nHuSC: an R package and shiny applet to visualise the scRNA-seq expression profiles of left ventricle human cardiac cells

Amanda Ho

17/08/2020

1. Background

The **nHuSC** R shiny applet and associated R package was developed to visualise the single-cell RNA-seq expression profiles of 78,264 human cardiac cells from the left ventricle. The background information on the scRNA-seq data, pre-processing of data, cell clustering and cell-type identification steps are briefly discussed in the subsequent paragraphs. **nHuSC**'s input is the outcome of the counts normalisation, guided clustering and differential gene analysis pipelines (the respective input format is described below). The full code used for generation of **nHuSC**'s input are available on Github at https://github.com/Amandahsr/UROPSBulkTissueDeconvolution.git under the "Guided Clustering (Seurat)" folder.

The R shiny applet enables the user to select a gene ID or symbol, and it generates a series of interactive plots for data viusalisation and exploratory analysis.

1.1 Single-cell RNA-seq data

The 10X scRNA-seq data was obtained from Tucker et al (Tucker et al, 2020). 56 single nuclei RNA-seq experiments were done on the four chambers of normal heart tissues donated by 7 individuals. In total, raw UMI counts of 287,269 cells and 33,694 genes were obtained in the scRNA-seq data.

1.2 Pre-processing of single-cell RNA-seq data

scRNA-seq data was normalised by dividing the counts for each transcript by the total counts per cell, multiplying the result by 10,000 before applying natural log. Pre-processing of genes was performed to retain genes that were consistently expressed across cells. Genes were retained if they met two criteria: 1) the gene had counts in at least 10 cells, and 2) the gene belongs to the top 30% highly expressed genes across cells. Additionally, only left ventricle cells were retained to focus on clustering cells from the largest tissue in the heart. As a result, 78,264 cells and 7479 genes remained for guided clustering using Seurat 3.1.5 (Butler et al, 2018).

1.3 Guided clustering using Seurat

HVGs were selected for clustering in Seurat to highlight meaningful biological signals. Out of 7479 genes, 5114 HVGs remained for clustering. Cells were clustered using the Louvain algorithm and `UMAP` was run on the normalised pre-processed data using the first 50 PCs (McInnes et al, 2018). Differential gene analysis was run on cell clusters (avg logFC >= 0.25).

1.4 Cell-type identification

The clusters were annotated based on evidence from the significant up-regulation of known cardiac cell-type markers generated by Seurat combined with pathway analysis of upregulated genes generated by DAVID 6.8 (Dennis et al, 2003). Nine distinct cell types were identified i.e. cardiomyocytes (36,035 cells), fibroblasts (19,638 cells), endothelial cells (9505 cells), smooth muscle cells (5508 cells), macrophages (3855 cells), pericytes (2194 cells), B-cells (686 cells), T-cells (563 cells) and neuronal cells (280 cells).

2. The input matrices

nHuSC expects as input a set of .txt files of a specific format briefly described here.

- CorrCounts: a matrix of the normalised gene counts of 5114 genes and 78,264 cells. The first column (column name: `Marker`) contains the gene IDs in the format *EnsemblID:GeneSymbol*. The column names of the other columns are the unique cell IDs.
- DesignTable: a matrix summarising the various predicted cell characteristics of our dataset. The first column (column name: `SampleID`) contains the unique cell IDs depicted in the CorrCounts matrix (in the same order). Some important predicted characteristics are the number of `detected genes` per cell, the UMAP dimensions (`UMAP_1` and `UMAP_2`) and the estimated `cell types`.
- DEs: a matrix with the differential expression analysis estimates. The first column (column name: `Marker`) has the gene IDs formatted as in the CorrCounts matrix. Other essential statistics are the `avg_logFC`, `PValue` and `FDR` that are directly obtained from Seurat. The last column, `CellType` indicates the predicted cell-type identity to characterise the pairwise tests. For example, CellType = `Fibroblast` indicates the comparison of differential genes expressed in Fibroblast cells vs differential genes expressed in the rest of the cell types.

3. The nHuSC R package

3.1 Package installation

The **nHuSC** package can be installed in R using the tar.gz file. This will automatically install and load all dependencies. After downloading the tar.gz file, the package can be installed in R using the command:

```
install.packages("pathdirectory/nHuSC.1.0.tar.gz", repos = NULL, type = "source")
```

The data visualisation is performed in 3 steps, i.e. data loading, gene selection and data visualisation. Here, the functions of interest are briefly described. Detailed information on package functions can be obtained from the package help pages (e.g. `?readData` for the help page of the readData() function).

