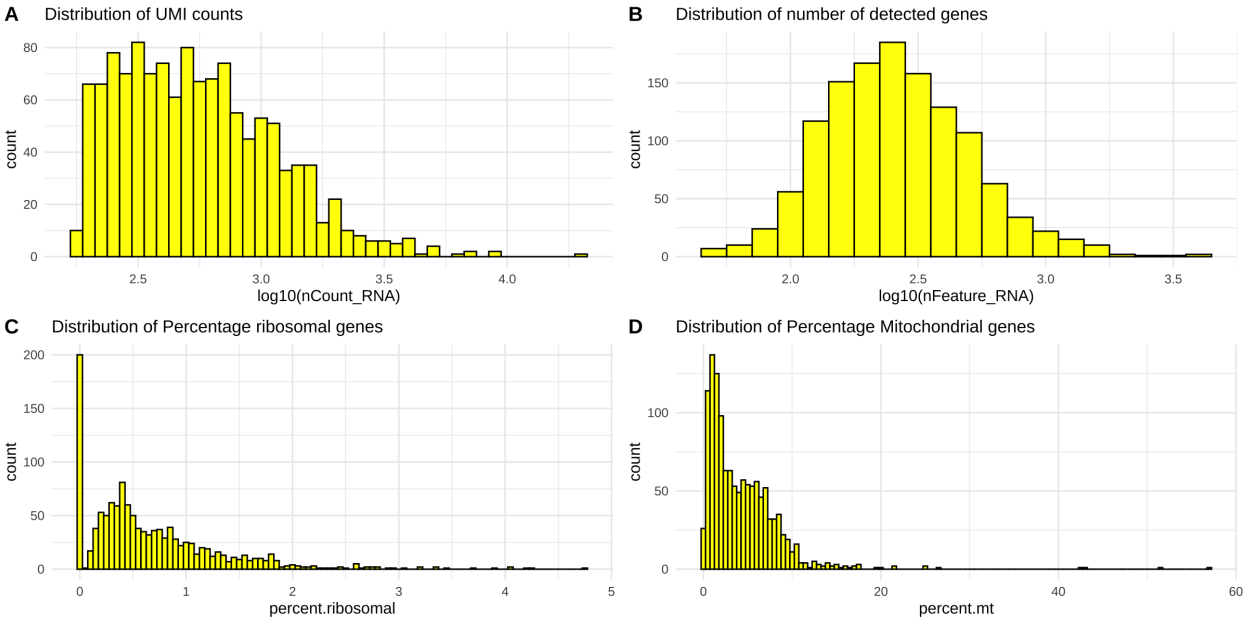

Figure 2: QC metrics for sublibrary2 KO-BL sample

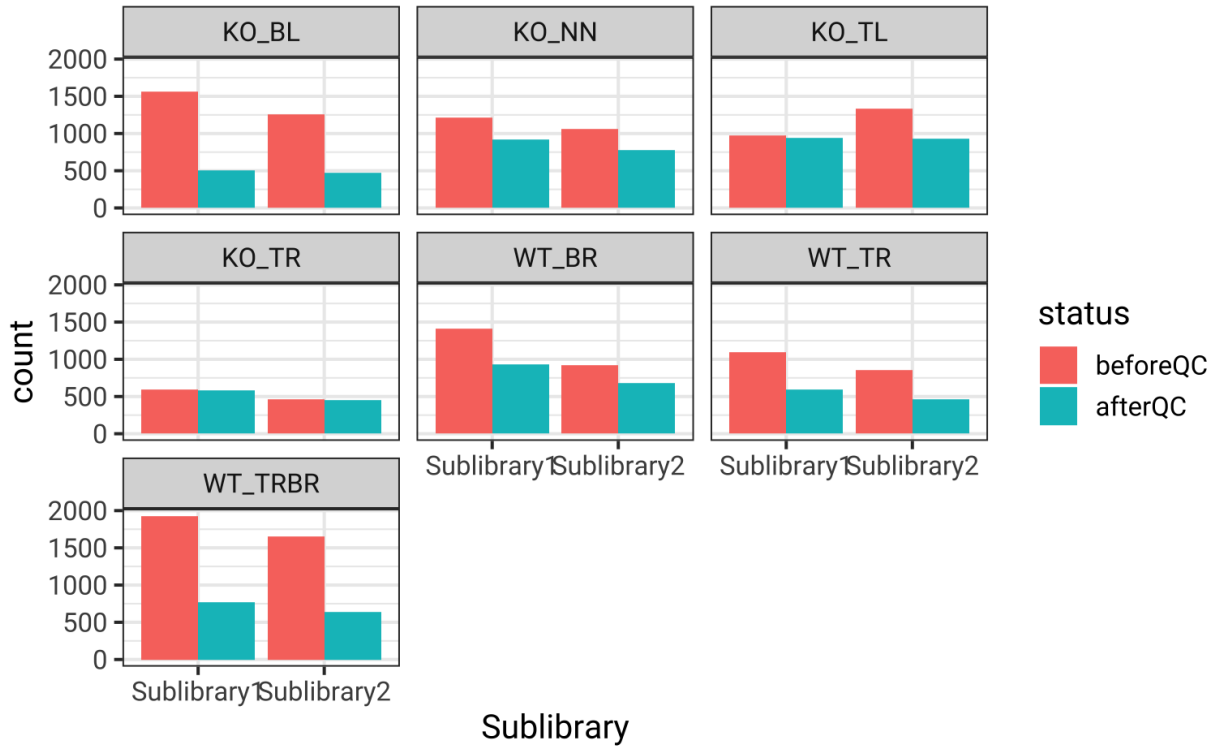

Figure 3: Number of cells before and after QC across both sublibraries

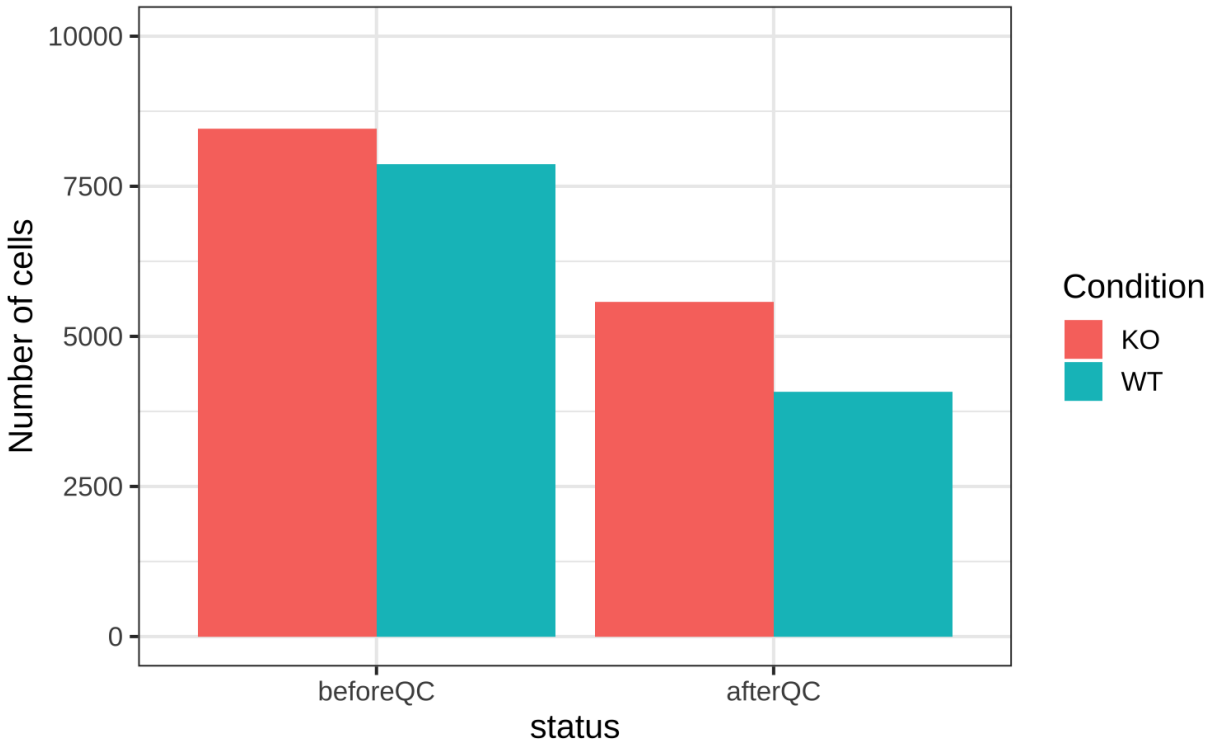

Figure 4: Number of cells before and after QC across both KO and WT conditions

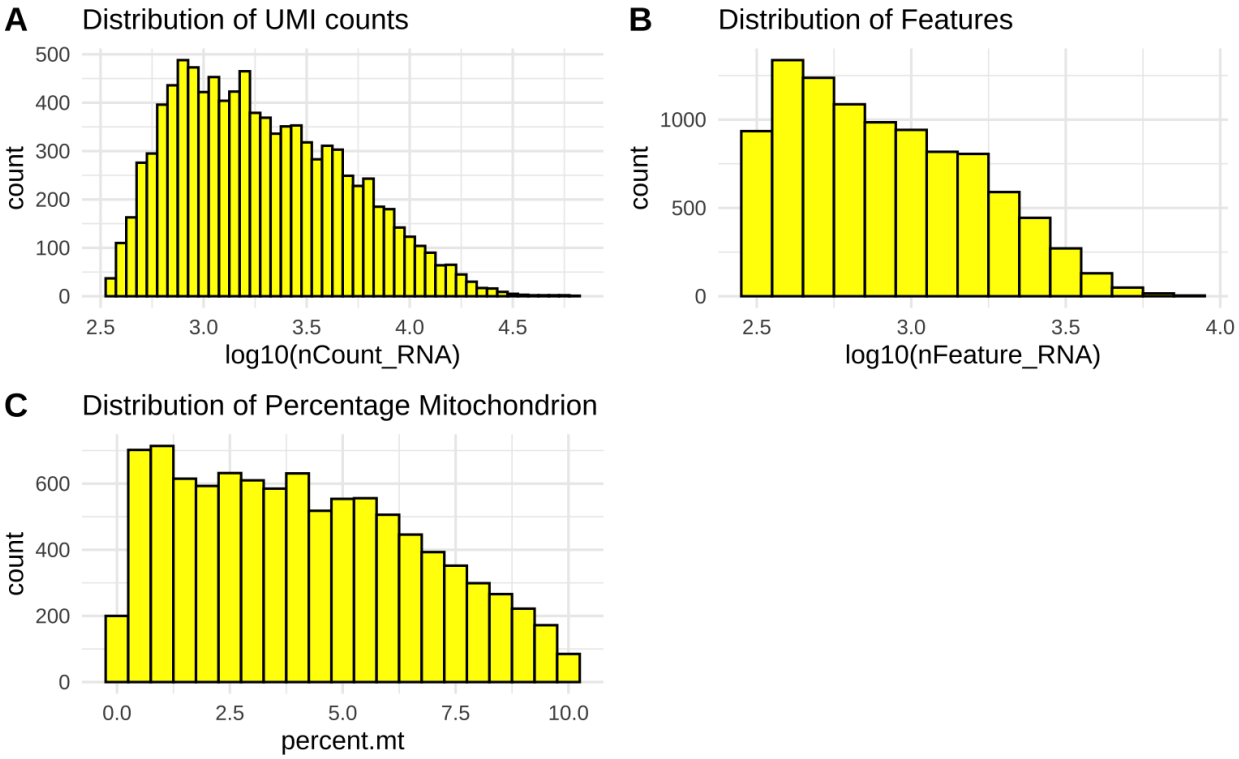

Figure 5: Post-QC metrics for all samples

3 Batch effect assessment and sample integration

First, we assessed whether substantial batch effects exist between the samples by inspecting whether cells cluster by biological or technical variables that are not of interest. From Figure 6, we observe in panels B and D that cells do not cluster by sublibrary or noticeably by their KO/WT status. In contrast, G2M cells cluster together in the lower quadrant of panel A and sample types appear unevenly distributed across the reduced UMAP dimensions in panel C. We note that as cell cycle status may be a relevant feature, at this stage of our analysis, we do not regress out cell-cycle effects.

3.1 Effects of integration

When we perform cell clustering with the louvain community detection algorithm and then visualise the clustering in reduced dimensions (UMAP space) without integrating the samples, we observe in the upper panel of Figure 7 that the distribution of the sample types across the 13 clusters appear to be quite variable. For instance, none