SALMON Implementation

Heeju Noh (heeju.noh@chem.ethz.ch)

2017-07-19

This is an example script for implementing SALMON.

Installation instruction

1. Download and unzip the SALMON package to a preferred folder.
2. Open R and install the package, using `install("yourpathname/salmon/R")`. Your package is inatalled in R library directory.
3. Load the package, using `library(salmon)`.

Example data

SALMON package includes microarray data from the chromatin targeting study using mouse pancreatic beta cells [1]:

`lfc`: log2FC data, pre-processed as described in SALMON manuscript

`glist`: The list of gene symbols corresponding to the rows in the log2FC data

`tobject`: The table of sample descriptions including time points (if in time-series) and group indices (same index for the same drug)

`tftg`: Transcription factor (TF)-gene network for mouse pancreas cells obtained from CellNet database [2]

`ppi`: protein-protein interactions for mouse cells obtained from STRING database [3]

Preparation for SALMON inputs

SALMON requires log2FC data and slope matrix (if data are time-series) and the adjacency matrix of protein-gene network (PGN).

The slope matrix can be calculated using `generateSlope` function.

```
tp <- tobject$time ## a vector of time points of the samples in the matrix lfc
group <- tobject$group ## a vector of indices of the grouped samples
slope <- generateSlope(lfc = lfc, tp = tp, group = group)
```

PGN is constructed using TF-gene and protein-protein interactions. Here, we used the same thretholds as described in the SALMON manuscript.

```
pgn <- generatePGN(glist = glist, tftg = tftg, ppi = ppi, tftg_thre = 0, ptf_thre = 0,
                   ppi_thre = 500)
```

Calculating protein perturbation scores by SALMON

SALMON can provide an overall score for each protein in the grouped samples. If you want a score for each sample, set each sample to a single group. Here, we grouped the sample for each different drug, and used 10-fold cross validation (default). The outcome of SALMON is a list of protein score matrix and weighted

PGN. To note, the rows in the outcome PGN matrix correspond to genes having at least one regulator based on the prior PGN (i.e. zeros for the others).

```
result <- salmon(lfc = lfc, slope = slope, pgn = pgn, grplist = group)
result$Pscore ## protein score matrix
result$A ## weigted PGN
```

Parallel computation is also available in SALMON. The following example shows parallel computation (par=TRUE) using 4 cores.

```
result <- salmon(lfc = lfc, slope = slope, pgn = pgn, grplist = group, par = TRUE,
                 numCores = 4)
```

Ranking the proteins based on the magnitudes of estimated scores

The greater magnitude of the scores by SALMON implies the higher perturbation caused by the drug. In this mouse dataset, trichostatin A is a well known histone deacetylase inhibitor. As an example, we can examine the ranks of histone deacetylases (Hdac1-Hdac11) for trichostatin.

```
oi <- order(abs(result$Pscore[,16]),decreasing=TRUE)## 16th column for trichostatin A
ranked.list <- glist[oi]
hdac <- glist[5096:5103] ## hdac genes

hdac.rank <- list(match(hdac,ranked.list))
names(hdac.rank) <- hdac
hdac.rank
```

REFERENCES:

[1] Kubicek, S., J. C. Gilbert, D. Fomina-yadlin, A. D. Gitlin, and Y. Yuan. 2012. Chromatin-targeting small molecules cause class-specific transcriptional changes in pancreatic endocrine cells.

[2] Cahan, P., H. Li, S. A. Morris, E. Lummertz Da Rocha, G. Q. Daley, and J. J. Collins. 2014. CellNet: Network biology applied to stem cell engineering. Cell 158 (4): 903-915.

[3] Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research 43 (D1): D447-D452.