pbcmc: Permutation-Based Confidence for Molecular Classification

Cristóbal Fresno
CONICET
Universidad Católica de Córdoba

Germán A González
CONICET
Universidad Católica de Córdoba

Andrea S Llera
CONICET
Fundación Instituto Leloir

Elmer A Fernández
CONICET
Universidad Católica de Córdoba

Abstract

The **pbcmc** package characterizes uncertainty assessment on gene expression classifiers, a. k. a. molecular signatures, based on a permutation test. In order to achieve this goal, synthetic simulated subjects are obtained by permutations of gene labels. Then, each synthetic subject is tested against the corresponding subtype classifier to build the null distribution. Thus, classification confidence measurement reports can be provided for each subject, to assist physician therapy choice. At present, it is only available for PAM50 implementation in **genefu** package but, it can easily be extend to other molecular signatures.

Keywords: PAM50, single subject classifier, clinical outcome, breast cancer subtype.

1. Introduction

Gene expression-based classifiers, known as molecular signatures (MS), are gaining increasing attention in oncology and market. The MS can be defined as a set of coordinately expressed genes and an algorithm that use these data to predict disease subtypes, response to therapy, disease risk and clinical outcome (**?**). Particularly in breast cancer market there exists many MS such as PAM50 (**??**), Prosigna® (**?**), Oncotype DX® (**?**) and MammaPrint® (**?**). In essence, most MS try to provide patient subtype classification or risk prediction which has been associated with distant metastasis free survival (DMFS) or relapse free survival (RFS). Consequently, they are intended to be used to support therapy choice. However, several authors have shown that data processing steps, technology, as well as population variability have an effect on measured gene expression and could bias subtype/risk subject assignment (**??????**). These effects suggest that, from a statistical point of view, MS are not robust for subject classification. In particular, there is no control over type I and II like classification error and subjects could potentially be assigned to a wrong subtype/risk class. Indeed, the lack of certainty in class assignation could lead to a misleading therapy affecting subject outcome. Hence, the development of methods for significance or certainty on MS class assignation is crucial in order to assist physicians' decision making (**??**).

1.1. PAM50 Molecular Siganture

The well-known breast cancer (BC) MS, PAM50 (**??**), is based on the comparison between the patient gene profiles (PGP) of 50 expressed genes, against five intrinsic genes profiles (IGP) representing: Basal, Her2-enriched, Luminal A, Luminal B and Normal-like subtypes using Spearman's ρ correlation. Then, the subject will be assigned to the $i-th$ subtype according to (**??**):

$$\arg\max_{i \in IGP} \rho(PGP, IGP_i) \tag{1}$$

Particularly, patients are assigned to the $i-th$ subtype which maximize $\rho(PGP, IGP_i)$, even if correlation is weak or there are similar/tight IGP correlations. The latter case has been addressed by **?**, where they have excluded subjects with a correlation difference between Luminal A and B of less than 0.1, considering them ambiguous pattern, as a way to control a kind of type II error. However, type I error control is still a debt.

1.2. genefu library

At present, the freely available PAM50 **genefu** algorithm implementation (**?**), offers a kind of subtype probability, $P(IGP_i)$, calculated as (**??**):

$$P(IGP_i) = \frac{\rho(PGP, IGP_i)}{\sum_i \rho(PGP, IGP_i)} \quad \forall \rho(PGP, IGP_i) > 0 \tag{2}$$

However, this probability does not take part in the classification rule. Even worse, a very low ρ (weak relationship) could reach a high subtype probability, for instance, if all the others ρ's are close to zero or are even negative.

1.3. Alternative proposal

In order to overcome these drawbacks, here a simple and reliable single subject classifier to control type I and II errors is proposed. Morevover, it provides a statistical significance on subtype assignation based on a gene label permutation test. Briefly, it evaluates if the observed ρ of each IGP can be achieved by chance, regarding subject observed MS expressed gene values. In addition, we propose a user-friendly subtype assignation panel to support physicians' decision making, enhancing PAM50 or commercial reports currently available in the market. The method is presented for PAM50 but can easily be extended to other MS algorithms.

2. Methods

The Permutation Based Confidence for Molecular Classification (**pbcmc**) package, estimates the statistical significance of ρ for each IGP. In other words, we want to see whether the observed ρ can be obtained by chance. In order to perform this task, the ρ null distribution for each IGP ($\rho_{H0_{IGP}}$) is obtained by evaluating β permutations of the SGP gene expression. Then, the observed (un-permuted) IGP correlations ($\rho_{u_{IGP}}$) are compared against their own

$\rho_{H0_{IGP}}$ in order to evaluate whether $H_0 : \rho_{u_{IGP}} \in \rho_{H0_{IGP}}$ versus $H_1 : \rho_{u_{IGP}} \notin \rho_{H0_{IGP}}$ according to the p-values (p_{IGP}) calculated as in (**??**):

$$p_{IGP} = \frac{\sum_i^{\beta} I(\rho_i, \rho_u)}{\beta}; \quad I(a,b) = \begin{cases} 1 & if \quad a > b \\ 0 & if \quad a \leq b \end{cases} \tag{3}$$

where IGP stands for Basal, Her2-Enriched, Luminal A, Luminal B and Normal-like. The resulting five p_{IGP}'s are adjusted to control multiple comparisons using False Discovery Rate, FDR (**?**). Then, assuming an acceptable type I error, α, the hypothesis test for all IGPs could result in:

i. No significant ρ_u for any IGP, i. e., all adjusted $p_{IGP} > \alpha$.

ii. A unique significant ρ_u.

iii. Multiple significant ρ_u.

For the first case, the subject cannot reliably be assigned to any IGP (not assigned - **NA**). In the second case, it is assigned (**A**) to the trustworthy current PAM50 subtype. In order to overcome the ambiguity of case iii., a correlation difference threshold of 0.1 between the top ones ($\rho_{u_{IGP_1}} > \rho_{u_{IGP_2}} > \ldots > 0$) was established, similarly as in **?** for Luminal subtypes. Then, if $(\rho_{u_{IGP_1}} - \rho_{u_{IGP_2}}) > 0.1$ the subject is assigned as in ii., otherwise it is considered as an ambiguous (**Amb**) subject.

3. Implementation

The S4 class hierarchy of the **pbcmc** package is based on the implementation of an abstract `MolecularPermutationClassifier` class, which can potentially be used for any MS as depicted in Figure **??**. Basically, it has been developed as an organized data processing framework for its heirs. The formers are supposed to implement the respective responsibilities. Once a heir object is implemented the user can:

loadBCDataset: Load one of the example breast cancer dataset available at Bioconductor. At present, it is possible to load breastCancerXXX where XXX can be "`upp`" (**?**), "`nki`" (**?**), "`vdx`" (**?**), "`mainz`" (**?**), "`transbig`" (**?**) or "`unt`" (**?**).

filtrate: Remove, from the `exprs` matrix, subjects not required by the classification algorithm.

classify: Generate subject classification according to the heir's implementation (PAM50, etc.).

permute: Obtains subject classification based on the null ρ_{H0} distribution by means of β permutation simulations.

subtype: Obtain the new classification using the permutation results.

subjectReport: Create a friendly report to assist physician treatment decision making.

databaseReport: Create a `pdf` with all `subjectReports`, if a database is available.

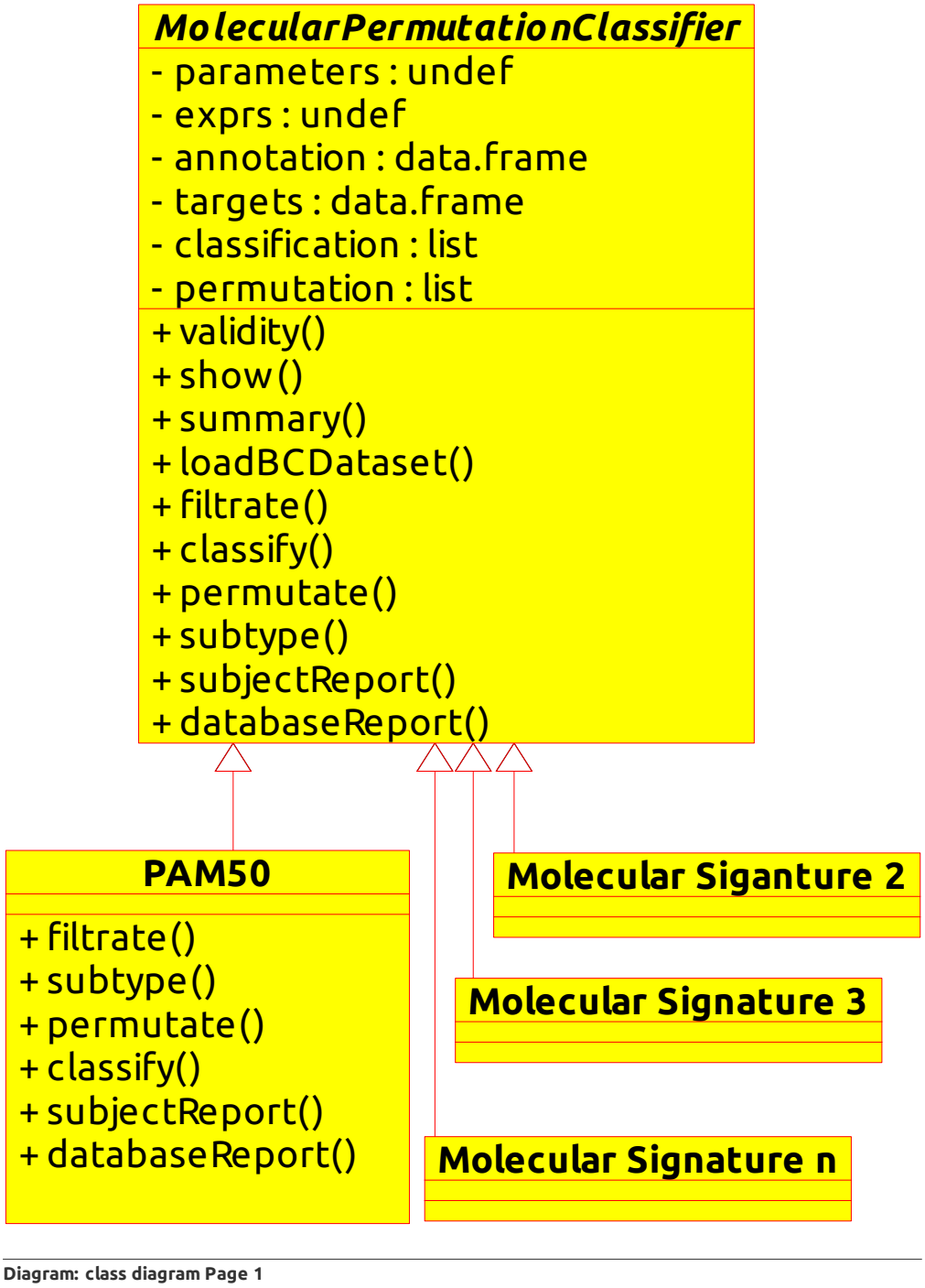

Figure 1: Package hierarchy. `MolecularPermutationClassifier` is the main S4 abstract class and `PAM50` implements it for the Molecular Signature (MS) of **??**. The other classes represent user-defined implementations of other MS. Note that complete operation signature have been omitted for simplicity.

At present, the only available heir is PAM50 based on **genefu** library (**?**). But, it can easily be extended to other MS such as Prosigna® (**?**) or others, just implementing `filtrate`, `classify`, `permutate`, `subtype`, `subjectReport` and `databaseReport` functions.

3.1. Example

In order to work with `PAM50` MS, the user must load a Bioconductor's breast cancer dataset. For example we can load NKI database (**?**), provided that the require library is installed, using the following code:

```
R> library("pbcmc")
R> library("BiocParallel")
R> object<-loadBCDataset(Class=PAM50, libname="nki", verbose=TRUE)
R> object

A PAM50 molecular permutation classifier object
Dimensions:
            nrow ncol
exprs      24481  337
annotation 24481   10
targets      337   21
```

The `object` is a PAM50 instance, which contains the `exprs` matrix with gene expression values, associated `annotation` and clinical data in `targets` data.frame. On the other hand, the user can use PAM50's constructor to create an object with his/her own data or convert a **limma** `MAList` object into PAM50 using `as.PAM50(MAList_object)` function. In the first case, the user will only need:

a) The **M** gene expression object, i. e., genes in rows and samples in columns.

b) The **annotation** `data.frame` which must include the compulsory fields: "`probe`", "`NCBI.gene.symbol`" and "`EntrezGene.ID`".

For example, we could use **genefu**'s PAM50 centroids to check if our implementation solves the proof of concept, where we *a priori* know the true class of each subject:

```
R> M<-pam50$centroids
R> genes<-pam50$centroids.map
R> names(genes)<-c("probe", "NCBI.gene.symbol", "EntrezGene.ID")
R> object<-PAM50(exprs=M, annotation=genes)
R> object

A PAM50 molecular permutation classifier object
Dimensions:
           nrow ncol
exprs        50    5
annotation   50    3
targets       0    0
```

Note that for the above output, the `targets` slot is empty, i. e., `nrow=0` and `ncol=0`. In addition, only the expression of the fifty genes and five IGP is available, with its corresponding `annotation` over the three compulsory fields. It is always a good idea to explore the slots content, to see whether they have been correctly loaded:

```
R> head(exprs(object))      ##The gene expression values for each subject

         Basal    Her2      LumA    LumB  Normal
ACTR3B  0.7183 -0.4817  0.009981 -0.1906  0.4657
ANLN    0.5374  0.2669 -0.579246  0.0988 -0.8369
BAG1   -0.5745 -0.4761  0.758221 -0.4055  0.3166
BCL2   -0.1188 -0.1579  0.287487 -0.4413  0.5340
BIRC5   0.3005  0.4057 -0.881434  0.6039 -0.8766
BLVRA  -0.6427  0.3353  0.042042  0.6912 -0.1634

R> head(annotation(object)) ##The compulsory annotation fields

        probe NCBI.gene.symbol EntrezGene.ID
ACTR3B ACTR3B           ACTR3B         57180
ANLN     ANLN             ANLN         54443
BAG1     BAG1             BAG1           573
BCL2     BCL2             BCL2           596
BIRC5   BIRC5            BIRC5           332
BLVRA   BLVRA            BLVRA           644

R> head(targets(object))    ##The clinical data, if available.

data frame with 0 columns and 0 rows
```

Just as we expected, the five centroids are loaded in `exprs` slot, with their corresponding "probe", "NCBI.gene.symbol" and "EntrezGene.ID" number in the `annotation` slot and no available data for the `targets`. Now, the user is ready to work with the data following the workflow suggested in section **??** (filtrate, classify and permutate):

```
R> object<-filtrate(object, verbose=TRUE)
R> object<-classify(object, std="none", verbose=TRUE)
R> object<-permutate(object, nPerm=10000, pCutoff=0.01, where="fdr",
+      corCutoff=0.1, keep=TRUE, seed=1234567890, verbose=TRUE,
+      BPPARAM=bpparam())

  |
  |                                                              |   0%
  |
  |============                                                  |  20%
  |
  |========================                                      |  40%
```

```
  |
  |==================================                    |  60%
  |
  |============================================          |  80%
  |
  |======================================================| 100%
```

The intention of `filtrate` function is to keep only the genes that will take place in the classification. In this example, it will not produce any change on the original `exprs` slot, given the fact that only the required fifty genes are present. But, if a complete microarray would have been present, then, probes that do not code for IGP will be removed. In addition, probes that code for the same gene (repeated or with similar annotation) will be treated as described in standardization (`std`) parameter.

Once genes are `filtrate`d, the user can `classify` them using the original PAM50 algorithm. However, here we propose to obtain subtype assignment confidence using at least $\beta = 10.000$ permutations over the SGP, using a type I error $\alpha = 0.01$ on the adjusted p-values ("`fdr`") and a correlation difference threshold of `corCutoff=0.1`. As a matter of fact, this process is computationally intensive, so we can take advantage of all the available computes cores using **BioParallel** package (`BPPARAM=bpparam()`) as we just did (`?`). In addition, the user can track the permutation progress bar by including `verbose=TRUE` option. If we now take a look at the `object`:

```
R> object

A PAM50 molecular permutation classifier object
Dimensions:
           nrow ncol
exprs        50    5
annotation   50    3
targets       0    0
Classification:
            nrow ncol
probability    5    5
correlation    5    5
$subtype

 Basal   Her2   LumA   LumB Normal
     1      1      1      1      1

Permutations test ran with following parameters:
 Permutations=10000, fdr<0.01, corCutoff>0.1, keep=TRUE
Permutation:
correlation available: TRUE
        nrow ncol
pvalues    5    5
fdr        5    5
subtype    5    5
```

we can see that it has been updated. First, the `classification` slot contains two datasets: one with the subtype *probability*, $P(IGP_i)$, as described in section **??** and the *correlation* of each subject with the five IGPs. The *$subtype* item shows a frequency table of the possible IGPs with the used subjects. In addition, the used `permutation parameters` are shown with the dimension of *pvalues*, *fdr* and *subtypes*. Note that in this case `keep=TRUE` option was used so, the simulated correlation null distribution data points ($\rho_{H0_{IGPs}}$) are available.

In this example we have used **genefu**'s PAM50 centroids, thus, only one subject is present (1) for each IGP cell in the `object` output. This result is also confirmed by the ones in the diagonal of `summary(object)` matrix between the original *Subtype* and the *Classes* found by the **pbcmc** package. Moreover, this toy example only shows assigned subjects (**A**) to the original PAM50 subtypes, whereas not assigned (**NA**) marginal row/column contains only zeros (0). If ambiguous (**AMB**) subjects would have been found, *Classes* column will have included additional rows with the classes in dispute (e. g., "LumA, Normal" or "Her2, LumB", etc.).

```
R> summary(object)
```

```
             Subtype
Classes       Basal Her2 LumA LumB Normal Not Assigned
  Basal           1    0    0    0      0            0
  Her2            0    1    0    0      0            0
  LumA            0    0    1    0      0            0
  LumB            0    0    0    1      0            0
  Normal          0    0    0    0      1            0
  Not Assigned    0    0    0    0      0            0
```

Finally, we can inspect the report of a single subject to see how the MS classification went (Figure **??**), in order to suggest an appropriate therapy for the physician:

```
R> subjectReport(object, subject=1)
```

The report of Figure **??** is a `grid.arrange` object which basically consists of three main parts:

tableGrob: A summary table which contains the following fields,

$Summary: `Subject` name and subtype obtained by `PAM50.Subtype` or the proposed methodology (`Permuted.Subtype`).

$Fields: For the five PAM50 subtypes,

- Correlation: The correlation of the $i-th$ PAM50 centroid with the observed subject exprs, $\rho(PGP, IGP_i)$.
- p-value: Permutation p-value obtained using the simulation data, p_{IGP}.
- FDR: Adjusted p-value using False Discovery Rate (**?**).

facet_wrap: Two rows to display **ggplot2** (**?**) scatter plots of subject exprs versus PAM50 centroids (**??**) and a linear regression fit (in blue). If the subject has an unique subtype, then the graph is colored in red. In addition, if simulated permutations were run with `keep=TRUE` option, then the null distribution boxplots are plotted with the corresponding observed un-permuted correlations as big round dots.

	Summary	Field	Basal	Her2	LumA	LumB	Normal
Subject	Basal	Correlation	1	−0.146	−0.649	−0.360	0.118
PAM50.Subtype	Basal	p−value	0	0.840	1.000	0.994	0.209
Permuted.Subtype	Basal	FDR	0	1.000	1.000	1.000	0.523