of the cells from both *BL* and *TRBR* samples are found in the bottom left set of clusters (clusters 1, 2, 3 and 10). In addition, we find that cells from both WT and KO samples are found across all clusters, as shown in (Figure 6 D). This indicates that there is no strong batch effect arising from the KO or WT conditions (e.g. if cells form clusters dominated by KO or WT samples). Finally, we notice that cluster 3 appears to be enriched in cells predicted to be in the G2/M and S cell cycle phases, which may suggest that this is a highly proliferative cell population.

When we integrated the dataset with Harmony across the sample types, we find that to a large degree, the above features are retained (Figure 8). Please note that the seurat clusters are assigned arbitrarily and do not correspond to the clusters indicated in the unintegrated umap plots (Figure 7). We also indicate how the proportion of sample types change across the clusters before and after integration with either Seurat or Harmony (Figure 9). From Figure 9, we observe that whilst integration may lead to a more equal distribution of sample types across the clusters, the effect is not pronounced. Therefore, it remains unclear at this stage of our investigation whether the batch effect integration with either Seurat or Harmony has led to an improvement in cell clustering.

4 Differential gene expression analysis

For simplicity, we focus on analysing the differentially expressed genes (cluster biomarkers) for each of the 14 clusters identified by the Harmony integration. Cluster biomarkers were identified via the `FindAllMarkers` function, using the Wilcoxon rank sum test for differential gene expression testing. This approach identifies genes that are either enriched or depleted in a particular cluster when compared against the average gene expression across all the remaining clusters. The full list of marker genes can be found in `results/clusterDEGs/harmony/all_markers.csv`, and genes that are enriched (but not those that are depleted) in the respective clusters are found in `results/clusterDEGs/harmony/all_markers_pos_only.csv`. The heatmap in Figure 10 shows the expression of the top 5 marker genes, ranked in increasing levels of their adjusted p-values, for each of the 14 clusters. The Dotplot in Figure 11 visualises the same information in a different format.

5 Summary

1. QC retains around 9600 cells for downstream analysis.
2. Sample type (e.g. BL,NN) appears to be a potential batch effect in the dataset.
3. Batch effect correction with Harmony or Seurat does not lead to an appreciable change in sample type distribution across clusters and it is unclear if they assist in meaningful cluster identification.
4. Cluster biomarkers for 14 clusters were identified and should now be used to identify the clusters.

