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));
+ })
> 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);
```

Session Info

```
> sessionInfo()

R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.3 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.14-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.14-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB              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    stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] edgeR_3.36.0         MIGSAdata_1.17.0     MIGSA_1.18.0
 [4] mGSZ_1.0             ismev_1.42           mgcv_1.8-38
 [7] nlme_3.1-153         MASS_7.3-54          limma_3.50.0
[10] GSA_1.03.1           BiocParallel_1.28.0  GSEABase_1.56.0
[13] graph_1.72.0         annotate_1.72.0      XML_3.99-0.8
[16] AnnotationDbi_1.56.0 IRanges_2.28.0       S4Vectors_0.32.0
[19] Biobase_2.54.0       BiocGenerics_0.40.0

loaded via a namespace (and not attached):
 [1] Category_2.60.0        bitops_1.0-7           matrixStats_0.61.0
 [4] bit64_4.0.5            httr_1.4.2             GenomeInfoDb_1.30.0
 [7] Rgraphviz_2.38.0       tools_4.1.1            utf8_1.2.2
[10] R6_2.5.1               vegan_2.5-7            DBI_1.1.1
[13] colorspace_2.0-2       permute_0.9-5          tidyselect_1.1.1
[16] bit_4.0.4              compiler_4.1.1         formatR_1.11
[19] ggdendro_0.1.22        labeling_0.4.2         scales_1.1.1
[22] genefilter_1.76.0      RBGL_1.70.0            digest_0.6.28
[25] stringr_1.4.0          AnnotationForge_1.36.0 XVector_0.34.0
[28] pkgconfig_2.0.3        fastmap_1.1.0          rlang_0.4.12
[31] rstudioapi_0.13        RSQLite_2.2.8          farver_2.1.0
[34] GOstats_2.60.0         generics_0.1.1         jsonlite_1.7.2
[37] dplyr_1.0.7            RCurl_1.98-1.5         magrittr_2.0.1
[40] GO.db_3.14.0           GenomeInfoDbData_1.2.7 futile.logger_1.4.3
[43] Matrix_1.3-4           Rcpp_1.0.7             munsell_0.5.0
```

```
[46] fansi_0.5.0           lifecycle_1.0.1       stringi_1.7.5
[49] zlibbioc_1.40.0       org.Hs.eg.db_3.14.0   plyr_1.8.6
[52] grid_4.1.1            blob_1.2.2            parallel_4.1.1
[55] crayon_1.4.1          lattice_0.20-45       Biostrings_2.62.0
[58] splines_4.1.1         KEGGREST_1.34.0       locfit_1.5-9.4
[61] pillar_1.6.4          reshape2_1.4.4        futile.options_1.0.1
[64] glue_1.4.2            lambda.r_1.2.4        data.table_1.14.2
[67] png_0.1-7             vctrs_0.3.8           gtable_0.3.0
[70] purrr_0.3.4           assertthat_0.2.1      cachem_1.0.6
[73] ggplot2_3.3.5         xtable_1.8-4          survival_3.2-13
[76] tibble_3.1.5          memoise_2.0.0         cluster_2.1.2
[79] ellipsis_0.3.2
```

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/