citccmst : CIT Colon Cancer Molecular Subtypes Prediction

<CITR@ligue-cancer.net>

January 10, 2014

Cartes d'Identité des Tumeurs research program - Ligue Nationale Contre le Cancer, Paris, France
(http://cit.ligue-cancer.net)

Overview

The work by Marisa et al[1] exploited a large multicenter and extensively characterized series of CC to establish a robust molecular classification based on transcriptome data. This package implements the approach used in the article to assign an expression profile to one of the 6 colon cancer molecular subtypes.

Contents

1 The CIT Subtypes Prediction Approach

To assign a sample to one of the 6 CIT colon cancer molecular subtypes from its expression profile, a centroid based approach is used. The centroids of the 6 CIT colon cancer molecular subtypes were defined on 443 samples and 57 probe sets corresponding to unique gene symbols. These data are in the object *citccmst* and is automatically loaded with the package :

```
> library(citccmst)
> summary(citccmst)
```

```
           Length Class      Mode
data       443    data.frame list
data.cl    443    -none-     character
data.annot  41    data.frame list
```

citccmst contains the following objects :

data the CIT discovery set normalized expression data matrix

data.cl the CIT discovery set colon cancer molecular subtypes

data.annot the CIT discovery set probe sets annotations provided by NetAffx (version na31) in order to map samples from other plateforms than Affymetrix

To assign a new sample to a CIT subtype (***cit.assignCcmst*** function) the following steps are performed:

1. mapping the probes from the new expression dataset to the 57 discrimating probe sets used in the CIT centroids (or to the 57 corresponding gene symbols, when the microarray platform of the new dataset is not Affymetrix HG U133 plus 2.0)

2. averaging expression measures per gene symbol both in the new dataset and in the CIT coreset, if step 1 is based on gene symbols. In any case, the external data and the CIT discovery set data are reduced to discrimating probes/genes measured in both datasets.

3. recomputing the CIT centroids of the 6 CIT subtypes using the CIT original data resulting from step 2

4. computing distance of the new sample(s) to those 6 centroids

5. assigning each sample to the subtype corresponding to the closest centroid
   In some case, (1) a sample can be close to several centroids or (2) the closest centroid can be too far to confidently assign the sample to a given subtype. In the first case, the sample will be considered as a *"mixed"* sample and in the second case as an outlier. In both cases, those samples may be classified as uncertain and removed from the analysis.

The output of ***cit.assignCcmst*** is a dataframe with n rows (n= the number of samples in the new dataset) and 4 columns :

citccmst the subtype of the closest centroid for each sample

citccmst.mixed the subtypes of the closest centroids for each sample (for "mixed" samples every subtypes are given)

citccmst.core subtypes of the closest centroid for "core" samples only ("outlier" and "mixed" samples are set to NA)

citccmst.confidence confidence annotation CORE, OUTLIER or MIXED

The cut-off values to define mixed and outlier samples are automatically computed but can be set manually (cf help(cit.assignCcmst)).

Only an expression data matrix/data.frame, with ids as rownames, is required (cf section2). Affymetrix external expression data should be normalized as CIT discovery set data, i.e. by RMA method (justRma function in *affy* R package with default parameters). No row centering is required as it will be computed during the assignment process. The approach has been defined for Affymetrix HG U133Plus2 chip expression data, therefore the prediction given for other platforms is expected to be less reliable.

