scAlign Tutorial

Nelson Johansen, Gerald Quon

2019-10-29

Introduction

This tutorial provides a guided alignment for two groups of cells from cellbench RNA mixture experiments. In this tutorial we demonstrate the unsupervised alignment strategy of `scAlign` described in Johansen et al, 2018 along with typical analysis utilizing the aligned dataset, and show how `scAlign` can identify and match cell types across platforms without using the labels as input.

Alignment goals

The following is a walkthrough of a typical alignment problem for `scAlign` and has been designed to provide an overview of data preprocessing, alignment and finally analysis in our joint embedding space. Here, our primary goals include:

1. Learning a low-dimensional cell state space in which cells group by function and type, regardless of condition (platform).
2. Accurately labeling old cells with cell cycle and cell type information using only the young cell annotations.

Installation

```
## Install scAlign
install.packages('devtools')
devtools::install_github(repo = 'quon-titative-biology/scAlign')
library(scAlign)

## Install Tensorflow
library(tensorflow)
install_tensorflow(version = "gpu") ## Removing version will install CPU version of Tensorflow
```

```
Guide to installing python and tensorflow on different operating systems.

On Windows:
Download Python 3.6.8. Note, newer versions of Python (e.g. 3.7) cannot use TensorFlow at this time.
Make sure pip is included in the installation.

On Ubuntu:
sudo apt update
sudo apt install python3-dev python3-pip

On MacOS (homebrew):
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
export PATH="/usr/local/bin:/usr/local/sbin:$PATH"
brew update
brew install python # Python 3
```

```
Further details at: https://www.tensorflow.org/install
```

Data setup

The gene count matrices used for this tutorial are hosted on the cellbench github: here.

First, we load in the normalized cellbench data. The data was normalized following the procedures defined on the cellbench github.

```
library(scAlign)
library(SingleCellExperiment)
library(ggplot2)

## Load in cellbench data
data("cellbench", package = "scAlign", envir = environment())

## Extract RNA mixture cell types
mix.types = unlist(lapply(strsplit(colnames(cellbench), "-"), "[[", 2))

## Extract Platform
batch = c(rep("CEL", length(which(!grepl("sortseq", colnames(cellbench)) == TRUE))),
          rep("SORT", length(which(grepl("sortseq", colnames(cellbench)) == TRUE))))
```

scAlign setup

The general design of scAlign's makes it agnostic to the input RNA-seq data representation. Thus, the input data can either be gene-level counts, transformations of those gene level counts or a preliminary step of dimensionality reduction such as canonical correlates or principal component scores. Here we create the `scAlign` object from the previously normalized cellbench data and perform CCA on the unaligned data.

```
## Create SCE objects to pass into scAlignCreateObject
youngMouseSCE <- SingleCellExperiment(
    assays = list(scale.data = cellbench[,batch=='CEL'])
)

oldMouseSCE <- SingleCellExperiment(
    assays = list(scale.data = cellbench[,batch=='SORT'])
)

## Build the scAlign class object and compute PCs
scAlignCB = scAlignCreateObject(sce.objects = list("CEL"=youngMouseSCE,
                                                   "SORT"=oldMouseSCE),
                                labels = list(mix.types[batch=='CEL'],
                                              mix.types[batch=='SORT']),
                                data.use="scale.data",
                                pca.reduce = FALSE,
                                cca.reduce = TRUE,
                                ccs.compute = 5,
                                project.name = "scAlign_cellbench")
```

```
## [1] "Computing CCA using Seurat."
```

```
## Centering and scaling data matrix
## Centering and scaling data matrix
```

```
## Running CCA

## Merging objects

## Warning: The following arguments are not used: scale.data
```

Alignment of cellbench RNAmixture

Now we align the cell populations from both protocols. `scAlign` returns a low-dimensional joint embedding space where the effect of platform is removed allowing us to use the complete dataset for downstream analyses such as clustering or differential expression.