3.2 Data loading

The data loading step is done as:

```
data(CorrCounts)
data(DesignTable)
data(DEs)
data<-readData(CorrCounts,DesignTable,DEs,Markers_file=NULL,is.Exact=TRUE,
               avg_logFC_cutoff=0,FDR_cutoff=1)
```

```
## [1] "***** The data have been successfully loaded! *****"
```

```
# listed components
names(data)
```

```
##  [1] "Counts"             "Design"             "DEstats"
##  [4] "Annotation"         "Status"             "Exact_Marker_Match"
##  [7] "avg_logFC"          "FDR"                "Output_Folder"
## [10] "Dimensions"         "temp"               "filteredData"
## [13] "Date_Stamp"
```

```
# the CorrCounts
data$Counts[1:5, 1:2]
```

```
##                          LV_1723_1_TGGGCGTAGACCACGA-1
## ENSG00000000460:C1orf112                            0
## ENSG00000000971:CFH                                 0
## ENSG00000001084:GCLC                                0
## ENSG00000001629:ANKIB1                              0
## ENSG00000001631:KRIT1                               0
##                          LV_1723_1_AGGTCCGAGCACACAG-1
## ENSG00000000460:C1orf112                       0.0000
## ENSG00000000971:CFH                            0.0000
## ENSG00000001084:GCLC                           0.0000
## ENSG00000001629:ANKIB1                         1.7576
## ENSG00000001631:KRIT1                          0.0000
```

```
# the design_table
data$Design[1:5, ]
```

```
##                                                SampleID      CellType
## LV_1723_1_TGGGCGTAGACCACGA-1 LV_1723_1_TGGGCGTAGACCACGA-1 Cardiomyocyte
## LV_1723_1_AGGTCCGAGCACACAG-1 LV_1723_1_AGGTCCGAGCACACAG-1 Cardiomyocyte
## LV_1723_1_GCATGCGCAGACAAAT-1 LV_1723_1_GCATGCGCAGACAAAT-1 Cardiomyocyte
## LV_1723_1_CATATTCCACGGTGTC-1 LV_1723_1_CATATTCCACGGTGTC-1 Cardiomyocyte
## LV_1723_1_CGTTGGGTCATTGCCC-1 LV_1723_1_CGTTGGGTCATTGCCC-1 Cardiomyocyte
##                                UMAP_1   UMAP_2 nCount_RNA Detected_genes
## LV_1723_1_TGGGCGTAGACCACGA-1 23.85389 4.233381   2071.952           1113
## LV_1723_1_AGGTCCGAGCACACAG-1 26.31149 4.850675   2188.501           1167
## LV_1723_1_GCATGCGCAGACAAAT-1 24.25173 4.761887   2144.312           1132
## LV_1723_1_CATATTCCACGGTGTC-1 24.13116 4.157863   2103.480           1091
## LV_1723_1_CGTTGGGTCATTGCCC-1 24.51045 4.367923   2146.733           1129
##                              Colors_CellType
## LV_1723_1_TGGGCGTAGACCACGA-1         #FF61C3
## LV_1723_1_AGGTCCGAGCACACAG-1         #FF61C3
## LV_1723_1_GCATGCGCAGACAAAT-1         #FF61C3
## LV_1723_1_CATATTCCACGGTGTC-1         #FF61C3
## LV_1723_1_CGTTGGGTCATTGCCC-1         #FF61C3
```

```
# the DE
data$DEstats[1:5, ]
```

```
##                     Marker     PValue  avg_logFC        FDR CellType
## 1 ENSG00000000460:C1orf112  -1.00e-01 -0.1000000  -1.00e-01   B_Cell
## 2      ENSG00000000971:CFH  -1.00e-01 -0.1000000  -1.00e-01   B_Cell
## 3     ENSG00000001084:GCLC  1.22e-109  0.3000465  9.14e-106   B_Cell
## 4   ENSG00000001629:ANKIB1  -1.00e-01 -0.1000000  -1.00e-01   B_Cell
## 5    ENSG00000001631:KRIT1  -1.00e-01 -0.1000000  -1.00e-01   B_Cell
```

```
# the data stamp (for file storage)
data$Date_Stamp
```

```
## [1] "Tue_Aug_18_2020_00.58.20"
```

Among the `data` components we find the CorrCounts (`Counts`), DesignTable (`Design`) and DE (`DEstats`) matrices as well as other automatically generated information such as: `Annotation` with the geneIDs, `avg_logFC` with the avg logFC cut-off for preliminary marker filtering based on the differential expression analysis results (default is 0), `FDR` with the FDR cut-off for preliminary marker filtering based on the differential expression analysis results (default is 1), `Output folder` specifying the folder that stores the results and `Date stamp` that assigns unique filenames for storage.