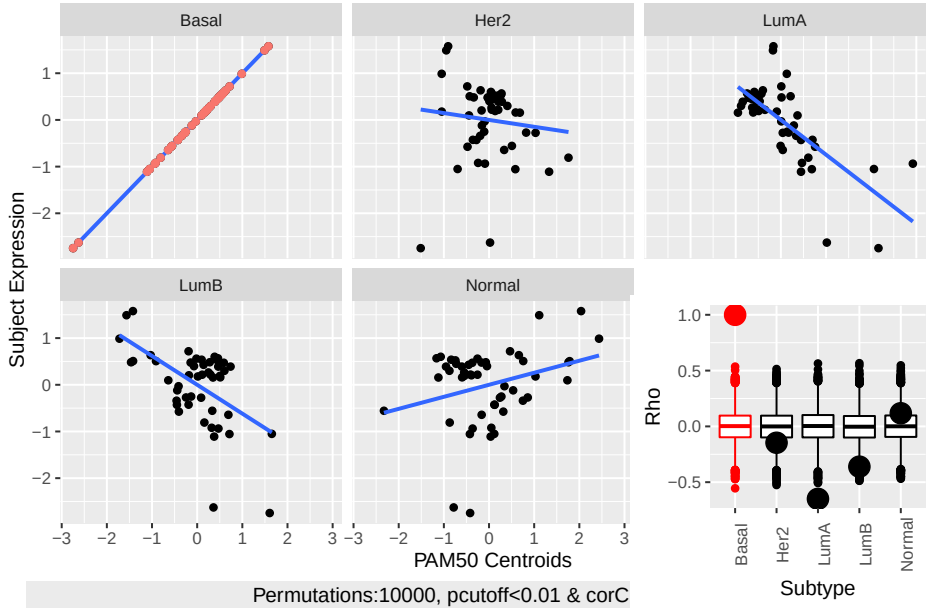

Figure 2: PAM50 permutation subject report for **genefu**'s "Basal" intrinsic gene profile (IGP). The top table summarize the results for the i-th subject, i. e., the correlation, p-value and false discovery rate (fdr) obtained for each IGP. In addition, scatter plots of the observed subject gene profiles against the IGP with the linear regression line (in blue). Red color indicates the assigned subtype. Finally, a boxplot for each IGP null permuted correlation distribution and big dots to represent the un-permuted observed correlations.

textGrob:The permutation `parameter` slot used in the simulation.

The **pbcmc** also includes the ability to get a `pdf` report for the complete database calling `databaseReport` function. In this context, the first page is a global summary of the database, i.e., a `summary` contingency table of the permuted test classes against the original PAM50 subtypes results. The following pages are the respective `subjectReport` outputs as the one shown in Figure **??**.

4. Conclusion

The **pbcmc** package characterizes uncertainty assessment on gene expression classifiers, a. k. a. molecular signatures, based on a permutation test. In order to achieve this goal, synthetic simulated subjects are obtained bypermutations of gene labels. Then, each synthetic subject is tested against the corresponding subtype classifier to build the null distribution. Thus, classification confidence measurement report can be provided for each subject, to assist physician therapy choice. At present, it is only available for PAM50 implementation in **genefu** package but, it can easily be extend to other molecular signatures.

Acknowledgements

Funding: This work was supported by Universidad Católica de Córdoba (PIP 800-201304-00047-CC to E.A.F.), Argentina and National Council of Scientific and Technical Research (CONICET), Argentina.

Session Info

```
R> sessionInfo()

R version 3.3.0 RC (2016-04-26 r70550)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.9.5 (Mavericks)

locale:
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
 [1] BiocParallel_1.6.0     pbcmc_1.0.0
 [3] genefu_2.4.0           AIMS_1.4.0
 [5] Biobase_2.32.0         BiocGenerics_0.18.0
 [7] e1071_1.6-7            iC10_1.1.3
```

```
 [9] iC10TrainingData_1.0.1 pamr_1.55
[11] cluster_2.0.4          biomaRt_2.28.0
[13] limma_3.28.0           mclust_5.2
[15] survcomp_1.22.0        prodlim_1.5.7
[17] survival_2.39-2

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.4.5         plyr_1.8.3            bitops_1.0-6
 [4] class_7.3-14          tools_3.3.0           gtable_0.2.0
 [7] RSQLite_1.0.0         lattice_0.20-33       Matrix_1.2-6
[10] DBI_0.4               bootstrap_2015.2      amap_0.8-14
[13] gridExtra_2.2.1       stringr_1.0.0         SuppDists_1.1-9.2
[16] S4Vectors_0.10.0      IRanges_2.6.0         stats4_3.3.0
[19] grid_3.3.0            cowplot_0.6.2         breastCancerNKI_1.9.0
[22] AnnotationDbi_1.34.0  XML_3.98-1.4          lava_1.4.3
[25] rmeta_2.16            magrittr_1.5          reshape2_1.4.1
[28] ggplot2_2.1.0         scales_0.4.0          survivalROC_1.0.3
[31] splines_3.3.0         colorspace_1.2-6      labeling_0.3
[34] KernSmooth_2.23-15    stringi_1.0-1         munsell_0.4.3
[37] RCurl_1.95-4.8
```

Affiliation:

Cristóbal Fresno, Germán A González & Elmer A Fernández
Bioscience Data Mining Group
Facultad de Ingeniería
Universidad Católica de Córdoba - CONICET
X5016DHK Córdoba, Argentina
E-mail: cfresno@bdmg.com.ar, ggonzalez@bdmg.com.ar, efernandez@bdmg.com.ar
URL: http://www.bdmg.com.ar/

Andrea S Llera
Laboratorio de Terapia Molecular y Celular
Fundación Instituto Leloir - CONICET
C1405BWE Ciudad Autónoma de Buenos Aires, Argentina
E-mail: allera@leloir.org.ar
URL: http://www.leloir.org.ar/podhajcer/