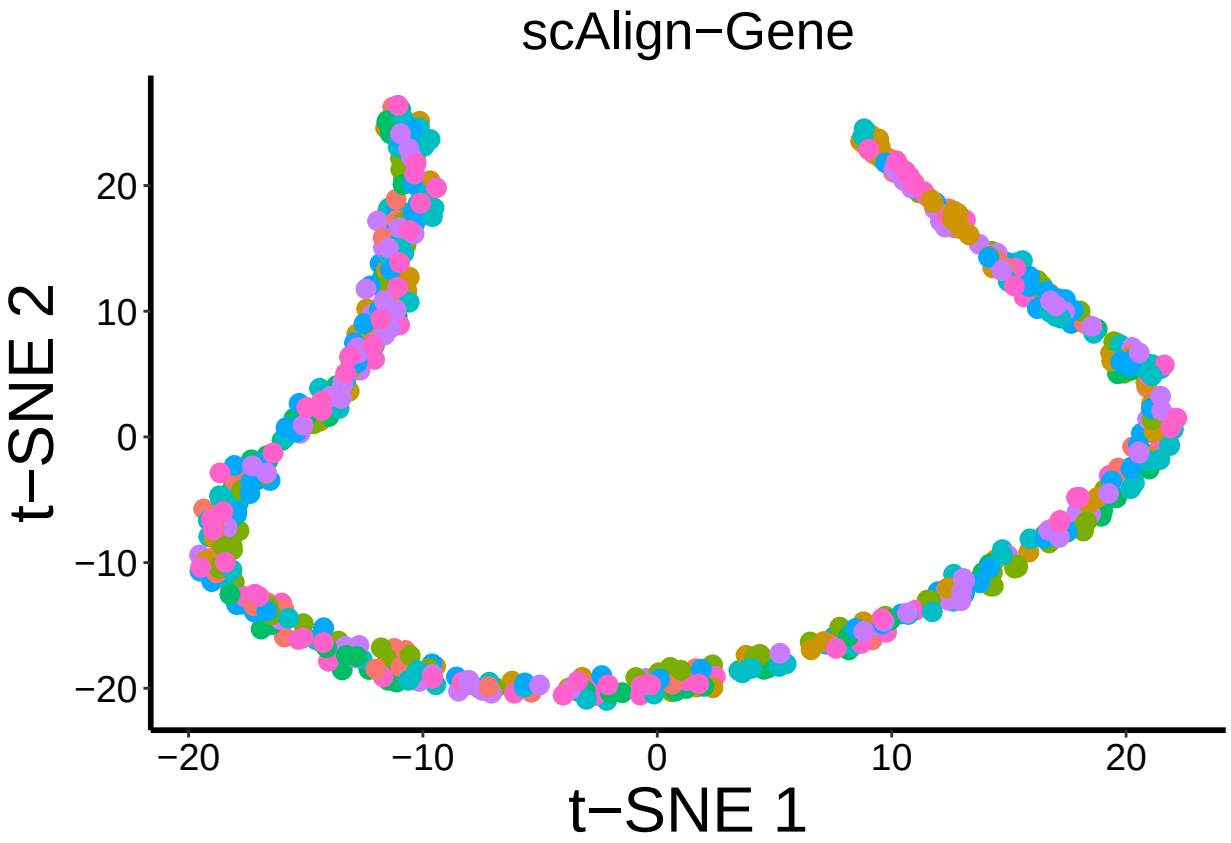
```
## Run scAlign with all_genes
scAlignCB = scAlign(scAlignCB,
                    options=scAlignOptions(steps=1000,
                                           log.every=1000,
                                           norm=TRUE,
                                           early.stop=TRUE),
                    encoder.data="scale.data",
                    supervised='none',
                    run.encoder=TRUE,
                    run.decoder=FALSE,
                    log.dir=file.path('~/models_temp','gene_input'),
                    device="CPU")
```

```
## [1] "============== Step 1/3: Encoder training ==============="
## [1] "Graph construction"
## [1] "Adding source walker loss"
## [1] "Adding target walker loss"
## [1] "Done random initialization"
## [1] "Step: 1    Loss: 10.0724"
## [1] "Step: 100    Loss: 9.0983"
## [1] "Step: 200    Loss: 9.0137"
## [1] "Step: 300    Loss: 8.7419"
## [1] "Step: 400    Loss: 8.5922"
## [1] "Step: 500    Loss: 8.6516"
## [1] "Step: 600    Loss: 8.576"
## [1] "Step: 700    Loss: 8.4378"
## [1] "Step: 800    Loss: 8.6551"
## [1] "Step: 900    Loss: 8.7798"
## [1] "Step: 1000    Loss: 8.6675"
## [1] "============== Alignment Complete =============="
```

```
# ## Additional run of scAlign with CCA
# scAlignCB = scAlign(scAlignCB,
#                     options=scAlignOptions(steps=1000,
#                                            log.every=1000,
#                                            norm=TRUE,
#                                            early.stop=TRUE),
#                     encoder.data="CCA",
#                     supervised='none',
#                     run.encoder=TRUE,
#                     run.decoder=FALSE,
#                     log.dir=file.path('~/models','cca_input'),
#                     device="CPU")
```

```
## Plot aligned data in tSNE space, when the data was processed in three different ways:
##   1) either using the original gene inputs,
##   2) after CCA dimensionality reduction for preprocessing.
##   Cells here are colored by input labels

set.seed(5678)
gene_plot = PlotTSNE(scAlignCB,
                     "ALIGNED-GENE",
                     title="scAlign-Gene",
                     perplexity=30)
## Show plot
gene_plot
```




```
# cca_plot = PlotTSNE(scAlignCB,
#                     "ALIGNED-CCA",
#                     title="scAlign-CCA",
#                     perplexity=30)
#
# multi_plot_labels = grid.arrange(gene_plot, cca_plot, nrow = 1)
```

