

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 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.5-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_1.4.0          genefu_2.8.0        AIMS_1.8.0
[4] e1071_1.6-8         iC10_1.1.3          iC10TrainingData_1.0.1
[7] pamr_1.55           cluster_2.0.6       biomaRt_2.32.1
[10] mclust_5.3          survcomp_1.26.0     proclim_1.6.1
[13] survival_2.41-3     GO.db_3.4.1         edgeR_3.18.1
[16] MIGSadata_1.0.0     MIGSA_1.0.1         mGSZ_1.0
[19] ismev_1.41          mgcv_1.8-17         nlme_3.1-131
[22] MASS_7.3-47         limma_3.32.2        GSA_1.03
[25] BiocParallel_1.10.1 GSEABase_1.38.0     graph_1.54.0
[28] annotate_1.54.0     XML_3.98-1.9        AnnotationDbi_1.38.1
[31] IRanges_2.10.2     S4Vectors_0.14.3   Biobase_2.36.2
[34] BiocGenerics_0.22.0
```

loaded via a namespace (and not attached):

```
[1] survivalROC_1.0.3          Category_2.42.1
[3] breastCancerUNT_1.14.0    bitops_1.0-6
[5] matrixStats_0.52.2        bit64_0.9-7
[7] Rgraphviz_2.20.0          tools_3.4.0
[9] vegan_2.4-3                KernSmooth_2.23-15
[11] DBI_0.7                    lazyeval_0.2.0
[13] colorspace_1.3-2          rmeta_2.16
[15] permute_0.9-4             gridExtra_2.2.1
[17] bit_1.1-12                compiler_3.4.0
[19] breastCancerNKI_1.14.0    labeling_0.3
[21] caTools_1.17.1           scales_0.4.1
[23] genefilter_1.58.1         RBGL_1.52.0
[25] stringr_1.2.0             digest_0.6.12
[27] breastCancerVDX_1.14.0    AnnotationForge_1.18.0
[29] pkgconfig_2.0.1          rlang_0.1.1
[31] RSQLite_2.0               SuppDists_1.1-9.4
[33] GOSTats_2.42.0           gtools_3.5.0
[35] magrittr_1.5              RCurl_1.95-4.8
[37] futile.logger_1.4.3       Matrix_1.2-10
[39] Rcpp_0.12.11             munsell_0.4.3
[41] stringi_1.1.5            RJSONIO_1.3-0
[43] org.Hs.eg.db_3.4.1        gplots_3.0.1
[45] plyr_1.8.4                breastCancerUPP_1.14.0
[47] grid_3.4.0                blob_1.1.0
[49] breastCancerTRANSBIG_1.14.0 gdata_2.18.0
[51] lattice_0.20-35          cowplot_0.7.0
[53] splines_3.4.0            locfit_1.5-9.1
[55] reshape2_1.4.2           futile.options_1.0.0
[57] lambda.r_1.1.9           data.table_1.10.4
[59] bootstrap_2017.2         gtable_0.2.0
[61] amap_0.8-14              ggplot2_2.2.1
[63] xtable_1.8-2             class_7.3-14
[65] tibble_1.3.3             memoise_1.1.0
[67] lava_1.5                  breastCancerMAINZ_1.14.0
```

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/>