# `TNBC.CMS`: prediction of TNBC consensus molecular subtype

Doyeong Yu[1], Jihyun Kim[1], In Hae Park[2], and Charny Park[1]

[1]Clinical Genomics Analysis Branch, Research Institute, National Cancer Center, Gyeonggi-do, Republic of Korea
[2]Center for Breast Cancer, Hospital, National Cancer Center, Gyeonggi-do, Republic of Korea

March 10, 2019

## Contents

# 1 Introduction

TNBC.CMS is a package for molecular subtype classification of triple-negative breast cancer (TNBC). While various classification strategies have been proposed, absence of precise subtype classifier was a limitation of patient diagnosis and TNBC studies. Our machine learning-based classifier model was derived from gene expression profiles of 957 TNBC patients. The TNBC.CMS package classifies patients into four consensus molecular subtypes (CMS): mesenchymal-like (MSL), immunomodulatory (IM), luminal AR (LAR) and stem-like (SL). It also provides a summary of genomic and clinical characteristics including survival, hazard ratio, pathway activities and drug responses.

# 2 Loading package and dataset for case studies

In this vignette, we walk through a case study of the breast cancer microarray dataset GSE25055 [1] to demonstrate the practical use of our package. The GSE25055 dataset was obtained from the `curatedBreastData` package. We filtered out samples which seemed to be positive for ER, PR, and HER2 based on immunohistochemistry results and the distribution of gene expression.

First, we load the package and the processed expression data. The dataset is contained in a `SummarizedExperiment` object, which includes expression profiles of 4,746 genes and 73 samples. Note that rows and columns correspond to genes and samples, and row names must be gene symbols.

```
library(TNBC.CMS)
data("GSE25055")
dim(assays(GSE25055)[[1]])

## [1] 4746   73

assays(GSE25055)[[1]][1:5, 1:5]

##       615397 615396 615394 615393 615392
## HBA1      15     16     16     15     13
## HBB       15     16     16     15     12
## B2M       13     14     14     13     12
## FTH1      12     13     12     13     12
## HLA-B     12     15     14     12     11
```

# 3 Case study: CMS classification

The `predictCMS` function assigns consensus molecular subtypes to TNBC samples based on input matrix or `SummarizedExperiment` object. If input is a `SummarizedExperiment` object, the first element in the assays list should be a matrix of gene expression. In any case, gene expression profiles should neither be scaled nor log-transformed. Class probabilities can be retrieved by accessing the `probabilities` attribute.

```
predictions <- predictCMS(GSE25055)
table(predictions)

## predictions
## MSL  IM LAR  SL
##  12  14  23  24

head(attr(predictions, "probabilities"))

##           MSL     IM    LAR    SL
## 615397 0.0050 0.0035 0.0068 0.985
## 615396 0.0264 0.0141 0.9397 0.020
## 615394 0.0563 0.0202 0.0204 0.903
## 615393 0.0010 0.0017 0.9566 0.041
## 615392 0.0021 0.0388 0.7918 0.167
## 615390 0.0292 0.0089 0.0050 0.957
```

# 4 Case study: summary of genomic and clinical characteristics

The `TNBC.CMS` package includes several functions for studying genomic and clinical characteristics of the consensus molecular subtypes. In this section, we apply these functions to the GSE25055 datasets of gene expression and clinical features.

The `computeGES` function calculates signature scores for the following 7 gene expression signatures: EMT (epithelial-mesencymal transition), stromal, immune, microenvironment, stemness, hormone, and CIN (chromosomal instability) [2-6]. For more details about the gene expression signatures, please see the manual page for `computeGES` function.

As shown in Figure 1, this function also draws boxplots of signature scores with p-values of comparison among the subtypes. Stromal, immune, hormone, and stemness scores are significantly higher in MSL, IM, LAR, and SL subgroups than in other subgroups, respectively.

```
resultGES <- computeGES(expr = GSE25055, pred = predictions,
                        rnaseq = FALSE)
```

```
resultGES[,1:4]

##                   615397  615396  615394  615393
## EMT                 6.04    6.19    6.35    5.73
## Stromal          -605.49 -642.15 -552.41 -929.42
## Immune            167.10  282.41  293.84 -141.40
## Microenvironment    0.30    0.26    0.34    0.22
## Stemness            0.55    0.30    0.40    0.61
## Hormone             4.96    4.95    5.59    5.45
## CIN                 7.39    6.76    7.23    7.52
```

