ADVOCATE

Pasquale Laise, Jing He and Andrea Califano
Department of Systems Biology, Columbia University, New York, USA

Overview

ADVOCATE is a machine learning based algorithm to combine weak evidence derived from the expression of individual genes into a model that provides an optimal estimate of the compartment-specific composition of a heterogeneous bulk tissue. The approach assumes that the gene expression probability density function (PDF) of the bulk is modeled as a mixture of two distinct PDFs, representing the stromal and epithelial compartments, with the optional inclusion of a third "residual" or "unspecified" compartment. The latter can be used to model either infiltration by an unknown tissue type in a specific sample or the contribution of a platform-specific bias. Inclusion of the additional compartment allows effectively addressing variability that is not statistically independent across all genes (i.e., uncorrelated noise), thus improving prediction of epithelial and stromal specific compartment representation. In this vignette, we show how to install and run ADVOCATE.

Installation of ADVOCATE package

ADVOCATE requires the R statistical software (https://www.r-project.org/) and the following dependencies: "foreach", "doMC", "iterators", "parallel", "mclust".

To install ADVOCATE open your shell and enter:

```
R CMD INSTALL ~/ADVOCATE_0.1.0.tar.gz
```

To install ADVOCATE dependencies, start R and enter:

```
install.packages("foreach")
install.packages("doMC")
install.packages('iterators')
install.packages('parallel')
install.packages('mclust')
```

Getting started

As first step, we need to load the ADVOCATE environment with:

```
library (ADVOCATE)
```

Run ADVOCATE

Here we show the basic steps to infer the percentage and virtual expression of epithelium and stroma compartments in PDA bulk RNAseq using ADVOCATE. After we loaded the ADVOCATE environment, we need to prepare the input data.

Input data

As an input, ADVOCATE expects the two following files:

1. The input_advocate_train file, containing the following data objects:

- a character vector containing a list of DEGs ("deg")
- the expression matrix with all the samples used for the training ("expmat")
- a numeric vector containing the fold change of each gene ("fc")
- a numeric vector containing the pvalue of each gene("pval")
- a data frame containing sample information ("sampleInfo"“). This file must be a 3 columns data frame where the 1st column is the names of the samples; the 2nd column the compartment.id (e.g."E" for epithelium, "S"" for stroma); the 3rd columns is numeric and indicates paired samples (e.g. paired samples between the two compartments will have the same number)

If not available, the input_advocate_train file can be readily generated by the "hedgeBeta_deg" function of the ADVOCATE package. Specifically, the "hedgeBeta_deg" uses raw LCM samples data (rowcounts) from epithelium and stroma samples to calculate differentially expressed genes and the associated weight metrics for ADVOCATE model training. ## Example: how to run the hedgeBeta_deg function ## hedgeBeta_deg(rawcounts, idxE, idxS, outfile = "input_advocate_train.rda") ## where idxE indicates Epithelium samples and idxS indicates Stroma samples.

2. A matrix containing normalized expression of all the bulk samples.

Predict the percentage and virtual expression of stroma and epithelium in PDA bulk RNAseq (2–component model). Once the input files have been assembled, we may compute the stroma and epithelium fraction and their virtual expression from PDA bulk RNAseq data, by performing the following three steps:

Step1: load the input data

```
trainDataADVOCATE="~/trainDataADVOCATE.rda"
bulkTCGA= "~/ bulkTCGA.rda"
load(trainDataADVOCATE)
load(bulkTCGA)
```

Step2: compute the fraction of epithelium and stroma compartments

```
res.prop<-predict_bulk(trainfile, bulkexp)
```

The predict_bulk function uses a probabilistic model to calculate the fraction of epithelium and compartment using gene expression.

Step3: Predict virtual gene expression

```
res.vexp = calCellTypeExpression(lcmexp, deg, fc, pval,sampleInfo, exp_bulkTCGA,
                                 res.prop, method = 'lcm')
```

The calCellTypeExpression function uses pre-generated LCM gene expression data and the output of "predict_bulk" function to infer virtual gene expression of epithelium and stroma compartments.

ADVOCATE with 3-component model

In the 2-component model, the bulk expression of each gene is represented exactly as the sum of its epithelial and stromal compartment expression. However, there may be cases where the expression of a gene may be outside the allowable range for the bulk probability density function (PDF), as represented by the weighed sum of the epithelial and stromal PDFs, at a given threshold. In such cases, we can only assume that such an expression is caused either by the presence of unknown cellular compartments in the specific sample or by technical differences (e.g., platform-speci c bias). To address this issue, we introduce a third "residual" or

"unspecified" compartment. In order to compute the percentage and virtual expression of 3 compartments we can use the following code:

```
res.prop3 = predict_bulk_3comp(trainDataADVOCATE, bulkTCGA, epsilon = 0.001)

res.vexp3 = calCellTypeExpression_3comp(lcmexp, deg, fc, pval,sampleInfo,
                                        bulkTCGA, res.prop, method = 'lcm')
```

Reference:

Jing He, H. Carlo Maurer, Sam R. Holmstrom, Pasquale Laise,Tao Su, Aqeel Ahmed, Hanina Hibshoosh, John A. Chabot, Paul E. Oberstein, Antonia R. Sepulveda, Jeanine M. Genkinger, Jiapeng Zhang, Alina C. Iuga, Andrea Califano, Mukesh Bansal, and Kenneth P. Olive. "Transcriptional deconvolution reveals consistent functional subtypes of pancreatic cancer epithelium and stroma" (manuscript under revision)