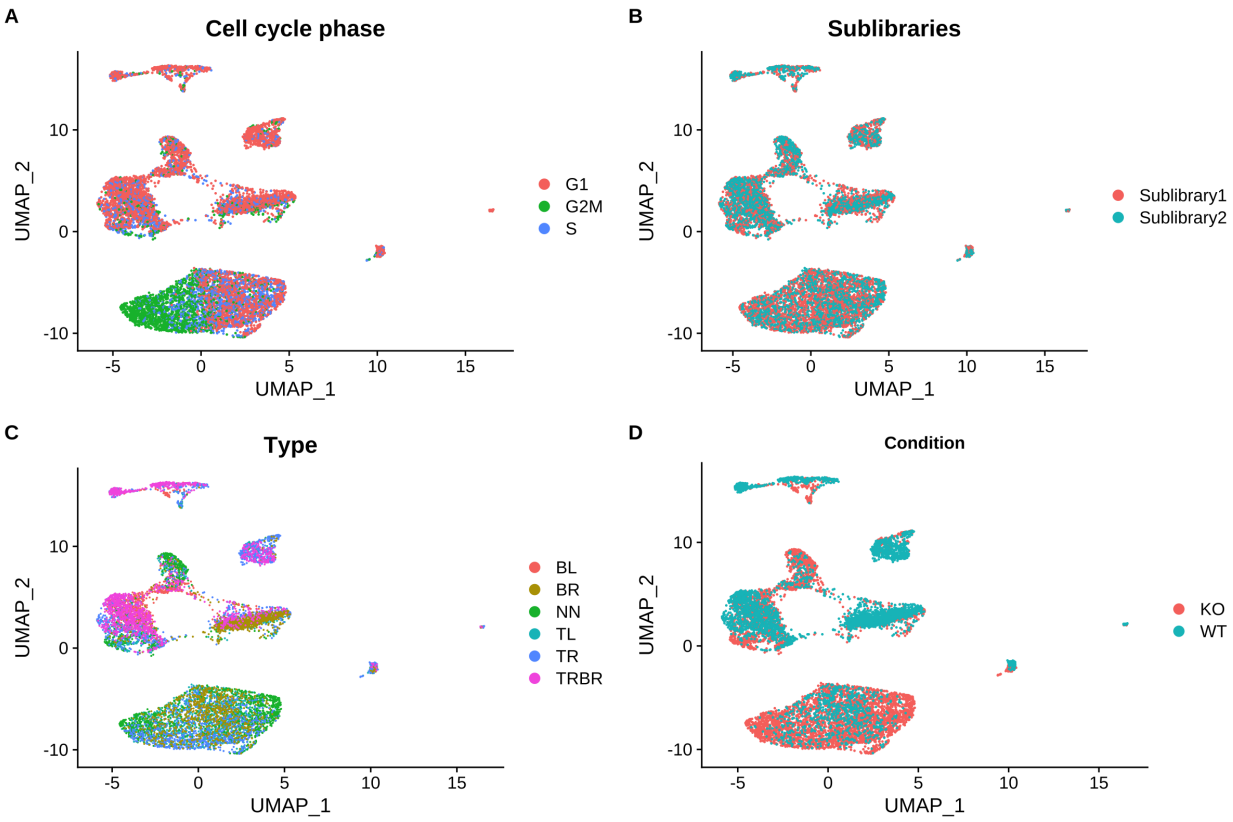

Figure 6: Assessing for batch effects in dataset

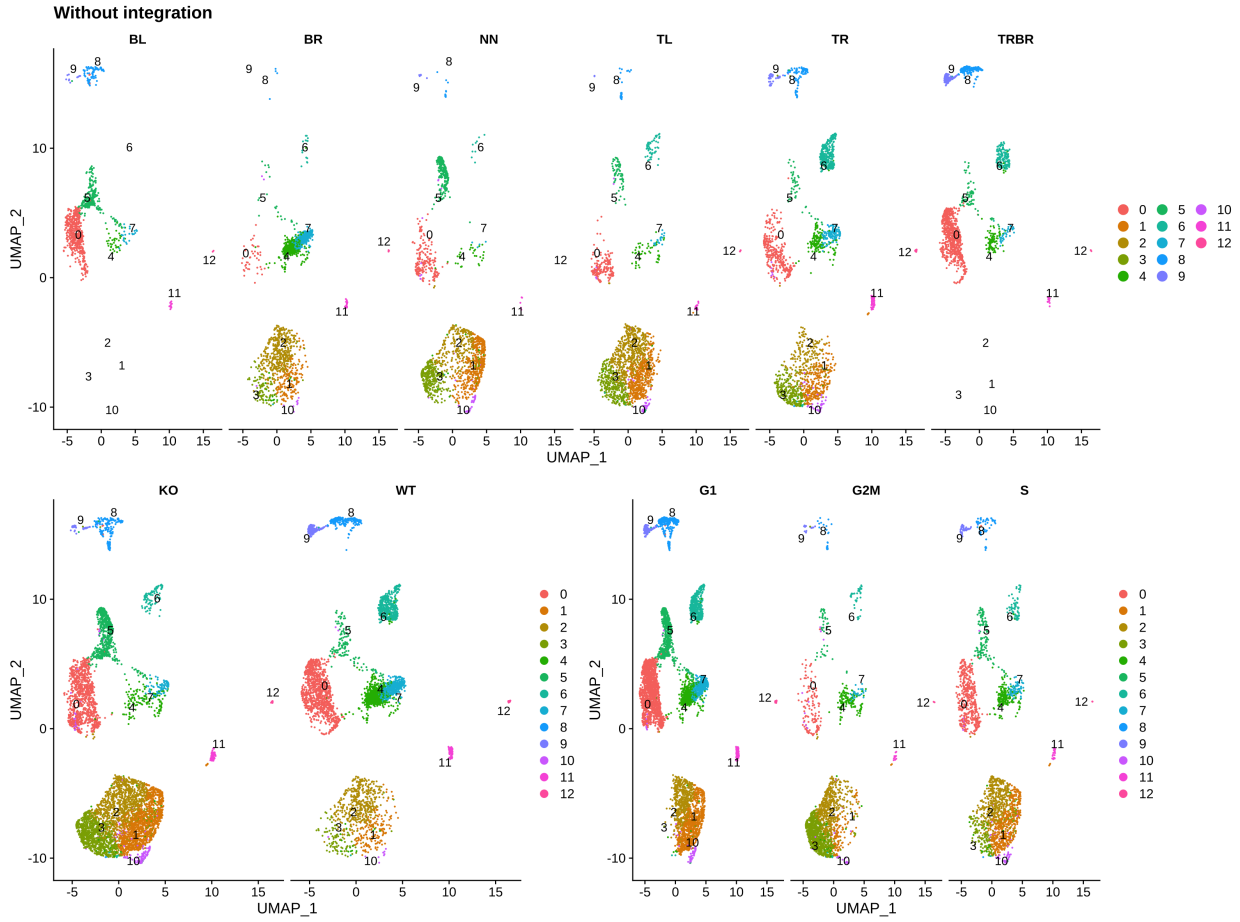