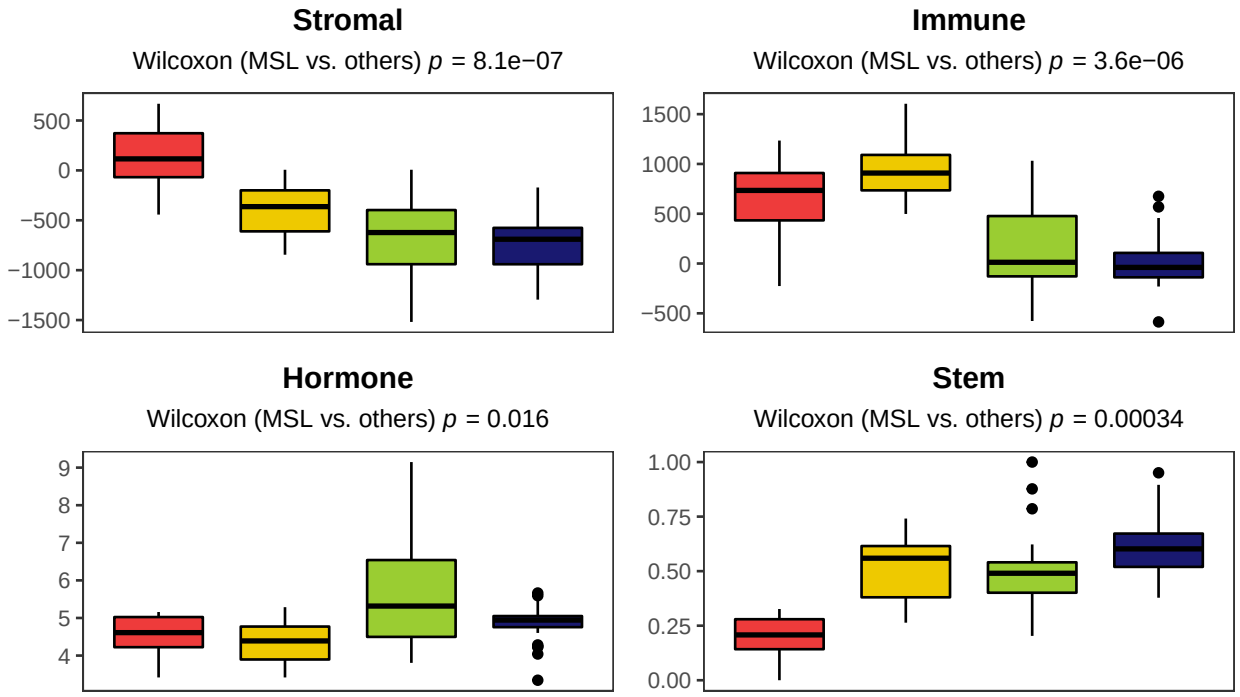


Figure 1: Gene expression signature scores

The `performGSVA` function performs gene set variation analysis on gene sets and produces a heatmap representing GSVA enrichment scores [7]. If gene sets are not given, the hallmark pathway gene sets are used [8]. The user can also choose a kernel for estimating the cumulative distribution function of expression values by setting the `gsva.kcdf` argument, which is set to `"Gaussian"` by default. If expression levels are integer counts, the `"Poisson"` is recommended.

Figure 2 shows differential activation of the hallmark pathways across the subtypes. The MSL subtype has high levels of EMT and P53 pathway activation, and the IM subtype shows the high interferon gamma response. AR and ER response pathways are highly activated in the LAR subtpe, and the expression of cell cycle associated pathway genes is up-regulated in the SL subtype.

```
resultGSVA <- performGSVA(expr = GSE25055, pred = predictions,
                          gene.set = NULL)
head(resultGSVA[,1:4])

##                                    615397 615396 615394 615393
## KRAS_SIGNALING                     -0.021  0.178   0.29  -0.16
## COAGULATION                        -0.353  0.217   0.24  -0.33
## EPITHELIAL_MESENCHYMAL_TRANSITION -0.181  0.039   0.38  -0.18
## MYOGENESIS                         -0.057  0.110   0.35  -0.26
## ANGIOGENESIS                        0.039 -0.120   0.45  -0.31
## APICAL_JUNCTION                    -0.127  0.108   0.34  -0.16
```

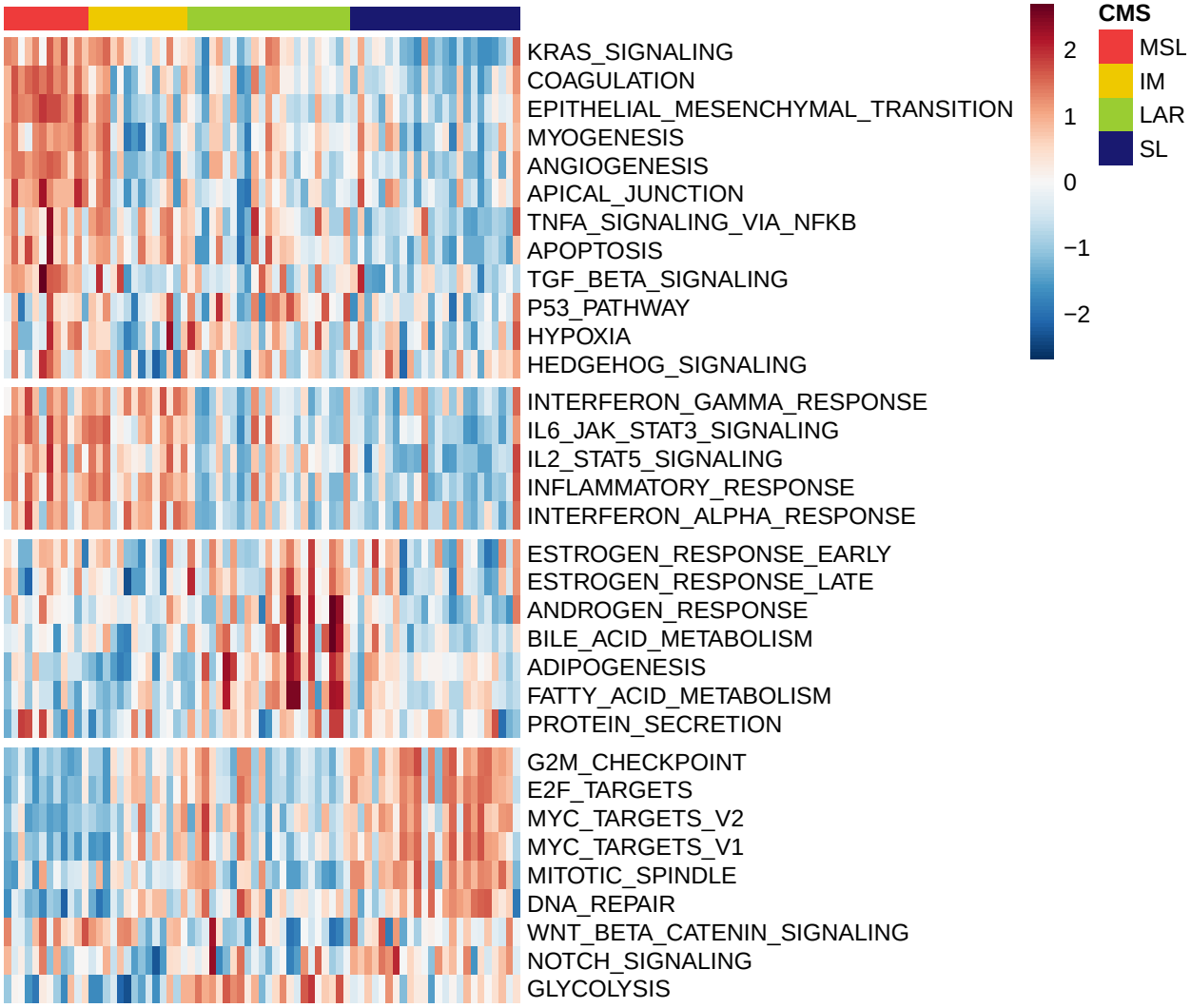


Figure 2: GSVA enrichment scores

The `TNBC.CMS` package provides two functions for survival analysis: `plotKM` and `plotHR`. Here, we use the survival data from the GSE25055 dataset to study the association between overall survival and the consensus molecular subtypes. The survival data is also included in the `SummarizedExperiment` object and can be accessed using the `colData` function.

