Reference Manual: Deconvolution with DWLS

Dampened weighted least squares (DWLS) is an estimation method for gene expression deconvolution, in which the cell-type composition of a bulk RNA-seq data set is computationally inferred. This method corrects common biases towards cell types that are characterized by highly expressed genes and/or are highly prevalent, to provide accurate detection across diverse cell types. To begin, the user must input a bulk RNA-seq data set, along with a labeled representative single-cell data set that will serve to generate cell-type-specific gene expression profiles. Ideally, the single-cell data set will contain cells from all cell types that may be found in the bulk data. DWLS will return the cell-type composition of the bulk data.

DWLS is written in the R programming language. To apply DWLS to any RNA-seq data set, first download the “DWLS” project. This folder contains the core functions, under the “Deconvolution\_functions.R” script, as well as example bulk and single-cell data sets from various sources.

Follow the steps below to run DWLS on an example data set (control mouse intestinal stem cells). This template can be easily customized for your own application:

## Paths

Create a new folder for your project. This is where all results and data will be placed. Change the “workdir” path to match this folder’s path.

#paths  
workdir = "/Users/Daphne/Documents/Yuan/Deconvolution/Deconvolution\_Github/ISC" # where you place the data and results  
setwd(workdir)

Create “data” and “results” folders within this project.

dir.create(file.path(workdir, "data"), showWarnings = FALSE) #folder to store data  
dir.create(file.path(workdir, "results"), showWarnings = FALSE) #folder to save results

Place your data into the “data” folder. In this example, we will use an intestinal stem cell bulk data set, along with a reference intestinal stem cell single-cell data set, which has been labeled by cell type.

## Load functions and associated packages

source("../Deconvolution\_functions.R")

## Load data

load("data/dataSC.RData") #load single-cell data  
load("data/dataBulk.RData") #read in bulk data for WT1 (control condition #1)  
load("data/labels.RData") #read in single-cell labels from clustering

## Build signature from single-cell data

Signature<-**buildSignatureMatrixMAST**(scdata=dataSC,id=labels,path="results",diff.cutoff=0.5,pval.cutoff=0.01)

## Deconvolution

#trim signature and bulk data to contain the same differentially expressed genes  
tr<-trimData(Signature,dataBulk)  
#estimate using dampened weighted least squares  
solDWLS<-solveDampenedWLS(tr$sig,tr$bulk)

## CycISC NonCycISC TA Goblet Paneth Ent PreEnt   
## 0.33397 0.26110 0.34681 0.00956 0.01113 0.00882 0.00000   
## Tuft EE   
## 0.02861 0.00000

We can also try running the deconvolution using other estimation methods.

#try running with other estimation methods  
solSVR<-solveSVR(tr$sig,tr$bulk) #nu-SVR

## CycISC NonCycISC TA Goblet Paneth Ent PreEnt   
## 0.00000 0.36151 0.62722 0.01126 0.00000 0.00000 0.00000   
## Tuft EE   
## 0.00000 0.00000

solOLS<-solveOLS(tr$sig,tr$bulk) #constrained ordinary least squares

## CycISC NonCycISC TA Goblet Paneth Ent PreEnt   
## 0.00000 0.74170 0.25541 0.00000 0.00289 0.00000 0.00000   
## Tuft EE   
## 0.00000 0.00000