3.3 Gene selection

Next, the user selects the gene for visualisation by its Ensembl ID or its gene symbol or the combination *EnsemblID:GeneSymbol*. Any of the genes present in the CorrCounts matrix can be selected. Below, we indicate some examples:

```
data_MYH7 <- MarkerQuery(Data = data, marker = "MYH7")
```

```
## [1] "Gene ENSG00000092054:MYH7 is selected for visualisation."
```

Parameter `Data` accepts the outcome of readData() function while parameter `marker` takes the ID of interest. The above example shows the selection of *MYH7* gene for further analysis. Alternatively, the user can select the respective Ensembl ID as:

```
data_MYH7_alt <- MarkerQuery(Data = data, marker = "ENSG00000092054")
```

```
## [1] "Gene ENSG00000092054:MYH7 is selected for visualisation."
```

or as a combination of the two:

```
data_MYH7_alt2 <- MarkerQuery(Data = data, marker = "ENSG00000092054:MYH7")
```

```
## [1] "Gene ENSG00000092054:MYH7 is selected for visualisation."
```

If the selected gene is not present in the data, **nHuSC** generates an appropiate error and the analysis stops, prompting the user to select another ID:

```
data_err <- MarkerQuery(Data = data, marker = "TP53")
```

```
## [1] "This gene ID does not match any of the existing IDs!"
```

3.4 Data visualisation

nHuSC can generate two types of interative plots for gene expression visualisation and exploration. First, users can select UMAP to depict the expression levels of the selected gene with a colour gradient (blue-to-red). Mouse-over on the plot reveals the estimated normalised expression level and the estimated cell type of each cell. Double clicking on any of the dots highlights all cells of the same cell type. The logical parameters `show.plot` and `save.plot` determine whether the plot will be shown on screen and/or stored as an html (interactive) file. Finally, clicking on the gene ID of the plot legend directs the user to the respective NCBI page of the gene.

```
scatter_MYH7 <- doScatterDR(Data = data_MYH7,
                            highlight.by = "CellType",
                            show.plot = TRUE,
                            save.plot = FALSE)
```

A violin plot for each cell type can be generated as:

```
violin_MYH7 <- doViolin(Data = data_MYH7,
                        grouping.by = "CellType",
                        show.plot = TRUE,
                        save.plot = FALSE)
```

```
## [1] "***** The grouping2.by variable does not exist in the Design file. Setting grouping2.by = NULL!
```

Similar to the UMAP, clicking on the gene ID of the plot legend redirects the user to the associated NCBI page of the gene.

4. The nHuSC R shiny applet

The **nHuSC** web applet can be accessed with:

```
run_nHuSC()
```

The applet offers computationally inexperienced users a simpler alternative way to visualise the cardiac dataset at the cost of less flexibility. It only requires a prior installation of R or R studio in their system.

The three input data tables are automatically loaded upon **nHuSC**'s initialization. By clicking on *Get Started!* the user is redirected to the Quick Analysis tab.

The right part of the screen allows the user to download the input matrices and see at a glance the package's functionality. As before, the user is enabled to select the gene ID or symbol of interest and load it in the system. The *Plot Options* tab will be subsequently activated enabling the user to select one of the three data visualisation plots.

The UMAP higlights the cell types (`highlight.by = CellType`) and the violin is done by cell type. All non-differentially expressed genes will reflect avg_logFC = -0.1 and FDR = -0.1. The user can either see the plot at the bottom of the screen (`Show data`) and/or save it (`Save data`) as an htlm interactive file located in the `Data/InteractivePlots` subfolder of the package.