The `plotKM` function produces a Kaplan-Meier curve for each consensus molecular subtype like Figure 3. The SL group showed the worst prognosis, which is consistent with our previous study.

```
time <- colData(GSE25055)$DFS.month
event <- colData(GSE25055)$DFS.status
plotKM(pred = predictions, time = time, event = event)
```

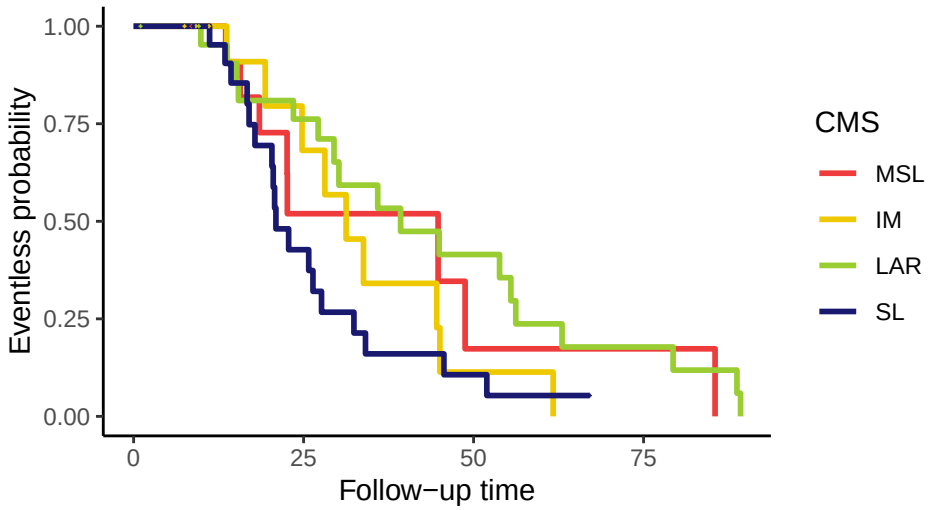


Figure 3: Overall survival

The `plotHR` produces a forest plot of hazard ratios for genes that the user provides. For each input gene, samples are divided into high and low groups based on its expression level and the 95% confidence interval for the hazard ratio is calculated. We selected 10 genes significantly associated with overall survival and generated a forest plot (Figure 4).

```
library(survival)

#Test for difference of survival between low and high expression groups
surv <- Surv(time, event)
GSE25055.exprs <- assays(GSE25055)[[1]]
chisq <- apply(GSE25055.exprs, 1, function(x) survdiff(surv ~ (x > median(x)))$chisq)
pval <- 1 - pchisq(chisq, 1)

#Select 10 genes with lowest p-values for the log-rank test
gs <- names(sort(pval)[1:10])
gs

##  [1] "RECK"   "RELN"   "EHD4"   "PRRX2"  "FOLR1"  "UGCG"   "GOT2"   "PRPF39"
##  [9] "OPHN1"  "CPE"

plotHR(expr = GSE25055, gene.symbol = gs, pred = predictions, time = time,
       event = event, by.subtype = FALSE)
```

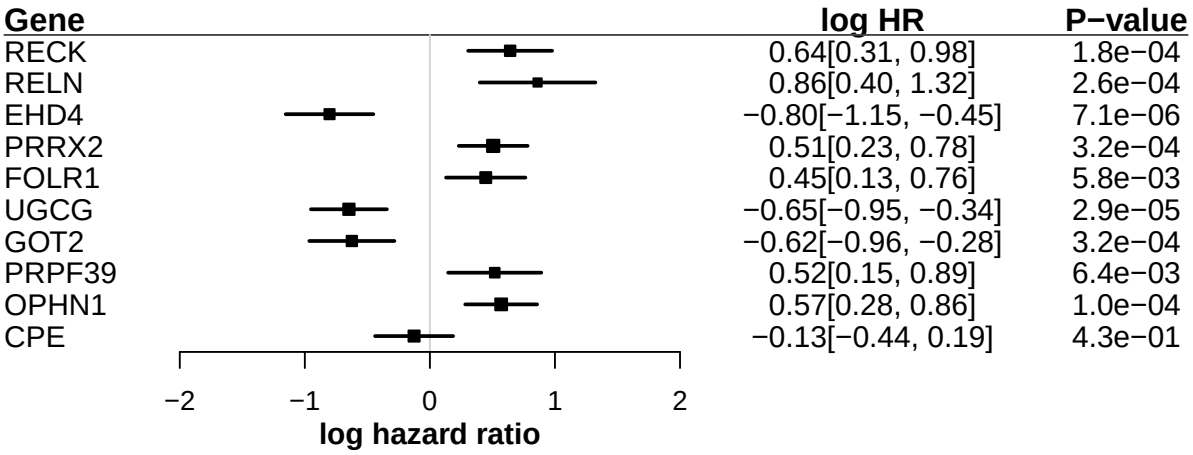


Figure 4: Forest plot of hazard ratios

Also, as shown in Figure 5, subtype-specific hazard ratios for genes of interest can be computed by setting the `by.subtype` argument.

```
plotHR(expr = GSE25055, gene.symbol = gs[1:4], pred = predictions,
       time = time, event = event, by.subtype = TRUE)
```

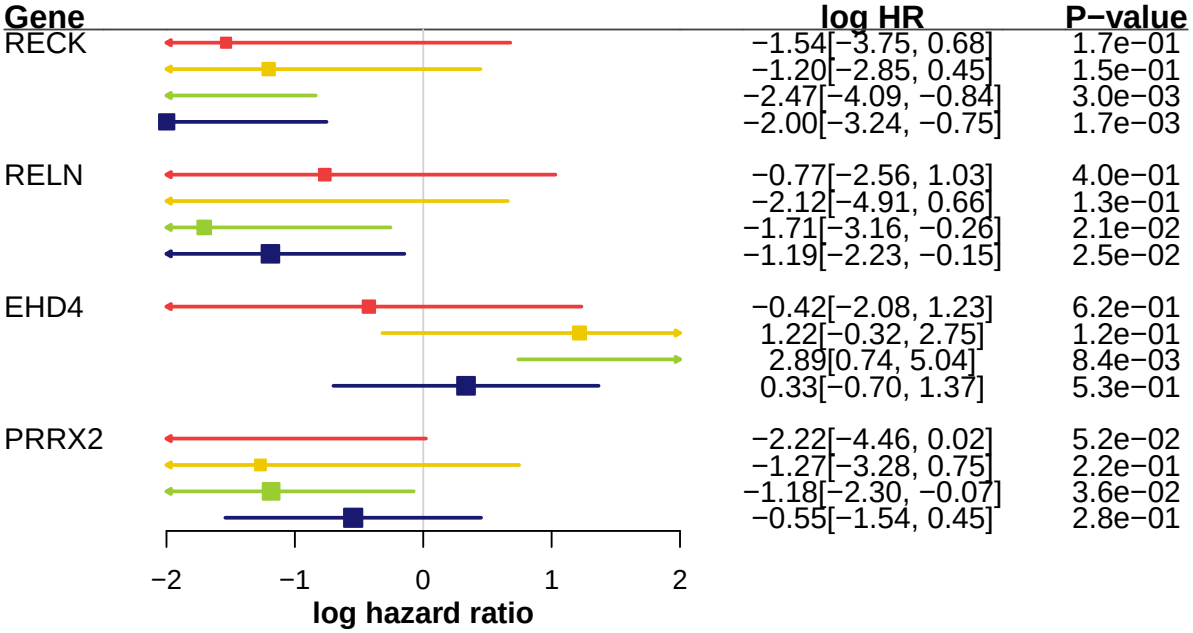


Figure 5: Forest plot of subtype-specific hazard ratios

# 5 Case study: drug response investigation

The `TNBC.CMS` package also provides a function for predicting drug responses. The `computeDS` function computes drug signature scores for the corresponding gene sets in the MSigDB CGP (chemical and genetic perturbations) collection [9] and draws a heatmap of the signature scores. Drug signature scores are calculated as the difference between the average expression values of gene sets associated with drug response and resistance. The higher a signature score is, the more likely a patient is to be responsive. The user can provide their own gene sets via the `gene.set` argument. Note that names of gene sets must follow the format of [DRUG NAME]_[RESPONSE/RESISTANCE]_[UP/DN] (e.g. CISPLATIN_RESISTANCE_UP).

Figure 6 shows a heatmap of drug signature scores for each sample. The MSL and SL subtypes appear to be resistant to dasatinib and doxorubicin, respectively. Also, the IM and LAR subtypes show higher levels of signature scores for androgen agonist and SB216763 (an inhibitor of GSK3B) than other subtypes, respectively.

```
resultDS <- computeDS(expr = GSE25055, pred = predictions)
head(resultDS[,1:4])

##              615397 615396 615394 615393
## APLIDIN       -0.89  -0.47  -0.81  -0.85
## CISPLATIN      0.96   0.69   1.26   0.25
## DASATINIB      0.55   0.70  -0.12   0.72
## FORSKOLIN      6.48   6.74   6.46   6.68
## IMATINIB      -6.35  -5.09  -5.48  -5.28
## TROGLITAZONE   6.88   6.72   6.72   6.68
```

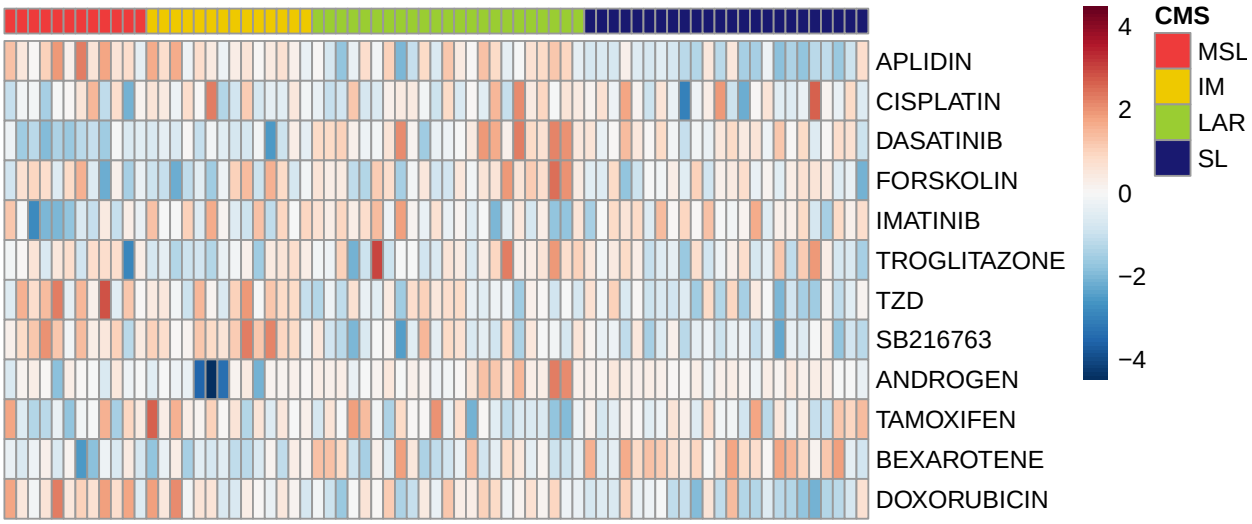


Figure 6: Drug signature scores

# 6 Saving results

For future analysis, it is useful to save the results of subtype assignment and characterization into a data frame and save it into a text file.

```
dfCMS <- data.frame(row.names = colnames(GSE25055.exprs), CMS = predictions, t(resultGES),
                    t(resultDS), stringsAsFactors = FALSE)
head(dfCMS)
```

```
##        CMS EMT Stromal Immune Microenvironment Stemness Hormone CIN APLIDIN
## 615397  SL 6.0    -605    167             0.30     0.55     5.0 7.4   -0.89
## 615396 LAR 6.2    -642    282             0.26     0.30     4.9 6.8   -0.47
## 615394  SL 6.4    -552    294             0.34     0.40     5.6 7.2   -0.81
## 615393 LAR 5.7    -929   -141             0.22     0.61     5.5 7.5   -0.85
## 615392 LAR 5.6   -1519   -560             0.17     1.00     4.2 8.2   -1.37
## 615390  SL 6.2    -607     32             0.27     0.59     4.6 7.5   -0.87
##        CISPLATIN DASATINIB FORSKOLIN IMATINIB TROGLITAZONE TZD SB216763
## 615397      0.96      0.55       6.5     -6.4          6.9 7.2      6.1
## 615396      0.69      0.70       6.7     -5.1          6.7 5.3      6.3
## 615394      1.26     -0.12       6.5     -5.5          6.7 6.4      6.0
## 615393      0.25      0.72       6.7     -5.3          6.7 6.3      5.8
## 615392      0.36      0.82       6.7     -4.9          7.0 5.5      5.7
## 615390      0.77      0.14       6.8     -5.0          6.9 7.5      6.0
##        ANDROGEN TAMOXIFEN BEXAROTENE DOXORUBICIN
## 615397      3.9      -6.0       0.33        -6.9
## 615396      4.0      -6.3       0.14        -6.3
## 615394      3.8      -6.3      -1.03        -6.7
## 615393      3.9      -5.9       0.15        -7.0
## 615392      4.0      -6.1      -0.13        -7.6
## 615390      4.2      -6.2      -0.88        -6.8
```

```
write.table(dfCMS, file = "GSE25055_CMS.txt")
```

