Using the *SIMLR* package

Bo Wang[*] Daniele Ramazzotti[†] Luca De Sano[‡] Junjie Zhu[§] Emma Pierson[¶]
Serafim Batzoglou[‖]

November 8, 2016

Overview. Single-cell RNA-seq technologies enable high throughput gene expression measurement of individual cells, and allow the discovery of heterogeneity within cell populations. Measurement of cell-to-cell gene expression similarity is critical to identification, visualization and analysis of cell populations. However, single-cell data introduce challenges to conventional measures of gene expression similarity because of the high level of noise, outliers and dropouts. We develop a novel similarity-learning framework, *SIMLR* (Single-cell Interpretation via Multi-kernel LeaRning), which learns an appropriate distance metric from the data for dimension reduction, clustering and visualization. *SIMLR* is capable of separating known subpopulations more accurately in single-cell data sets than do existing dimension reduction methods. Additionally, *SIMLR* demonstrates high sensitivity and accuracy on high-throughput peripheral blood mononuclear cells (PBMC) data sets generated by the GemCode single-cell technology from 10x Genomics.

In this vignette, we give an overview of the package by presenting some of its main functions.

[*]Department of Computer Science, Stanford University, Stanford, CA , USA.
[†]Department of Pathology, Stanford University, Stanford, CA , USA.
[‡]Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi Milano Bicocca Milano, Italy.
[§]Department of Electrical Engineering, Stanford University, Stanford, CA , USA.
[¶]Department of Computer Science, Stanford University, Stanford, CA , USA.
[‖]Department of Computer Science, Stanford University, Stanford, CA , USA.

Contents

1 Changelog

1.0 implements SIMLR and SIMLR feature ranking algorithms.

2 Algorithms and useful links

Acronym	Extended name	Reference
SIMLR	Single-cell Interpretation via Multi-kernel LeaRning	Paper

3 Using SIMLR

We first load the data provided as an example in the package.

```
library(SIMLR)
data(BuettnerFlorian)
```

The external R package igraph is required for the computation of the normalized mutual information to assess the results of the clustering.

```
library(igraph)
```

We now run SIMLR as an example on an input dataset from Buettner, Florian, et al. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells." Nature biotechnology 33.2 (2015): 155-160. For this dataset we have a ground true of 3 cell populations, i.e., clusters.

```
set.seed(11111)
example = SIMLR(X = BuettnerFlorian$in_X, c = BuettnerFlorian$n_clust, cores.ratio = 0)
```

```
## Computing the multiple Kernels.
## Performing network diffiusion.
## Iteration:  1
## Iteration:  2
## Iteration:  3
## Iteration:  4
## Iteration:  5
## Iteration:  6
## Iteration:  7
## Iteration:  8
## Iteration:  9
## Iteration:  10
## Iteration:  11
## Performing t-SNE.
## Epoch: Iteration # 100  error is:  0.1140084
## Epoch: Iteration # 200  error is:  0.06181848
## Epoch: Iteration # 300  error is:  0.05932105
## Epoch: Iteration # 400  error is:  0.05923113
## Epoch: Iteration # 500  error is:  0.05914954
## Epoch: Iteration # 600  error is:  0.05907505
## Epoch: Iteration # 700  error is:  0.05900841
## Epoch: Iteration # 800  error is:  0.05894696
## Epoch: Iteration # 900  error is:  0.05889082
```

```
## Epoch: Iteration # 1000  error is:  0.0588387
## Performing Kmeans.
## Performing t-SNE.
## Epoch: Iteration # 100  error is:  10.36092
## Epoch: Iteration # 200  error is:  1.167142
## Epoch: Iteration # 300  error is:  0.8673864
## Epoch: Iteration # 400  error is:  1.463917
## Epoch: Iteration # 500  error is:  0.7682029
## Epoch: Iteration # 600  error is:  0.6115922
## Epoch: Iteration # 700  error is:  0.4790367
## Epoch: Iteration # 800  error is:  1.289853
## Epoch: Iteration # 900  error is:  0.793667
## Epoch: Iteration # 1000  error is:  0.6030175
```

We now compute the normalized mutual information between the inferred clusters by SIMLR and the true one. This measure with values in [0,1], allows us to assess the performance of the clustering with higher values reflecting better performance.

```
nmi_1 = compare(BuettnerFlorian$true_labs[,1], example$y$cluster, method="nmi")
print(nmi_1)
```

```
## [1] 0.888298
```

As a further understanding of the results, we now visualize the cell populations in a plot.

```
plot(example$ydata,
     col = c(topo.colors(BuettnerFlorian$n_clust))[BuettnerFlorian$true_labs[,1]],
     xlab = "SIMLR component 1",
     ylab = "SIMLR component 2",
     pch = 20,
     main="SIMILR 2D visualization for Test_1_mECS")
```

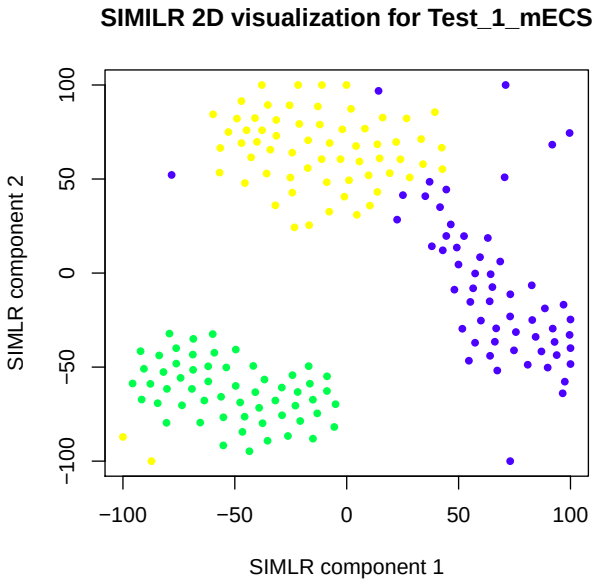

Figure 1: Visualization of the 3 cell populations retrieved by SIMLR.

