R Packages for Single-Cell RNA-Seq

v2019-06

Simon Andrews
simon.andrews@babraham.ac.uk
Major Package Systems

Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R


SEURAT
R toolkit for single cell genomics

https://satijalab.org/seurat/

Monocle 3
An analysis toolkit for single-cell RNA-seq.

https://cole-trapnell-lab.github.io/monocle3/
What do they provide?

• Data Structure for modelling scRNA-Seq
  – Counts
  – Normalisations
  – Metadata
  – Clusters
• Convenience methods
  – Data access
  – Data parsing
  – Data access
  – Simple transformations
What do they provide?

• Implementations of common methods
  – Data Normalisation
  – Dimensionality reduction
    • PCA
    • tSNE
    • UMAP
• Plotting
  – Projections
  – QC
  – Standard graphs (scatterplots, violin plots, stripcharts)
What do they provide?

• Statistics
  – Enriched genes
  – Differential expression

• Novel functionality
  – Seurat
    • Feature anchors to match datasets
  – Monocle
    • Trajectory mapping
Seurat

• Probably the most popular choice (monocle is gaining though)

• Used to be a bit of a mess
  – Version 3 fixes a lot of issues and is nicer

• Lots of built in functionality

• Lots of nice examples on their web pages
Seurat Data Structure

• Single object holds all data
  – Build from text table or 10X output

<table>
<thead>
<tr>
<th>data</th>
<th>S4 [15969 x 5058] (Seurat::Seurat)</th>
<th>S4 object of class Seurat</th>
</tr>
</thead>
<tbody>
<tr>
<td>assays</td>
<td>list [1]</td>
<td>List of length 1</td>
</tr>
<tr>
<td>meta.data</td>
<td>list [5058 x 3] (S3: data.frame)</td>
<td>A data.frame with 5058 rows and 3 columns</td>
</tr>
<tr>
<td>active.assay</td>
<td>character [1]</td>
<td>'RNA'</td>
</tr>
<tr>
<td>active.ident</td>
<td>factor</td>
<td>Factor with 5058 levels: &quot;course&quot;, &quot;course&quot;, &quot;course&quot;, &quot;course&quot;, &quot;course&quot;, &quot;cour,...</td>
</tr>
<tr>
<td>graphs</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
<tr>
<td>neighbors</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
<tr>
<td>reductions</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
<tr>
<td>project.name</td>
<td>character [1]</td>
<td>'course'</td>
</tr>
<tr>
<td>misc</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
<tr>
<td>version</td>
<td>list [1] (S3: package_version, num)</td>
<td>List of length 1</td>
</tr>
<tr>
<td>commands</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
<tr>
<td>tools</td>
<td>list [0]</td>
<td>List of length 0</td>
</tr>
</tbody>
</table>
Seurat Data Structure

• Metadata
  – QC metrics
  – Imported classifications
  – Derived clusters
  – Some defined – can add your own

• Access directly or indirectly
  – `data$my.qc.metric`
  – `data@meta.data$my.qc.metric`
Seurat Data Structure

• Counts
  – Top level is a matrix (rows = genes, cols = cells)
  – Shortcut to `data@assays$RNA@counts`

• Normalised data
  – A second independent matrix
  – `data@assays$RNA@data`

• Can filter by subsetting the top level matrix
Seurat Data Structure

• Reductions
  – `data$projections`
  – Rows = cells, Cols = Projection axes
    • PCA
    • tSNE
    • UMAP

• Graphs
  – `data$graphs`
  – (Sparse) Distance matrices
  – Used for graph based clustering
Seurat Methods

• Data Parsing
  – Read10X
  – CreateSeuratObject

• Data Normalisation
  – NormalizeData
  – ScaleData
Seurat Graphics

• Violin Plot – metadata or expression ($VlnPlot$)
• Feature plot ($FeatureScatter$)
• Projection Plot ($DimPlot$, $DimHeatmap$)
Seurat Statistics

- Select Variable Genes `FindVariableFeatures`
- Build nearest neighbour graph `FindNeighbors`
- Build graph based cell clusters `FindClusters`
- Find genes to classify clusters (multiple tests) `FindMarkers`