Figure 7: Faceted UMAP plots for original, unintegrated samples

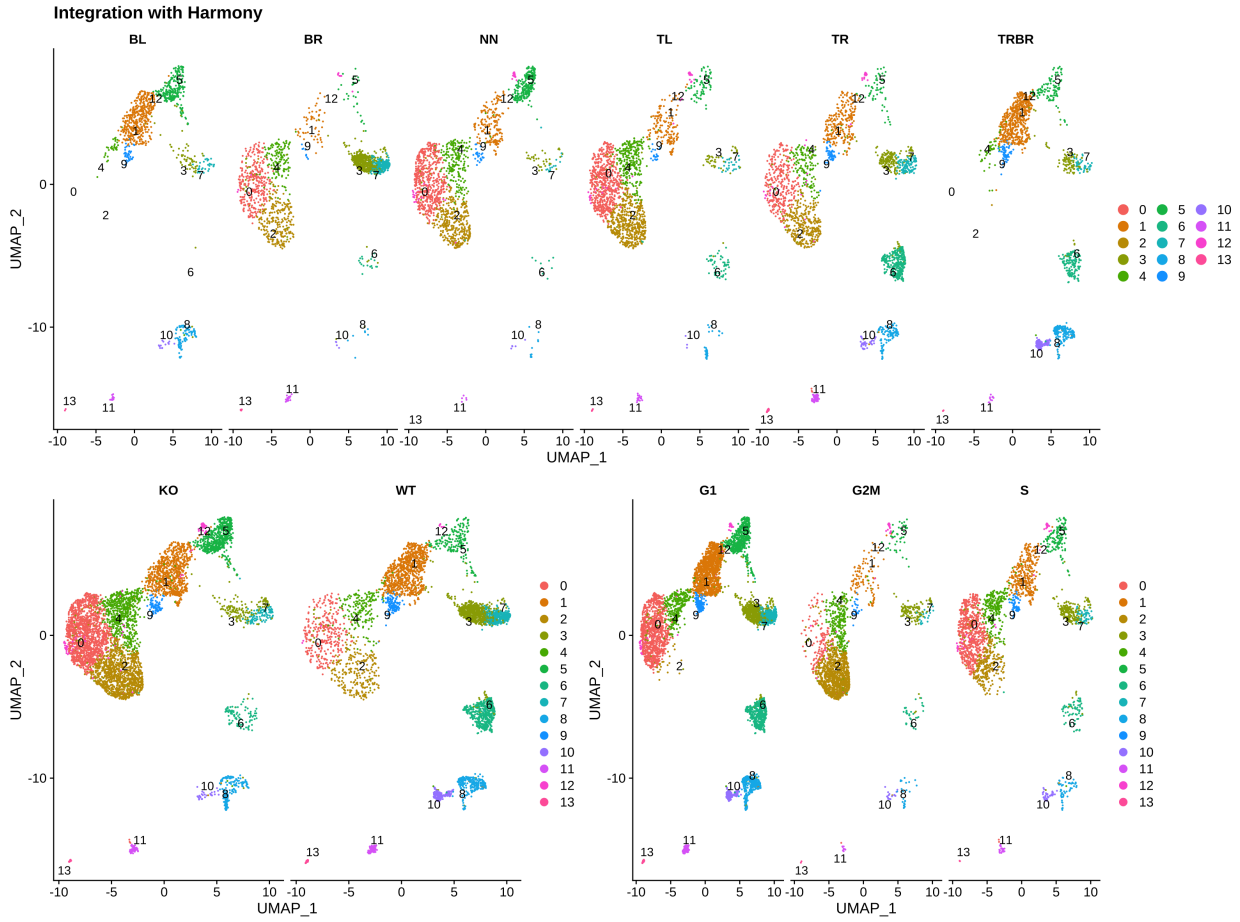

Figure 8: Faceted UMAP plots for Harmony integrated samples

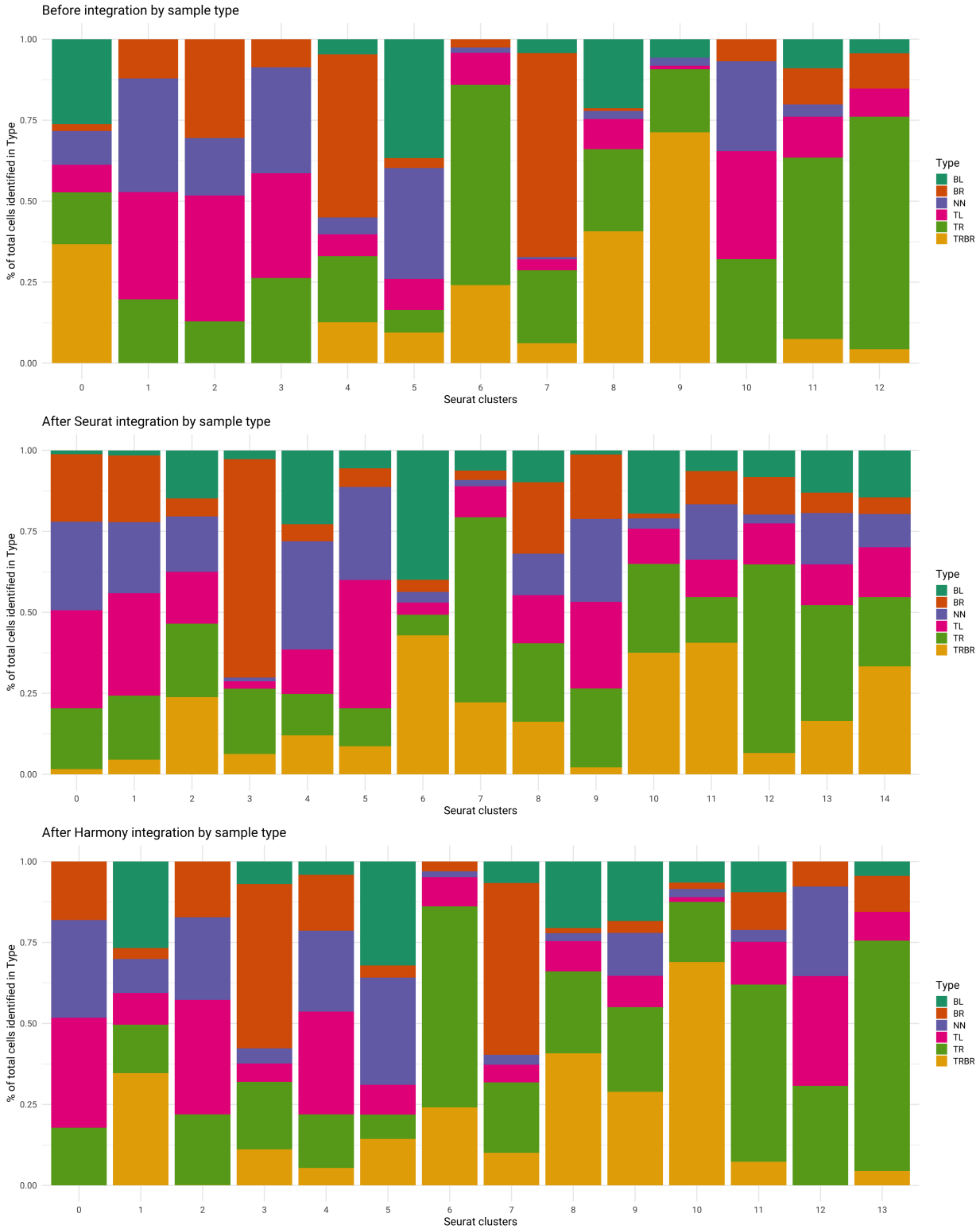

Figure 9: Proportion of sample types across clusters before and after data integration