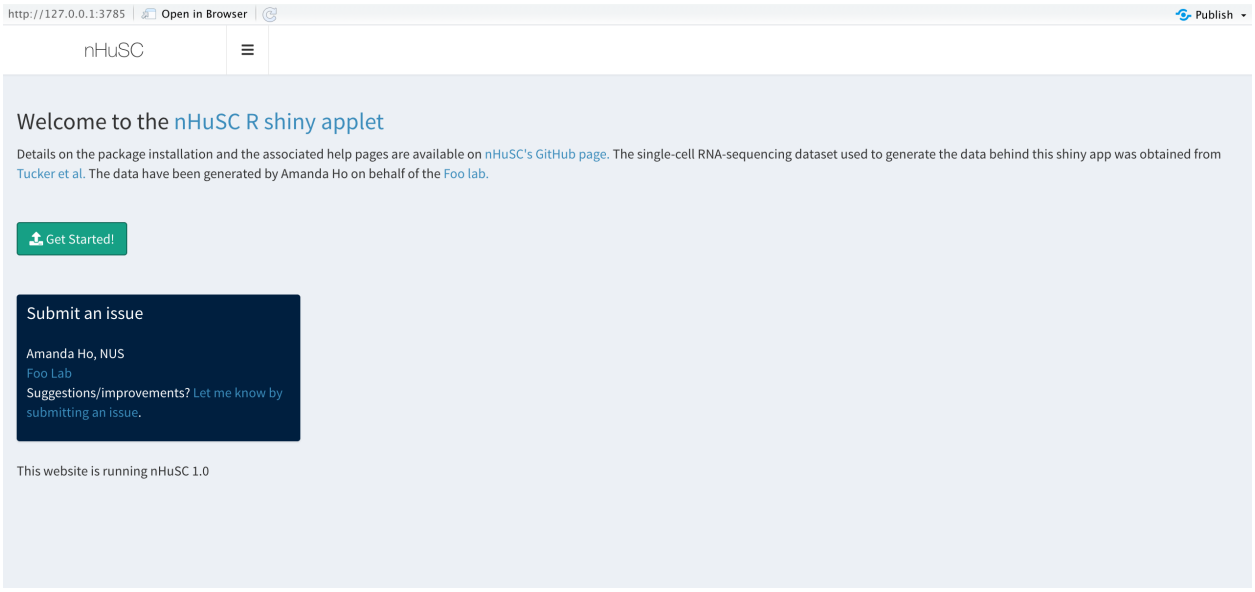

Figure 1: nHuSC welcome screen

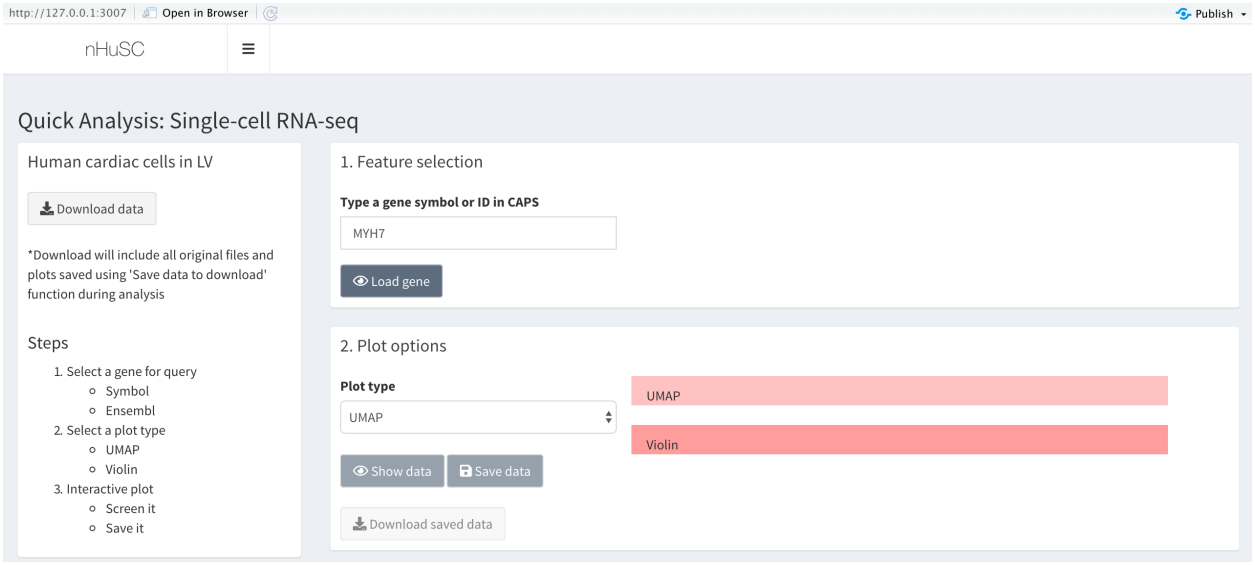

Figure 2: nHuSC quick analysis screen

5. References

- Tucker et al. "Transcriptional and Cellular Diversity of the Human Heart". *Circulation* 2020; 142:466–482.
- McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". *The Journal of Open Source Software* 2018; 3:861.
- Butler et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". *Nature Biotechnology* 2018; 36:411–420.
- Dennis et al. "DAVID: Database for Annotation, Visualization, and Integrated Discovery". *Genome Biology* 2003; 4:R60.