SIMRL supports SCESet objects. We now create an example object and then run SIMLR on it.

```
library(scran)
ncells = 100
ngenes = 50
```

```
mu <- 2^runif(ngenes, 3, 10)
gene.counts <- matrix(rnbinom(ngenes*ncells, mu=mu, size=2), nrow=ngenes)
rownames(gene.counts) = paste0("X", seq_len(ngenes))
sce = newSCESet(countData=data.frame(gene.counts))
output = SIMLR(X = sce, c = 8, cores.ratio = 0)
```

```
## X is and SCESet, converting to input matrix.
## Computing the multiple Kernels.
## Performing network diffiusion.
## Iteration:  1
## Iteration:  2
## Iteration:  3
## Iteration:  4
## Iteration:  5
## Iteration:  6
## Iteration:  7
## Iteration:  8
## Iteration:  9
## Iteration:  10
## Performing t-SNE.
## Epoch: Iteration # 100  error is:  0.4468889
## Epoch: Iteration # 200  error is:  0.4825569
## Epoch: Iteration # 300  error is:  0.2033292
## Epoch: Iteration # 400  error is:  0.1401524
## Epoch: Iteration # 500  error is:  0.03213967
## Epoch: Iteration # 600  error is:  0.04962836
## Epoch: Iteration # 700  error is:  0.03475453
## Epoch: Iteration # 800  error is:  0.1202234
## Epoch: Iteration # 900  error is:  0.09088354
## Epoch: Iteration # 1000  error is:  0.04854141
## Performing Kmeans.
## Performing t-SNE.
## Epoch: Iteration # 100  error is:  24.90758
## Epoch: Iteration # 200  error is:  2.094652
## Epoch: Iteration # 300  error is:  3.645437
## Epoch: Iteration # 400  error is:  3.072863
## Epoch: Iteration # 500  error is:  1.986776
## Epoch: Iteration # 600  error is:  1.575832
## Epoch: Iteration # 700  error is:  3.012081
## Epoch: Iteration # 800  error is:  2.325609
## Epoch: Iteration # 900  error is:  1.769769
## Epoch: Iteration # 1000  error is:  1.268874
```

We finally run SIMLR feature ranking on the same inputs to get a rank of the key genes with the related pvalues.

```
ranks = SIMLR_Feature_Ranking(A=BuettnerFlorian$results$S,X=BuettnerFlorian$in_X)
```

```
head(ranks$pval)
```

```
## [1] 4.928188e-131  7.017070e-91  4.491575e-80  2.145926e-77  7.135361e-76  3.949742e-70
```

```
head(ranks$aggR)
```

```
## [1] 5701 1689 7549   57 2653 7595
```

4 sessionInfo()

```
toLatex(sessionInfo())
```

- R version 3.3.1 (2016-06-21), `x86_64-apple-darwin13.4.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: Biobase 2.34.0, BiocGenerics 0.20.0, BiocParallel 1.8.1, SIMLR 1.0.1, ggplot2 2.1.0, igraph 1.0.1, knitr 1.14, scater 1.2.0, scran 1.2.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.36.0, BiocStyle 2.2.0, DBI 0.5-1, IRanges 2.8.1, Matrix 1.2-7.1, R6 2.2.0, RCurl 1.95-4.8, RSQLite 1.0.0, Rcpp 0.12.7, S4Vectors 0.12.0, XML 3.98-1.4, assertthat 0.1, beeswarm 0.2.3, biomaRt 2.30.0, bitops 1.0-6, chron 2.3-47, colorspace 1.2-7, data.table 1.9.6, digest 0.6.10, dplyr 0.5.0, dynamicTreeCut 1.63-1, edgeR 3.16.2, evaluate 0.10, formatR 1.4, ggbeeswarm 0.5.0, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, highr 0.6, htmltools 0.3.5, httpuv 1.3.3, lattice 0.20-34, limma 3.30.2, locfit 1.5-9.1, magrittr 1.5, matrixStats 0.51.0, mime 0.5, munsell 0.4.3, plyr 1.8.4, reshape2 1.4.2, rhdf5 2.18.0, rjson 0.2.15, scales 0.4.0, shiny 0.14.2, shinydashboard 0.5.3, statmod 1.4.26, stats4 3.3.1, stringi 1.1.2, stringr 1.1.0, tibble 1.2, tools 3.3.1, tximport 1.2.0, vipor 0.4.4, viridis 0.3.4, xtable 1.8-2, zlibbioc 1.20.0, zoo 1.7-13