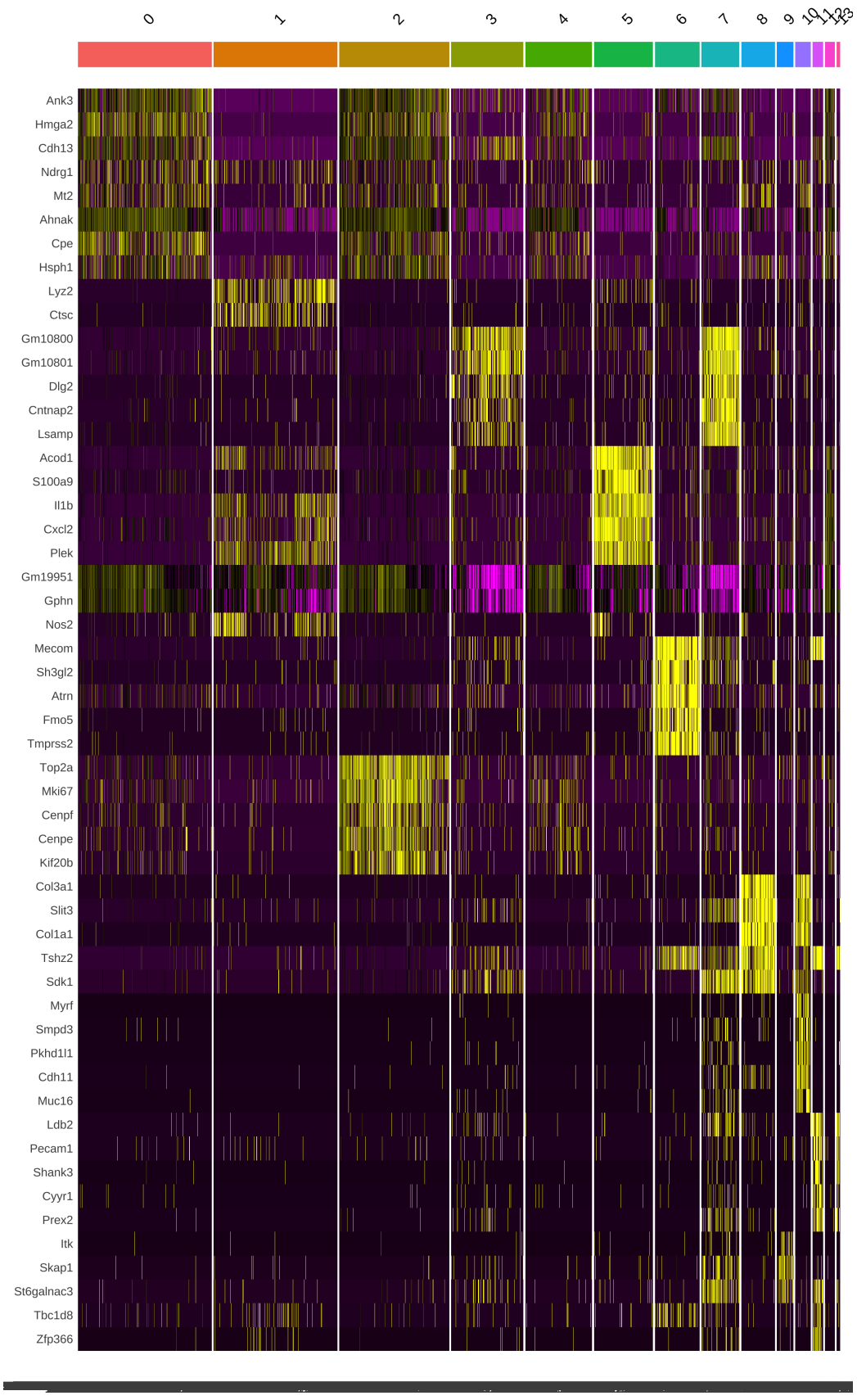

Figure 10: Heatmap of differentially expressed genes for Harmony integrated scRNA-seq dataset

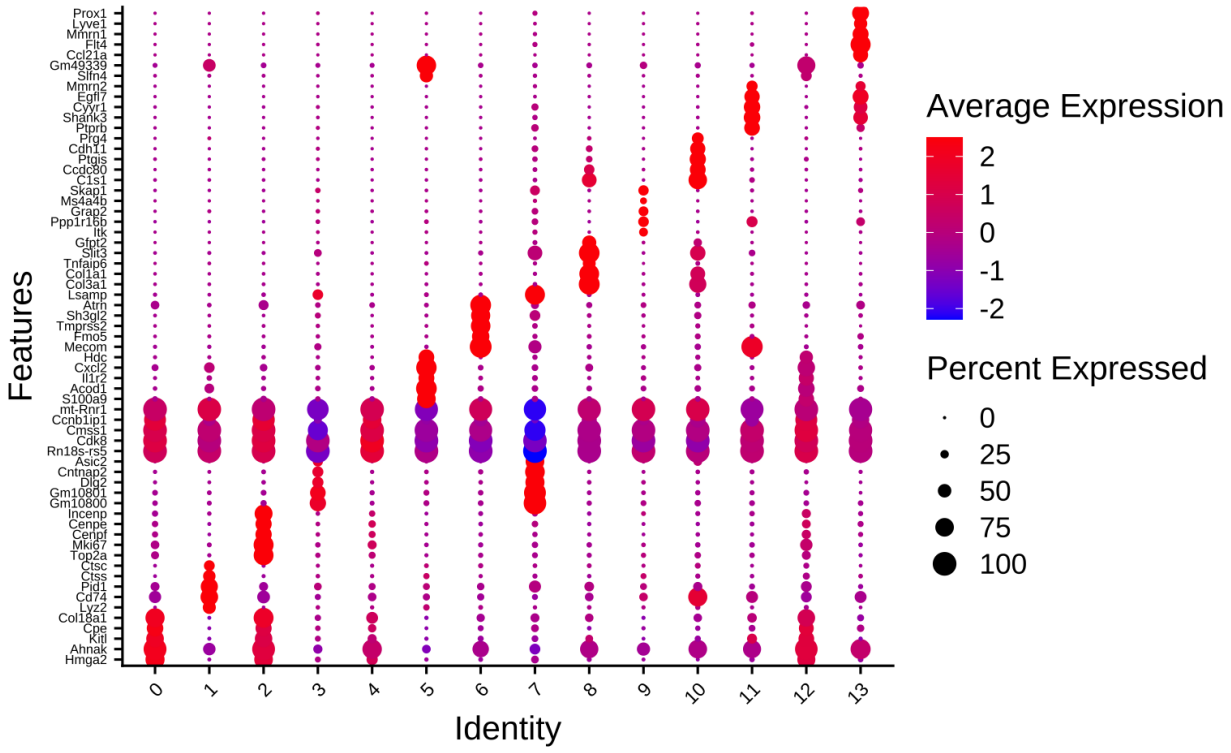

Figure 11: Dotplot of differentially expressed genes for Harmony integrated scRNA-seq dataset

