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 3.5.0 (2018-04-23)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2012 R2 x64 (build 9600)

Matrix products: default

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

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

other attached packages:
 [1] pbcmc_1.8.0            genefu_2.12.0          AIMS_1.12.0
 [4] e1071_1.6-8            iC10_1.1.3             iC10TrainingData_1.0.1
 [7] pamr_1.55              cluster_2.0.7-1        biomaRt_2.36.0
[10] mclust_5.4             survcomp_1.30.0        prodlim_2018.04.18
[13] survival_2.42-3        edgeR_3.22.0           MIGSAdata_1.3.0
[16] MIGSA_1.4.0            mGSZ_1.0               ismev_1.41
[19] mgcv_1.8-23            nlme_3.1-137           MASS_7.3-50
[22] limma_3.36.0           GSA_1.03               BiocParallel_1.14.0
[25] GSEABase_1.42.0        graph_1.58.0           annotate_1.58.0
[28] XML_3.98-1.11          AnnotationDbi_1.42.0   IRanges_2.14.0
[31] S4Vectors_0.18.0       Biobase_2.40.0         BiocGenerics_0.26.0

loaded via a namespace (and not attached):
 [1] survivalROC_1.0.3           Category_2.46.0
 [3] breastCancerUNT_1.17.0      bitops_1.0-6
 [5] matrixStats_0.53.1          bit64_0.9-7
```

```
 [7] progress_1.1.2              httr_1.3.1
 [9] Rgraphviz_2.24.0            tools_3.5.0
[11] R6_2.2.2                    vegan_2.5-1
[13] KernSmooth_2.23-15          DBI_0.8
[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] bit_1.1-12                  compiler_3.5.0
[23] formatR_1.5                 breastCancerNKI_1.17.0
[25] ggdendro_0.1-20             labeling_0.3
[27] scales_0.5.0                genefilter_1.62.0
[29] RBGL_1.56.0                 stringr_1.3.0
[31] digest_0.6.15               breastCancerVDX_1.17.0
[33] AnnotationForge_1.22.0      pkgconfig_2.0.1
[35] rlang_0.2.0                 RSQLite_2.1.0
[37] SuppDists_1.1-9.4           GOstats_2.46.0
[39] RCurl_1.95-4.10             magrittr_1.5
[41] GO.db_3.6.0                 futile.logger_1.4.3
[43] Matrix_1.2-14               Rcpp_0.12.16
[45] munsell_0.4.3               stringi_1.1.7
[47] RJSONIO_1.3-0               org.Hs.eg.db_3.6.0
[49] plyr_1.8.4                  breastCancerUPP_1.17.0
[51] grid_3.5.0                  blob_1.1.1
[53] breastCancerTRANSBIG_1.17.0 lattice_0.20-35
[55] cowplot_0.9.2               splines_3.5.0
[57] locfit_1.5-9.1              pillar_1.2.2
[59] reshape2_1.4.3              futile.options_1.0.1
[61] lambda.r_1.2.2              data.table_1.10.4-3
[63] bootstrap_2017.2            gtable_0.2.0
[65] amap_0.8-14                 assertthat_0.2.0
[67] ggplot2_2.2.1               xtable_1.8-2
[69] class_7.3-14                tibble_1.4.2
[71] snow_0.4-2                  memoise_1.1.0
[73] lava_1.6.1                  breastCancerMAINZ_1.17.0
```

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/