citbcmst : a package to assign CIT Breast Cancer Molecular SubTypes from expression data

<CITR@ligue-cancer.net>

May 24, 2011

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

Overview

The existence of breast cancer molecular subtypes was shown by Perou et al (Nature 2000) using transcriptomic data. The work by Guedj et al [3] yields a new breast cancer molecular classification - called the CIT classification. This package permits to assign a transcriptomic profile to one of the 6 CIT breast cancer molecular subtypes with a confidence index.

Contents

1 CIT Subtypes Assignation Approach

To assign a sample to one of the 6 CIT breast cancer molecular subtypes from expression data, a centroid based approach is used. Centroids of the 6 CIT breast cancer molecular subtypes are defined on a subset (coreset) of 355 samples and 375 probesets (257 HUGO gene symbols). These data are in the object *citbcmst* that can be accessed by :

```
> library(citbcmst)
> data(citbcmst)
> summary(citbcmst)
```

```
           Length Class      Mode
data       355    data.frame list
data.cl    355    -none-     character
data.annot  41    data.frame list
```

citbcmst contains the following objects :

data CIT coreset RMA (uncentered) normalized expression data matrix

data.cl CIT coreset breast cancer molecular subtypes

data.annot CIT coreset probesets annotations provided by NetAffx (version na30) in order to map samples from other plateforms than Affymetrix

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

1. mapping the probes from the new expression dataset to the 375 discrimating probe sets used in the CIT centroids (or to the 257 corresponding HUGO gene symbols, when the microarray platform of the new dataset is not Affymetrix HG U133 plus 2.0)

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

3. recomputing the CIT centroids of the 6 CIT subtypes using the CIT coreset 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 so removed from the analysis.

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

citbcmst subtype of the closest centroid for each sample

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

citbcmst.core subtype of the closest sample for "core" samples (not "outlier" and not mixed", "outlier" and "mixed" samples are set to NA)

citbcmst.confidence confidence annotation CORE, OUTLIER or MIXED

The cutoffs to define mixed and outlier samples are automatically computed but can be set manually (cf help(cit.bcmst)).

Only an expression data matrix/data.frame, with ids as rownames, is required (cf section2). Affymetrix external expression data should be normalized as CIT coreset data, i.e. by RMA method (justRma function in *affy* R package with default parameters), without centering data. Even if other platforms are managed, the approach has been defined for Affymetrix HG U133Plus2 chip expression data, consequently the assignation given for other platforms is expected to be less reliable.

2 Example on an Affymetrix dataset

Here is an example on the public dataset from *Bertheau et al 2007*[1], using Affymetrix HG U133A platform. In this case, mapping to the coreset data can be done on probe sets ids as HG U133A probe sets are included in the HG U133Plus2 chip. Raw data were first normalized by RMA method (justRma function in *affy* R package with default parameters).

```
> data(exp.norm.bertheau07)
> exp.annot.bertheau07 <- data.frame(id = rownames(exp.norm.bertheau07),
+     stringsAsFactors = F, row.names = rownames(exp.norm.bertheau07))

> citbcmst.bertheau07 <- cit.assignBcmst(data = exp.norm.bertheau07,
+     data.annot = exp.annot.bertheau07, data.colId = "id", data.colMap = "id",
+     citbcmst.colMap = "Probe.Set.ID", dist.method = "dlda", plot = TRUE)

Mapping - 241/375 original probes.
Classification - 98.3% of cit data well classified after data reduction.

> str(citbcmst.bertheau07)

'data.frame':        46 obs. of  4 variables:
 $ citbcmst           : chr  "mApo" "mApo" "basL" "basL" ...
 $ citbcmst.mixed     : chr  "mApo" "mApo" "basL" "basL" ...
 $ citbcmst.core      : chr  "mApo" "mApo" "basL" "basL" ...
 $ citbcmst.confidence: chr  "CORE" "CORE" "CORE" "CORE" ...
 - attr(*, "distmethod")= chr "dlda"
 - attr(*, "nb.mapped.probes")= chr "241/375"
 - attr(*, "citmc")= chr "98.3%"
 - attr(*, "scoreGroup")= Named num  519 503 666 503 623 ...
  ..- attr(*, "names")= chr  "normL" "lumA" "mApo" "lumA" ...

> table(citbcmst.bertheau07$citbcmst)

basL lumA lumB lumC mApo
   9    2   15    6   14

> table(citbcmst.bertheau07$citbcmst.mixte)

character(0)

> table(citbcmst.bertheau07$citbcmst.core)

basL lumA lumB lumC mApo
   9    2   13    5   13

> table(citbcmst.bertheau07$citbcmst.confidence)

   CORE   MIXED OUTLIER
     42       1       3
```

```
Mapping - 241/375 original probes.
Classification - 98.3% of cit data well classified after data reduction.
```