6 Next steps

To assess the quality of the data and the effectiveness of the data integration and clustering procedure, we will need to determine the extent to which the clusters identified are biologically meaningful, and whether we are able to recover known cell populations in the bladder tissue. Supervised annotation can be carried out by the investigators to identify the most likely cell type that each cluster corresponds to. This can be done by examining the expression of the biomarkers in each cluster present in the excel files, and comparing them against marker genes from the literature. In addition, for each sample type, we can ask whether the absence of cells within particular clusters are possibly meaningful. For instance, we will want to check whether it makes sense for cells in the BL and TRBR samples to be absent from clusters 1,2,3 and 10.

We also provide investigators with an anndata h5ad output file, which they can upload into the cellxgene tool to interactively explore the clusters, sample annotations and carry out differential gene expression between subsets of cells in the web browser (Figure 12).

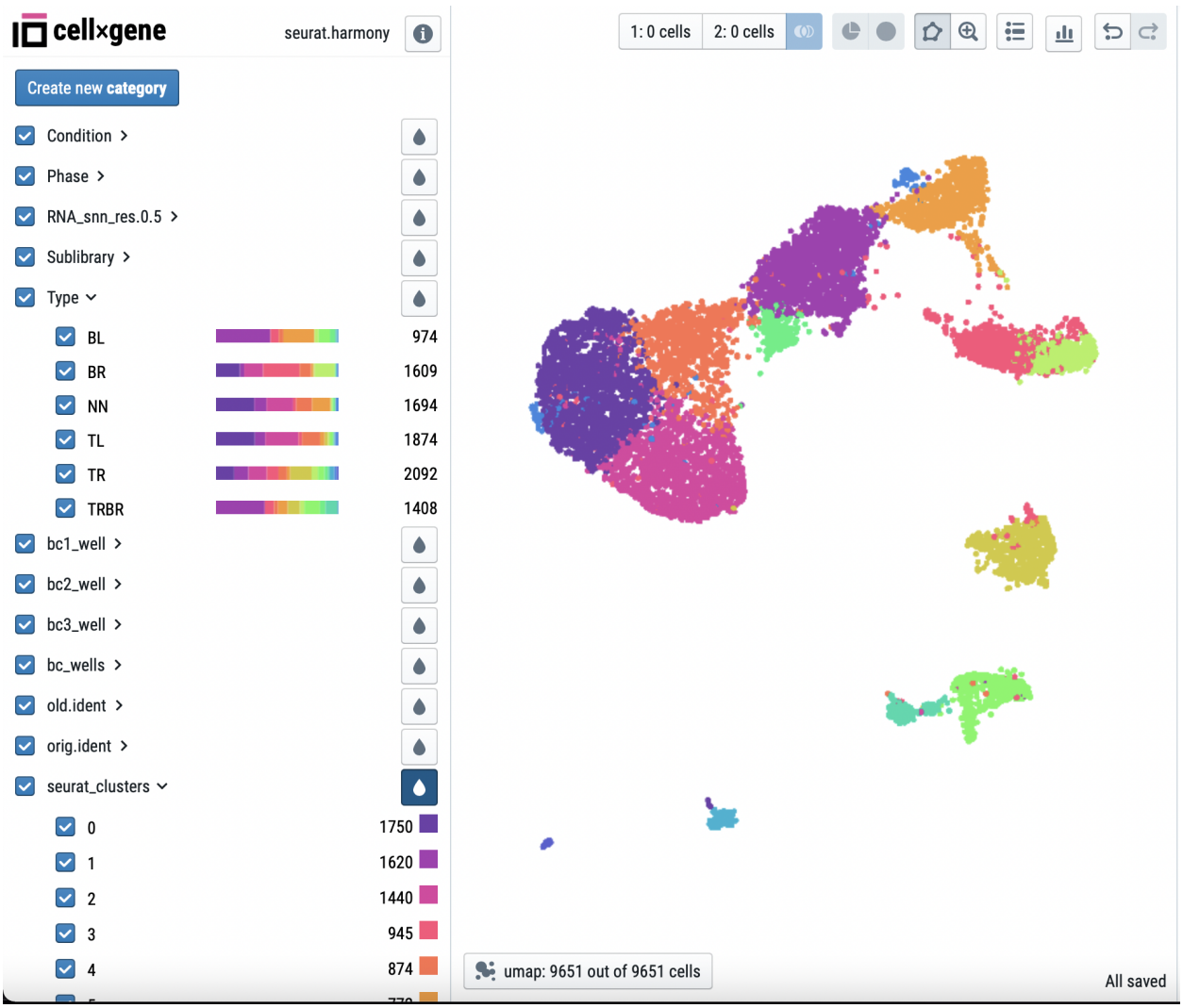

Figure 12: cellxgene browser interface for Harmony integrated scRNAseq dataset