scRNA Results

The pipeline is based on the standard Seurat (version 4) workflow and consists of a number of stages. Below is a list of the files (mostly plots/pdfs) and a short description of their contents.

Details on the Seurat package: https://satijalab.org/seurat/

The various files are arranged in folders by stage to help organize. Also the files will start with the ID of your project so will have the form `stage1/{ProjNo}_plt_01_QC.pdf` so for example:

```
stage1/p23456_C_plt_01_QC.pdf
stage2/p23456_C_plt_21_PCADimMetric.pdf
stage3/p23456_C_plt_31_ClusterMarkers_SCT_snn_res.0.1_FDR_0.05_logFC_1.
```

are some of the files for projct `p23456_C`.

Stage I - QC/Cell Level filtering

This stage looks at some QC values including cell level filter thresholds, cell cycle bias and most variable genes.

- stage1/{ProjNo}_plt_01_QC.pdf: Plots showing a number of metrics to assess the quality of cells and help in the selection for threshold for cell filtering. The key metrics plotted are:
  - nFeature_RNA: the numnber of genes detected in a given cell
  - nCounts_RNA: the number of unique molecules found in each cell
  - percent.mt: the number of reads that map to mitochondrial genes.
- stage1/{ProjNo}_plt_02_CellCycle.pdf: A plot of the first two PCA coordinates colored by the cell cycle phase of the cells which is used to assess the extent of the cell cycle signal in the dataset.

- `stage1/{ProjNo}_plt_11_PostFilterQCTbls`: Tables showing the number of cells that are filtered/kept for the downstream anaylsis.
- `stage1/{ProjNo}_plt_12_PostIntegrateCC`: Plot of top two PCA coordinates, as the previous plot, but now after normalization and scaling with the cell cycle score regression done.
- `stage1/{ProjNo}_plt_13_VariableFeatures`: The final plot show the genes with highest variability. Should be inspected to determine if there are any potential artifacts that should be filterd out (such as an ribosomal or mitochondrial genes)

Stage II - Linear reduction, clustering and projection

At this stage the PCA transformation is done on the scaled data using the top (typically 2000) more variable features (genes) determined in the previous step. The top N PCA coordianates (typically 50) are then used to both cluster the data and also to computed the 2D projection (UMAP). The clustering is done at a number of different resolutions.

- `stage2/{ProjNo}_plt_21_PCADimMetric`: Variance as a function of pca coordinate
- `stage2/{ProjNo}_plt_22_UMAP_20`: UMAP's colored by:
  - Cluster number at various resolutions
  - Samples: look for possible sample batch effects
  - Cell Cycle Phase: another assesment of how well cell cycle effects were removed
- `stage2/{ProjNo}_plt_23_ClusterChart_20`: Fraction of samples per cluster, cluster per samples, cell cycle phase per clusters. QC for possible batch effects and to help decided the optimal cluster number.
- `stage2/{ProjNo}_plt_24_SampleBias`: This plots are similar to `plt_23_ClusterChart_20` but focus on the fraction of each sample in each cluster for each resolution. They are important to inspect for potential sample bias. If there is a sample bias then the first two stages should be repeated with changes to the:
  - cell filtering

    - addition of gene filtering
    - switch from simple sample merging to Seurat more complex integration mode

Stage III - Cluster specific marker genes

For a given cluster resolution; either choosen by default of selected after review of the cluster QC plots we find the genes that are differentially expressed between the clusters. We show the distribution of these marker genes expressions amount the cluster. Also included is a table of the marker genes.

- `stage3/{ProjNo}_plt_31_ClusterMarkers_SCT_snn_res.0.1_FDR_0.05_logFC_1`:
- `stage3/{ProjNo}_plt_32_ClusterMarkersDot_SCT_snn_res.0.1_FDR_0.05_logFC_1`:
- `stage3/{ProjNo}_plt_33_ClusterHeatmap_SCT_snn_res.0.1_FDR_0.05_logFC_1`:
- `stage3/{ProjNo}_plt_34_ClusterUMAP_SCT_snn_res.0.1_FDR_0.05_logFC_1`:

Stage IV - Cell Types

Cell type level analysis. UMAPs showing the various cell types as identified using the `singleR` package with the one of the following RNAseq databases:

- ImmGenData: The ImmGen reference consists of microarray profiles of pure mouse immune cells from the project of the same name (Heng et al. 2008). This is currently the most highly resolved immune reference - possibly overwhelmingly so, given the granularity of the fine labels.
- Mouse RNA-seq: This reference consists of a collection of mouse bulk RNA-seq data sets downloaded from the gene expression omnibus (Benayoun et al. 2019). A variety of cell types are available, again mostly from blood but also covering several other tissues.
- `stage4/{ProjNo}_plt_b_01_CellTypes_SingleR_ImmGenData_fine`:

- stage4/{ProjNo}_plt_b_01_CellTypes_SingleR_MouseRNAseqData_fine: