# MIGSA: Getting pbcmc datasets

**Juan C Rodriguez**
CONICET
Universidad Católica de Córdoba
Universidad Nacional de Córdoba

**Cristóbal Fresno**
Instituto Nacional de Medicina Genómica

**Andrea S Llera**
CONICET
Fundación Instituto Leloir

**Elmer A Fernández**
CONICET
Universidad Católica de Córdoba
Universidad Nacional de Córdoba

---

**Abstract**

In this vignette we are going to show how we got the RData *pbcmcData.RData* which can be loaded via the **MIGSAdata** package using data(pbcmcData).

*Keywords*: singular enrichment analysis, over representation analysis, gene set enrichment analysis, functional class scoring, big omics data.

---

## 1. Getting the data

Following we give the used code to download this data and their PAM50 subtypes.

```
> library(limma);
> library(pbcmc);
> # datasets included in BioConductor repository
> libNames <- c("mainz", "nki", "transbig", "unt", "upp", "vdx");
> # let's load them!
> pbcmcData <- lapply(libNames, function(actLibName) {
+     print(actLibName);
+
+     # the pbcmc package provides an easy way to download and classify them
+     actLib <- loadBCDataset(Class=PAM50, libname=actLibName, verbose=FALSE);
+     actLibFilt <- filtrate(actLib, verbose=FALSE);
+     actLibFilt <- classify(actLibFilt, std="none", verbose=FALSE);
+     actSubtypes <- classification(actLibFilt)$subtype;
+
+     # get the expression matrix and the annotation
+     actExprs <- exprs(actLib);
+     actAnnot <- annotation(actLib);
+
```

```
+     # we recommend working allways with Entrez IDs, let's match them with
+     # expression matrix rownames (and modify them)
+     if (all(actAnnot$probe == rownames(actExprs))) {
+         actExprs <- actExprs[!is.na(actAnnot$EntrezGene.ID),];
+         actAnnot <- actAnnot[!is.na(actAnnot$EntrezGene.ID),];
+         rownames(actExprs) <- as.character(actAnnot$EntrezGene.ID);
+     } else {
+         matchedEntrez <- match(rownames(actExprs), actAnnot$probe);
+         # all(rownames(actExprs) %in% actAnnot$probe == !is.na(matchedEntrez));
+
+         stopifnot(all(
+             actAnnot$probe[!is.na(matchedEntrez)] ==
+             rownames(actExprs)[!is.na(matchedEntrez)]));
+
+         actExprs <- actExprs[!is.na(matchedEntrez),];
+         actAnnot <- actAnnot[!is.na(matchedEntrez),];
+         stopifnot(all(actAnnot$probe == rownames(actExprs)));
+         actExprs <- actExprs[!is.na(actAnnot$EntrezGene.ID),];
+         actAnnot <- actAnnot[!is.na(actAnnot$EntrezGene.ID),];
+         rownames(actExprs) <- as.character(actAnnot$EntrezGene.ID);
+     }
+
+     # average repeated genes expression
+     actExprs <- avereps(actExprs);
+
+     stopifnot(all(colnames(actExprs) == names(actSubtypes)));
+     # filtrate only these two conditions
+     actExprs <- actExprs[, actSubtypes %in% c("Basal", "LumA")];
+     actSubtypes <- as.character(
+         actSubtypes[actSubtypes %in% c("Basal", "LumA")]);
+
+     return(list(geneExpr=actExprs, subtypes=actSubtypes));
+ })

[1] "mainz"
[1] "nki"
[1] "transbig"
[1] "unt"
[1] "upp"
[1] "vdx"

> names(pbcmcData) <- libNames;
```

And let's check it is the same data.

```
> # save the just created pbcmcData to newPbcmcData
> newPbcmcData <- pbcmcData;
```

```
> library(MIGSAdata);
> # and load the MIGSAdata one.
> data(pbcmcData);
> all.equal(newPbcmcData, pbcmcData);

[1] TRUE
```