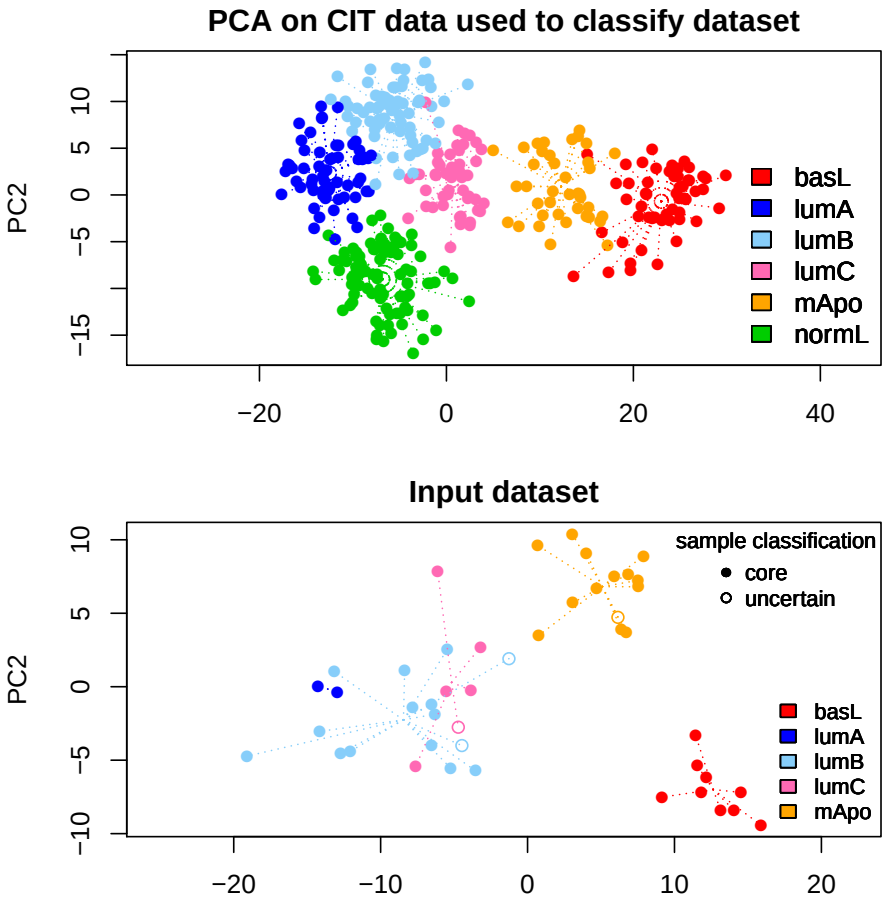

Figure 1: PCA of CIT data and Bertheau et al 2007 dataset

3 Example on an non-Affymetrix dataset

Here is an example on the public dataset from *Chanrion et al. 2008* [2], using 70-mer oligonucleotide microarrays. The CIT classification is optimised for Affymetrix dataset, so the results on other platforms can be less relevant. For non-Affymetrix dataset, the use of Pearson metric is recommended. Nethertheless on this ER+ dataset we obtained very consistent results with no basL samples and only 3 mApo which presented higher ERBB2 expression for 2 of them. This dataset was first aggregated on Gene Symbol annotations given with the data and used to map to CIT data, i.e. `citbcmst.colMap` is set to "Gene.Symbol".

```
> data(exp.norm.chanrion08)
> exp.annot.chanrion08 <- data.frame(id = rownames(exp.norm.chanrion08),
+     gs = rownames(exp.norm.chanrion08), stringsAsFactors = F,
+     row.names = rownames(exp.norm.chanrion08))

> citbcmst.chanrion08 <- cit.assignBcmst(data = exp.norm.chanrion08,
+     data.annot = exp.annot.chanrion08, data.colId = "id", data.colMap = "gs",
+     citbcmst.colMap = "Gene.Symbol", dist.method = "pearson",
+     plot = TRUE)

Mapping - 68/257 original probes.
Classification - 94.4% of cit data well classified after data reduction.

> str(citbcmst.chanrion08)

'data.frame':        155 obs. of  4 variables:
 $ citbcmst           : chr  "lumA" "lumA" "lumA" "lumA" ...
