scNPF: An integrative framework assisted by network propagation and network fusion for pre-processing of single-cell RNA-seq data

2018-12-18

1 Overview

scNPF is a R package for pre-processing of single-cell RNA-seq data by leveraging the context-specific topology inherent in the given data and the network information from priori gene-gene interaction networks. *scNPF* consists of two modules (**Figure 1**), including scNPF-propagation for imputing dropouts based on random walk with restart (RWR) and scNPF-fucion for fusing multiple smoothed expression matrices to a cell-cell similarity matrix. *scNPF* is highly integrative and flexible in that the two modules are independent but interconnected. *scNPF* can be used as a general and flexible pre-processing step prior to downstream analyses of scRNA-seq data for recovering gene expression lost, correcting gene expression measurements, and learning similarities between cells.

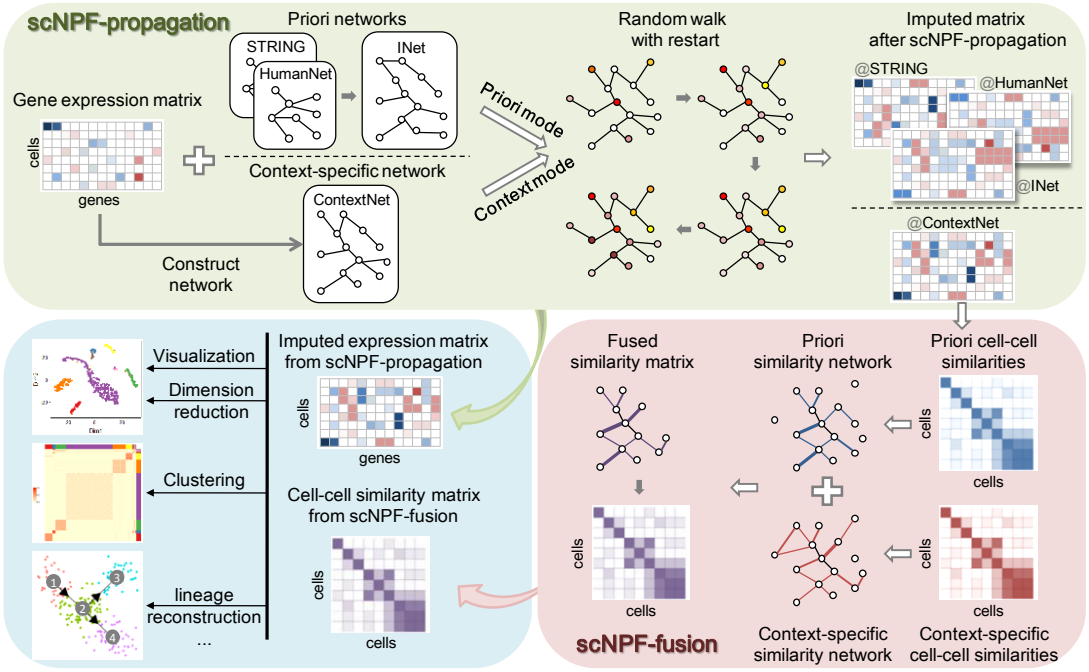

Figure 1. **Schematic diagram of the scNPF framework**. scNPF consists of two modules, scNPF-propagation for imputing dropouts and scNPF-fusion for fusing multiple smoothed expression matrices to a cell-cell similarity matrix. The outputs from scNPF-propagation and scNPF-fusion can be used for downstream analyses of scRNA-seq data, such as visualization,

dimension reduction, clustering, and lineage reconstruction.

2 Installation

You can install scNPF from github with:

```
install.packages("devtools")
library(devtools)
install_github("BMILAB/scNPF")
library(scNPF)
```

3 Preparations

3.1 Gene expression matrix

The input to *scNPF* is matrix of gene expression count. The rows correspond to genes and the columns correspond to cells. In this study, we will use the human embryonic stem cells data from [1] as example.

```
>load(system.file("data","yan.data",package = "scNPF"))
>exp.data <- yan$data
>dim(exp.data)
[1] 17916    90
>exp.data[5:7,1:3]
        Oocyte..1.RPKM. Oocyte..2.RPKM. Oocyte..3.RPKM.
PNMA1             0.679          1.343          2.125
MMP2              0.000          0.000          0.000
TMEM216           8.869         12.539         13.851
```

3.2 Gene-gene interaction network

A gene-gene interaction network (a adjacency matrix) is used to smooth expression values in the scNPF-propagation model. If users use priori mode, they should provide gene co-expression network from publicly available database or a specific gene-gene network established on your own method. In this package, we provided three human gene-gene interaction networks from different databases, including String (v9.1)[2], HumanNet (v1)[3] and an integrated network (INet)[4]. Here is an example to show the format of a adjacency matrix.

```
>load(system.file("data","string.Rdata",package = "scNPF"))
```

```
> str(string)
Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  ..@ i   : int [1:1002817] 71 150 158 643 744 800 840 1389 1656 1704 ...
  ..@ p  : int [1:16661] 0 141 312 346 393 460 525 529 552 568 ...
  ..@ Dim: int [1:2] 16660 16660
  ..@ Dimnames:List of 2
  .. ..$ : chr [1:16660] "ARF5" "M6PR" "ESRRA" "FKBP4" ...
  .. ..$ : chr [1:16660] "ARF5" "M6PR" "ESRRA" "FKBP4" ...
  ..@ x   : num [1:1002817] 1090 1536 1824 1800 1876 ...
  ..@ factors : list()
> class(string)
[1] "dgCMatrix"
attr(,"package")
[1] "Matrix"
```

If users use context mode, a gene co-expression network is automatically generated in the *scNPF.pro* function.

4 Standard analysis work-flow

4.1 Data preprocessing (optional)

scNPF is a purely pre-processing tool for scRNA-seq data. To reduce computational time, users can use R package scater[5] or Seurat[6] to remove low quality cells and filter out genes with low expression for downstream analysis.

4.2 scNPF-propagation

scNPF-propagation (*scNPF.pro*) involves a network propagation process based on RWR on a given gene-gene interaction network to obtain a distribution for each node (gene), which captures its relevance to all other genes in the network. In this step, users can use priori mode or context mode. For priori mode, user can use publicly available molecular networks to smooth scRNA-seq expression data. In this package, we provided three human gene-gene interaction networks from different databases, including String (v9.1)[2], HumanNet (v1)[3] and an integrated network (INet)[4].

```
##For priori mode
##Using String network to smooth expression values.
>load(system.file("data","string.Rdata",package = "scNPF"))
>string.data <- scNPF.pro(x=exp.data, network=string,nThreads=8)
```

```
>dim(string.data)
[1] 17916    90
>string.data[5:7,1:3]
         Oocyte..1.RPKM. Oocyte..2.RPKM. Oocyte..3.RPKM.
PNMA1          0.9134392       1.338069       1.877562
MMP2          20.7332068      21.277234      21.560455
TMEM216       21.5610074      23.772617      24.292170
##Or using HumanNet network
>load(system.file("data","humannet.Rdata",package = "scNPF"))
>hm.data <- scNPF.pro(x=exp.data,network=humannet,nThreads=8)

##Or using integrated network
>load(system.file("data","integrated.Rdata",package = "scNPF"))
>inter.data <- scNPF.pro(x=exp.data,network=INet,nThreads=8)
```

For context model, a context-specific gene-gene network is constructed from the scRNA-seq data set using the WGCNA package[7].

```
##For context mode
>context.data<- scNPF.pro(x=exp.data, network="context",nThreads=8)
```

The output of function *scNPF.pro* is a propagated gene-cell expression matrix , which could be used as input for scNPF-fusion (*scNPF.fus*), and also as the input for many other single cell tools to perform downstream analyses like dimension reduction, clustering, and visualization.

4.3 scNPF-fusion

scNPF-fusion (*scNPF.fus*) constructs a sample-similarity network for each propagated expression matrix and then integrates these networks into a single cell-cell similarity network based on a nonlinear combination method. This process consists of two main steps for data integration. First, scNPF-fusion constructs a cell-by-cell similarity matrix for the output of scNPF-propagation using two different modes, respectively. Then, both similarity matrices are iteratively and gradually fused to a coherent and combined network, employing the non-linear method of message passing theory[8].Finally, weak similarities which may be potential noises are discarded, and strong similarities are added. For example, we takes the propagated matrices from scNPF-propagation using the priori mode with the String network and the context mode as inputs and learns a matrix of similarities between cells by network fusion.

```
##Construction a cell-by-cell similarity matrix.
>similarity<-scNPF.fus(data=list(string=string.data,
context=context.data))
```

Then, by function *plotHeatmap*, we can observe the difference between

different similarities or distance metrics. Here, we compared distance metrics learned from Euclidian measure and scNPF (**Figure 2**).

```
library(gplots)
library(RColorBrewer)
#Heatmap for distance learned by scNPF-fusion
#Turn similarity to distance
plotHeatmap(1/(similarity+1),yan$label)
#Heatmap for distance learned by Euclidean distance.
data.dist <- as.matrix(dist(t(yan$data)))
plotHeatmap(data.dist,yan$label)
```

