Running ICeDT on Hugo et al. Data

Douglas Roy Wilson Jr., Chong Jin

3/21/2019

This document explains how to run ICeDT on Hugo et al. data using `./program/running_ICeDT_on_Hugo_data.R`.

Packages needed

To run the script `running_ICeDT_on_Hugo_data.R`, 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 `./programs/EPIC-master`.

Preprocessing

The purpose of Sections 2 - 4 (in the code being referred to) is to consolidate gene names and obtain TPMs from gene counts.

Gene set and weights

```
Geneset = "Original"
```

Here `Weights = "Original"` says only 98 EPIC Genes will count as signature genes. Another option, `Geneset = "Revised"`, means that signature genes will include EPIC Genes, LM22 Genes and MCP-Counter genes, whose total number is 473. We provide separate `.Rout` files to illustrate the result of running the script with any of these two options.

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.

Illustration of results

Section 8 of the code produces plots of inferred cellular proportions and Section 9 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 by their model-based probability of being consistent into 3-quantiles (cutoffs are listed below). For each group, a scatterplot of observed gene expression and expected gene expression from the model in $\log(1 \times 10^{-5} + \text{TPM})$ is saved as files in `./figures` folder starting with `"probConsistent"`. 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.

The cutoffs used are:

- If `Geneset=="Revised"`:
  - Consistent probability cutoffs for model w/o weight: 0.558, 0.680
  - Consistent probability cutoffs for model w/ weight: 0.946, 0.983
- If Geneset==“Original”:
  - Consistent probability cutoffs for model w/o weight: 0.509, 0.549
  - Consistent probability cutoffs for model w/ weight: 0.512, 0.538

Folder structure

- ./data/
  - ./TRef_Data/
    - Contains Datasets generated to construct variance weights.
    - Validated Against TRef.
    - TRef_purData.RData:
      Contains the purified reference sample data utilized by EPIC, one matrix per cell type with expressions across 23681 genes and 7 cell types. Data is stored as:
      - XXXdat_r:
        Each XXX will define a different cell type (B=B-cells, CAF = Cancer Associated fibroblasts, CD4 = CD4+ T-cells, CD8 = CD8+ T-cells, E = Endothelials, M = Macrophages, NK = Natural Killer cells.
  - ./FlowCytometry/
    - LM22.txt: gene signature matrix provided in CIBERSORT
    - Contains Datasets generated by Dr. Wei Sun for CIBERSORT Analysis.
    - Corrected Cell Type labels and cleaned files.
  - SunMelanoma_Exp_Full_032118.txt
    - Contains the TPM normalized RNA-seq data for the PD1 immunotherapy trial.
  - Patient_info.txt
    - Contains the Patient information for the PD1 immunotherapy trial.
  - FPKM_and_counts_filtered.RData
    - Contains the gene-level read counts of the PD1 immunotherapy data.
  - Gene_lengths_v27.rds + gencode.v27.genes.txt
    - Contains the gene lengths computed by Dr. Wei Sun for the PD1 immunotherapy data.
  - CIBERSORT.Output_SunMelExt_TRefNQNorm.csv:
    - Contains the CIBERSORT results, without quantile normalization, for the PD1 immunotherapy data using TRef reference matrix from EPIC.
  - CIBERSORT.Output_SunMelExt_TRefNQNorm.csv:
    - Contains the CIBERSORT results, without quantile normalization, for the PD1 immunotherapy data using LM22 reference matrix.
  - Melanoma_Exp_Full_orig.txt
    - Contains the TPM normalized RNA-seq data for the EPIC Melanoma Validation Dataset.
  - CIBERSORT.Output_Rescale_TRef:

    - Contains the CIBERSORT output for the EPIC melanoma validation dataset using the rescaled data and TRef.
  - CIBERSORT.Output_FULL_lm22.csv:
    - Contains the CIBERSORT output for the EPIC melanoma validation dataset when fit using the entire TPM matrix and LM22 reference matrix.
- . /programs/
  - running_ICeDT_on_Hugo_data.R
    - Contains the code necessary to fit ICeDT models using weight and without weight, and produce the figures.
  - ./EPIC_TitrationData/
    - Code necessary to fit the EPIC melanoma Validation Dataset.
  - ./EPIC-Master/
    - Contains the R library for the EPIC functions.
  - ./FlowCytometry/
    - Code necessary to perform the fits to the Flow Cytometry Validation data.
  - ./Simulation_Fits/
    - Saved Output from Simulation Fitting.
  - EPIC_Extract.R
    - Contains the code necessary to save EPIC's optimization pieces (design matrix, outcome data, etc) and output it to a list object. It contains two functions [EPIC_Extract() and scaleCounts()].
      - EPIC_Extract():

      EPIC_Extract runs the EPIC model fit but has been edited to output the utilized reference expression matrix and mixture expression profiles so that ICeDT and EPIC can be fit to the same data using the same data. Returns the EPIC fit and edited to provided:
      $bulk: The mixture expression matrix (nG by nS)
      $ref : The utilized version of the TRef matrix
      - scaleCounts():

      This function was designed by the EPIC authors (Racle et al) to renormalize mixture and reference data across common genes. It is presented unedited but must be utilized outside the EPIC package as the function was not made available to users.
- ./figures/
  - Contains the requested figures from EPIC and ICeD-T and these figures are labeled.