Running ICeDT on Hugo et al. Data

Chong Jin

3/14/2019

Packages needed

To run this package, make sure you have these packages installed:

```
library(nnls)
library(quantreg)
library(hqreg)
library(gplots)
library(org.Hs.eg.db)
library(alabama)
library(EPIC)
library(clinfun)
library(ICeDT)
```

For convenience, a clone of `EPIC_1.1.2` is stored inside the folder.

Preprocessing

In the code, the purpose of Sections 2 - 4 is mainly to consolidate gene names and obtain TPMs from gene counts.

Gene set and weights

```
Geneset = "Revised"
Weights = "Revised"
rescale = TRUE
```

We recommend to use rescaled data, and weights based on rescaled data (hence `Weights = "Revised"`). Here `Geneset = "Revised"` means that signature genes will include EPIC Genes, LM22 Genes, MCP-Counter genes, which total number is 473. If `Weights = "Original"`, only 98 EPIC Genes will count as signature genes.

Running ICeDT

```
# Using ICeDT package
# no weight
fitnw = ICeDT::ICeDT(Y=bulk, Z=refProfiles, tumorPurity=NULL, refVar=NULL,
                     rhoInit=NULL, maxIter_prop = 500, maxIter_PP = 250, rhoConverge = 1e-3)
# with weight
fitw0 = ICeDT::ICeDT(Y=bulk, Z=refProfiles, tumorPurity=NULL, refVar=refVar,
                     rhoInit=NULL, maxIter_prop = 500, maxIter_PP = 250, rhoConverge = 1e-3)
```

The code block illustrates a typical way to use ICeDT. Running the code will take a couple of minutes. Since the code uses `auglag` to do augmented Lagranian method, the program may prompt warning messages, which is no indication of actual failure of ICeDT.

Analysis of results

Section 7 produces plots of inferred cellular proportions and Section 8 gives a summary of how consistent (as opposed to abberant) the sample-gene pairs are.

For each model w/ weight and no weight, we divide sample-gene pairs into three equally numbered groups using cutoffs of probability being consistent. For each group, we have a scatterplot of observed gene expression and expected gene expression from the model in $\log(1 \times 10^{-5} + \mathrm{TPM})$ saved as the `"scatterplotConsistent"` figures. The plot agrees with our assumption that among more consistent sample-gene pairs, the model-predicted gene expression is aligned more closely with observed gene expression.