2 Example

Here is an example on the expression profiles of colon tumors used in the validation set of Marisa et al[1].

```
> load(list.files(system.file("extdata", package="citccmst"), full.names=TRUE)) # load citva
> citvalid.exp.annot <- data.frame(id=rownames(citvalid.exp.norm), stringsAsFactors=FALSE, r
```

```
> citvalid.citccmst <- cit.assignCcmst(   data=citvalid.exp.norm,
+                                         data.annot=citvalid.exp.annot,
+                                         data.colId="id",
+                                         data.colMap="id" ,
+                                         citccmst.colMap="Probe.Set.ID",
+                                         dist.method="dqda",
+                                         plot=T
+                 )
```

```
Mapping - 57/57 original probes.
```

```
> str(citvalid.citccmst)
```

```
'data.frame':        123 obs. of  4 variables:
 $ citccmst           : chr  "C3" "C4" "C2" "C3" ...
 $ citccmst.mixed     : chr  "C3" "C4" "C2" "C3" ...
 $ citccmst.core      : chr  "C3" "C4" "C2" "C3" ...
 $ citccmst.confidence: chr  "CORE" "CORE" "CORE" "CORE" ...
 - attr(*, "distmethod")= chr "dqda"
 - attr(*, "nb.mapped.probes")= chr "57/57"
 - attr(*, "citmc")= chr "65.7%"
 - attr(*, "scoreGroup")= Named num  164 150 204 164 204 ...
  ..- attr(*, "names")= chr  "C3" "C4" "C2" "C3" ...
```

```
> table(citvalid.citccmst$citccmst)
```

```
C1 C2 C3 C4 C5 C6
21 21 19 13 34 15
```

```
> table(citvalid.citccmst$citccmst.mixed)
```

```
  C1 C1C6   C2   C3   C4   C5   C6
  20    1   21   19   13   34   15
```

```
> table(citvalid.citccmst$citccmst.core)
```

```
C1 C2 C3 C4 C5 C6
20 21 19 13 34 15
```

```
> table(citvalid.citccmst$citccmst.confidence)
```

```
 CORE MIXED
  122     1
```

References

[1] Marisa L et al. (2013). *Gene Expression Classification of Colon Cancer into Six Molecular Subtypes: Characterization, Validation and Prognostic value.* PLoS Medicine.

```
Mapping - 57/57 original probes.
```

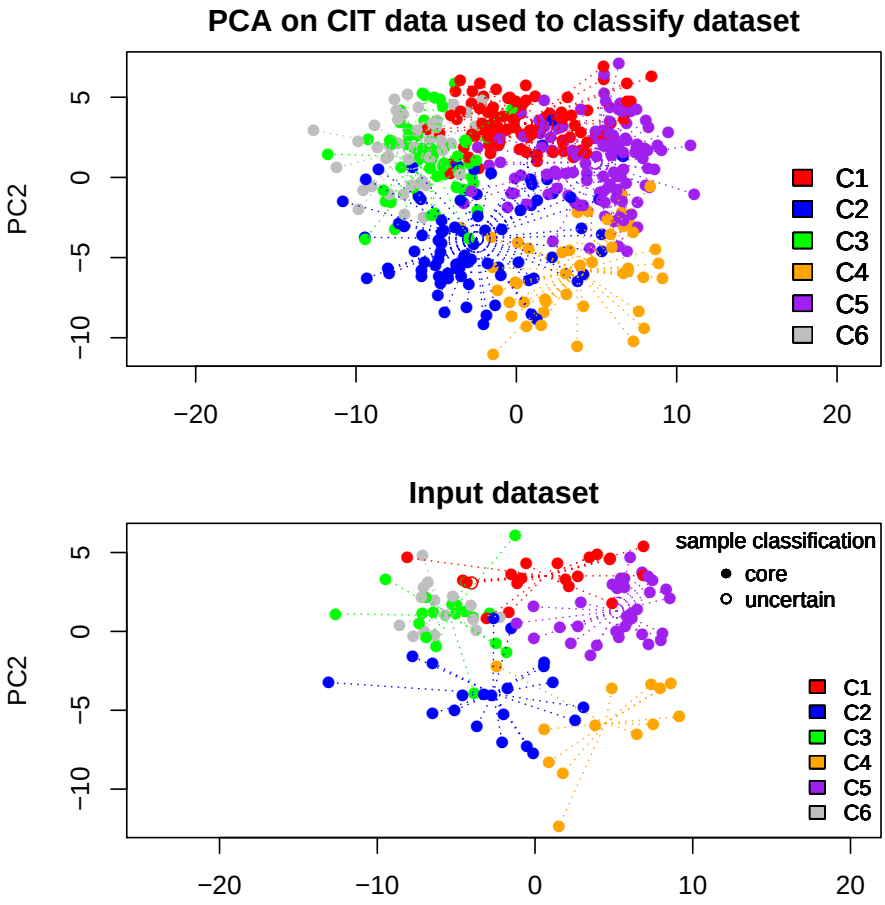

Figure 1: PCA of CIT discovery set data and validation set data