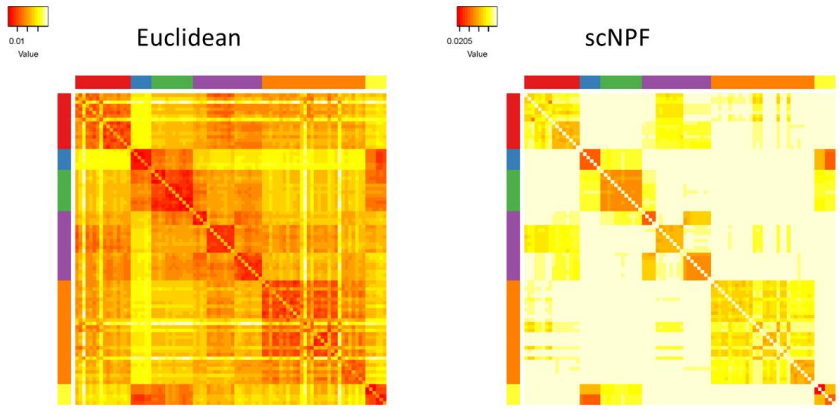

Figure 2. **Distance metrics**. Heatmaps for distances metrics from the yan data by Euclidean distances (left) and scNFP fusion (right). Cells with the same cell type (annotated by the colored axes) are grouped together.

5 Session information

```
>sessionInfo()
R version 3.5.0 (2018-04-23)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1
Matrix products: default
locale:
[1] LC_COLLATE=Chinese (Simplified)_People's Republic of China.936
[2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936
[3] LC_MONETARY=Chinese(Simplified)_People's Republic of China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_People's Republic of China.936
attached base packages:
[1] parallel stats  graphics grDevices utils datasets methods base
other attached packages:
```

```
 [1] scNPF_0.1.0   plyr_1.8.4  Matrix_1.2-14    doParallel_1.0.14
 [5] iterators_1.0.9 foreach_1.4.4 WGCNA_1.63   fastcluster_1.1.25
 [9] dynamicTreeCut_1.63-1 igraph_1.2.1
loaded via a namespace (and not attached):
 [1] Biobase_2.40.0   bit64_0.9-7  splines_3.5.0   Formula_1.2-3
 [5] assertthat_0.2.0 stats4_3.5.0 latticeExtra_0.6-28   blob_1.1.1
 [9] fit.models_0.5-14 yaml_2.1.19 robustbase_0.93-3 impute_1.54.0
[13] pillar_1.2.3  RSQLite_2.1.1  backports_1.1.2  lattice_0.20-35
[17] glue_1.2.0 digest_0.6.15  RColorBrewer_1.1-2  checkmate_1.8.5
[21] colorspace_1.3-2 htmltools_0.3.6 preprocessCore_1.42.0 pcaPP_
1.9-73
[25] pkgconfig_2.0.1  purrr_0.2.5   GO.db_3.6.0 mvtnorm_1.0-8
[29] scales_0.5.0 htmlTable_1.12  tibble_1.4.2  IRanges_2.15.18
[33] ggplot2_3.0.0 nnet_7.3-12 BiocGenerics_0.27.1 lazyeval_0.2.1
[37] survival_2.42-3  magrittr_1.5  memoise_1.1.0  MASS_7.3-50
[41] foreign_0.8-70tools_3.5.0 data.table_1.11.4matrixStats_0.53.1
[45] stringr_1.3.1 S4Vectors_0.19.22 munsell_0.5.0 cluster_2.0.7-
[49] AnnotationDbi_1.42.1 bindrcpp_0.2.2 compiler_3.5.0rlang_0.2.
[53] grid_3.5.0 rstudioapi_0.7 htmlwidgets_1.2.1  robust_0.4-18
[57] base64enc_0.1-3  gtable_0.2.0   codetools_0.2-15  DBI_1.0.0
[61] rrcov_1.4-4  R6_2.2.2    gridExtra_2.3    knitr_1.20
[65] dplyr_0.7.5    bit_1.1-14    bindr_0.1.1   Hmisc_4.1-1
[69] stringi_1.1.7  Rcpp_0.12.19    rpart_4.1-13   acepack_1.4.1
[73] DEoptimR_1.0-8       tidyselect_0.2.4
```

References

1. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J *et al*: **Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.** *Nature Structural &Amp; Molecular Biology* 2013, **20**:1131.
2. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C *et al*: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41**(Database issue):D808-D815.
3. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM: **Prioritizing candidate disease genes by network-based boosting of genome-wide association data.** *Genome research* 2011, **21**(7):1109-1121.
4. Yang F, Wu D, Lin L, Yang J, Yang T, Zhao J: **The integration of weighted gene association networks based on information entropy.** *PLoS ONE* 2017, **12**(12):e0190029.
5. McCarthy DJ, Campbell KR, Lun ATL, Wills QF: **Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R.** *Bioinformatics* 2017, **33**(8):1179-1186.
6. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data.** *Nature Biotechnology* 2015, **33**:495.
7. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC bioinformatics* 2008, **9**:559-559.
8. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale.** *Nat Methods* 2014, **11**(3):333-337.