## Session Info

```
> sessionInfo()

R version 3.5.1 Patched (2018-07-12 r74967)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.5 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.8-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.8-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] pbcmc_1.9.0            genefu_2.14.0          AIMS_1.14.0
 [4] e1071_1.7-0            iC10_1.4.2             iC10TrainingData_1.3.1
 [7] pamr_1.55              cluster_2.0.7-1        biomaRt_2.38.0
[10] mclust_5.4.1           survcomp_1.32.0        prodlim_2018.04.18
[13] survival_2.43-1        edgeR_3.24.0           MIGSAdata_1.5.0
[16] MIGSA_1.6.0            mGSZ_1.0               ismev_1.42
[19] mgcv_1.8-25            nlme_3.1-137           MASS_7.3-51
[22] limma_3.38.0           GSA_1.03               BiocParallel_1.16.0
[25] GSEABase_1.44.0        graph_1.60.0           annotate_1.60.0
[28] XML_3.98-1.16          AnnotationDbi_1.44.0   IRanges_2.16.0
[31] S4Vectors_0.20.0       Biobase_2.42.0         BiocGenerics_0.28.0

loaded via a namespace (and not attached):
```

```
 [1] survivalROC_1.0.3           Category_2.48.0
 [3] breastCancerUNT_1.19.0      bitops_1.0-6
 [5] matrixStats_0.54.0          bit64_0.9-7
 [7] httr_1.3.1                  progress_1.2.0
 [9] Rgraphviz_2.26.0            tools_3.5.1
[11] R6_2.3.0                    vegan_2.5-3
[13] KernSmooth_2.23-15          DBI_1.0.0
[15] lazyeval_0.2.1              colorspace_1.3-2
[17] rmeta_3.0                   permute_0.9-4
[19] gridExtra_2.3               prettyunits_1.0.2
[21] tidyselect_0.2.5            bit_1.1-14
[23] compiler_3.5.1              formatR_1.5
[25] breastCancerNKI_1.19.0      ggdendro_0.1-20
[27] labeling_0.3                scales_1.0.0
[29] genefilter_1.64.0           RBGL_1.58.0
[31] stringr_1.3.1               digest_0.6.18
[33] breastCancerVDX_1.19.0      AnnotationForge_1.24.0
[35] pkgconfig_2.0.2             rlang_0.3.0.1
[37] RSQLite_2.1.1               SuppDists_1.1-9.4
[39] bindr_0.1.1                 GOstats_2.48.0
[41] dplyr_0.7.7                 RCurl_1.95-4.11
[43] magrittr_1.5                GO.db_3.7.0
[45] futile.logger_1.4.3         Matrix_1.2-14
[47] Rcpp_0.12.19                munsell_0.5.0
[49] stringi_1.2.4               RJSONIO_1.3-0
[51] org.Hs.eg.db_3.7.0          plyr_1.8.4
[53] breastCancerUPP_1.19.0      grid_3.5.1
[55] blob_1.1.1                  breastCancerTRANSBIG_1.19.0
[57] crayon_1.3.4                lattice_0.20-35
[59] cowplot_0.9.3               splines_3.5.1
[61] hms_0.4.2                   locfit_1.5-9.1
[63] pillar_1.3.0                reshape2_1.4.3
[65] futile.options_1.0.1        glue_1.3.0
[67] lambda.r_1.2.3              data.table_1.11.8
[69] bootstrap_2017.2            gtable_0.2.0
[71] purrr_0.2.5                 amap_0.8-16
[73] assertthat_0.2.0            ggplot2_3.1.0
[75] xtable_1.8-3                class_7.3-14
[77] tibble_1.4.2                memoise_1.1.0
[79] bindrcpp_0.2.2              lava_1.6.3
[81] breastCancerMAINZ_1.19.0
```

**Affiliation:**

Juan C Rodriguez & Elmer A Fernández
Bioscience Data Mining Group
Facultad de Ingeniería
Universidad Católica de Córdoba - CONICET
X5016DHK Córdoba, Argentina
E-mail: jcrodriguez@bdmg.com.ar, efernandez@bdmg.com.ar
URL: http://www.bdmg.com.ar/