 $ citbcmst.mixed     : chr  "lumA" "lumA" "lumA" "lumA" ...
 $ citbcmst.core      : chr  "lumA" "lumA" "lumA" NA ...
 $ citbcmst.confidence: chr  "CORE" "CORE" "CORE" "OUTLIER" ...
 - attr(*, "distmethod")= chr "pearson"
 - attr(*, "nb.mapped.probes")= chr "68/375"
 - attr(*, "citmc")= chr "94.4%"
 - attr(*, "scoreGroup")= Named num  0.244 0.334 0.286 0.334 0.271 ...
  ..- attr(*, "names")= chr  "normL" "lumA" "mApo" "lumA" ...

> table(citbcmst.chanrion08$citbcmst)

 lumA  lumB  lumC  mApo normL
   93    23    16     3    20

> table(citbcmst.chanrion08$citbcmst.mixte)

character(0)

> table(citbcmst.chanrion08$citbcmst.core)

lumA lumB lumC
  25    1    1

> table(citbcmst.chanrion08$citbcmst.confidence)

   CORE   MIXED OUTLIER
     27       9     119
```

```
Mapping - 68/257 original probes.
Classification - 94.4% of cit data well classified after data reduction.
```

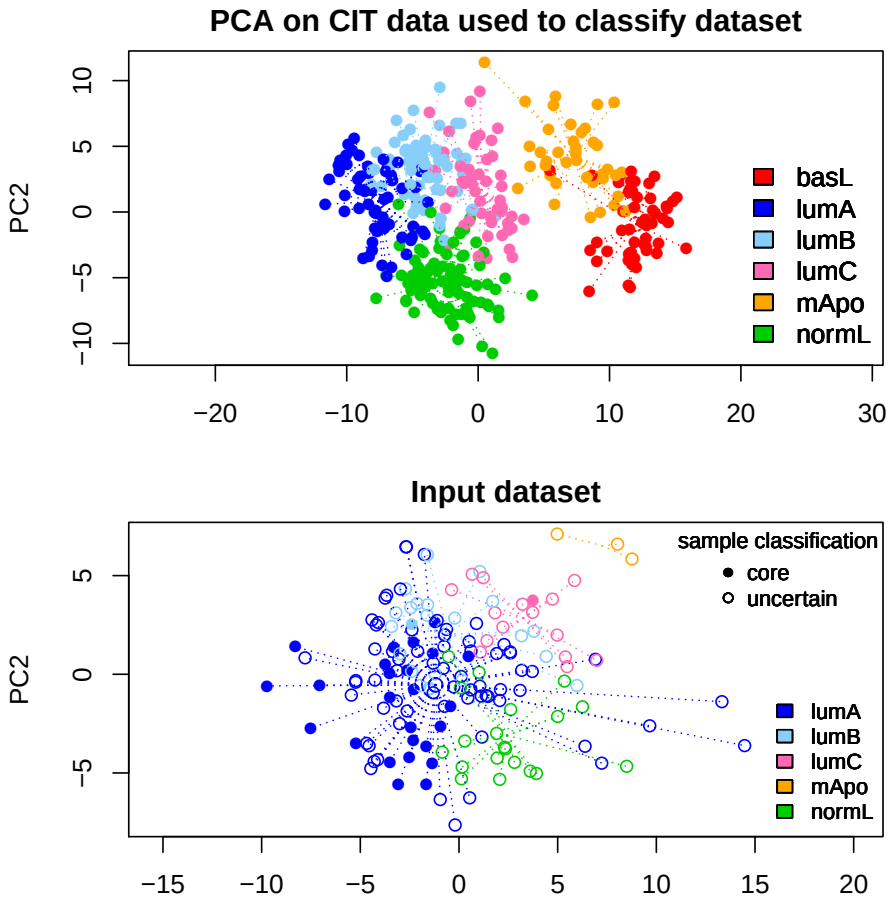

Figure 2: PCA visualisation of CIT data and Chanrion et al 2008 dataset

References

[1] Bertheau P, Turpin E, Rickman DS, Espie M, de Reynies A, Feugeas JP, et al. (2007). *Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen.* PLoS Med, 4: e90.

[2] Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, et al. (2008). *A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer.* Clin Cancer Res, 14: 1744-1752.

[3] Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, et al. (2011). *A refined molecular taxonomy of breast cancer.* Oncogene, submitted.