Session Info

```
sessionInfo()
```

```
## R Under development (unstable) (2019-10-24 r77329)
```

```
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.3 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.11-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.11-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] scAlign_1.3.0               PMA_1.1
##  [3] FNN_1.1.3                   ggplot2_3.2.1
##  [5] Rtsne_0.15                  irlba_2.3.3
##  [7] Matrix_1.2-17               purrr_0.3.3
##  [9] tensorflow_2.0.0            Seurat_3.1.1
## [11] SingleCellExperiment_1.9.0  SummarizedExperiment_1.17.0
## [13] DelayedArray_0.13.0         BiocParallel_1.21.0
## [15] matrixStats_0.55.0          Biobase_2.47.0
## [17] GenomicRanges_1.39.0        GenomeInfoDb_1.23.0
## [19] IRanges_2.21.0              S4Vectors_0.25.0
## [21] BiocGenerics_0.33.0
##
## loaded via a namespace (and not attached):
##   [1] colorspace_1.4-1       ggridges_0.5.1         XVector_0.27.0
##   [4] base64enc_0.1-3        leiden_0.3.1           listenv_0.7.0
##   [7] npsurv_0.4-0           ggrepel_0.8.1          codetools_0.2-16
##  [10] splines_4.0.0          R.methodsS3_1.7.1      impute_1.61.0
##  [13] lsei_1.2-0             knitr_1.25             config_0.3
##  [16] zeallot_0.1.0          jsonlite_1.6           ica_1.0-2
##  [19] cluster_2.1.0          tfruns_1.4             png_0.1-7
##  [22] R.oo_1.22.0            uwot_0.1.4             sctransform_0.2.0
##  [25] compiler_4.0.0         httr_1.4.1             backports_1.1.5
##  [28] assertthat_0.2.1       lazyeval_0.2.2         htmltools_0.4.0
##  [31] tools_4.0.0            rsvd_1.0.2             igraph_1.2.4.1
##  [34] gtable_0.3.0           glue_1.3.1             GenomeInfoDbData_1.2.2
##  [37] RANN_2.6.1             reshape2_1.4.3         dplyr_0.8.3
##  [40] Rcpp_1.0.2             vctrs_0.2.0            gdata_2.18.0
##  [43] ape_5.3                nlme_3.1-141           gbRd_0.4-11
##  [46] lmtest_0.9-37          xfun_0.10              stringr_1.4.0
##  [49] globals_0.12.4         lifecycle_0.1.0        gtools_3.8.1
##  [52] future_1.14.0          zlibbioc_1.33.0        MASS_7.3-51.4
##  [55] zoo_1.8-6              scales_1.0.0           RColorBrewer_1.1-2
##  [58] yaml_2.2.0             reticulate_1.13        pbapply_1.4-2
##  [61] gridExtra_2.3          stringi_1.4.3          caTools_1.17.1.2
```

```
##  [64] bibtex_0.4.2           Rdpack_0.11-0          SDMTools_1.1-221.1
##  [67] rlang_0.4.1            pkgconfig_2.0.3        bitops_1.0-6
##  [70] evaluate_0.14          lattice_0.20-38        ROCR_1.0-7
##  [73] labeling_0.3           htmlwidgets_1.5.1      cowplot_1.0.0
##  [76] tidyselect_0.2.5       RcppAnnoy_0.0.13       plyr_1.8.4
##  [79] magrittr_1.5           R6_2.4.0               gplots_3.0.1.1
##  [82] withr_2.1.2            whisker_0.4            pillar_1.4.2
##  [85] fitdistrplus_1.0-14    survival_2.44-1.1      RCurl_1.95-4.12
##  [88] tibble_2.1.3           future.apply_1.3.0     tsne_0.1-3
##  [91] crayon_1.3.4           KernSmooth_2.23-16     plotly_4.9.0
##  [94] rmarkdown_1.16         grid_4.0.0             data.table_1.12.6
##  [97] metap_1.1              digest_0.6.22          tidyr_1.0.0
## [100] R.utils_2.9.0          RcppParallel_4.4.4     munsell_0.5.0
## [103] viridisLite_0.3.0
```