# 7 Session Info

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.13-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.13-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] grid      parallel  stats4    stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
##  [1] survival_3.2-11             ggpubr_0.4.0
##  [3] ggplot2_3.3.3               TNBC.CMS_1.8.0
##  [5] SummarizedExperiment_1.22.0 Biobase_2.52.0
##  [7] GenomicRanges_1.44.0        GenomeInfoDb_1.28.0
##  [9] IRanges_2.26.0              S4Vectors_0.30.0
## [11] BiocGenerics_0.38.0         MatrixGenerics_1.4.0
## [13] matrixStats_0.58.0          quadprog_1.5-8
## [15] e1071_1.7-6                 knitr_1.33
##
## loaded via a namespace (and not attached):
##   [1] colorspace_2.0-1            ggsignif_0.6.1
##   [3] ellipsis_0.3.2              class_7.3-19
##   [5] rio_0.5.26                  XVector_0.32.0
##   [7] proxy_0.4-25                farver_2.1.0
##   [9] bit64_4.0.5                 AnnotationDbi_1.54.0
##  [11] fansi_0.4.2                 splines_4.1.0
##  [13] R.methodsS3_1.8.1           sparseMatrixStats_1.4.0
##  [15] cachem_1.0.5                broom_0.7.6
##  [17] annotate_1.70.0             png_0.1-7
##  [19] R.oo_1.24.0                 pheatmap_1.0.12
##  [21] graph_1.70.0                HDF5Array_1.20.0
##  [23] compiler_4.1.0              httr_1.4.2
##  [25] backports_1.2.1             assertthat_0.2.1
##  [27] Matrix_1.3-3                fastmap_1.1.0
##  [29] BiocSingular_1.8.0          tools_4.1.0
##  [31] rsvd_1.0.5                  gtable_0.3.0
##  [33] glue_1.4.2                  GenomeInfoDbData_1.2.6
##  [35] dplyr_1.0.6                 Rcpp_1.0.6
##  [37] carData_3.0-4               cellranger_1.1.0
##  [39] vctrs_0.3.8                 Biostrings_2.60.0
##  [41] rhdf5filters_1.4.0          DelayedMatrixStats_1.14.0
##  [43] xfun_0.23                   stringr_1.4.0
##  [45] openxlsx_4.2.3              beachmat_2.8.0
##  [47] lifecycle_1.0.0             irlba_2.3.3
##  [49] rstatix_0.7.0               XML_3.99-0.6
##  [51] zlibbioc_1.38.0             scales_1.1.1
##  [53] hms_1.1.0                   rhdf5_2.36.0
##  [55] RColorBrewer_1.1-2          SingleCellExperiment_1.14.0
##  [57] curl_4.3.1                  memoise_2.0.0
##  [59] reshape_0.8.8               stringi_1.6.2
##  [61] RSQLite_2.2.7               GSVA_1.40.0
##  [63] highr_0.9                   ScaledMatrix_1.0.0
##  [65] checkmate_2.0.0             zip_2.1.1
##  [67] BiocParallel_1.26.0         rlang_0.4.11
##  [69] pkgconfig_2.0.3             bitops_1.0-7
##  [71] pracma_2.3.3                evaluate_0.14
##  [73] lattice_0.20-44             purrr_0.3.4
##  [75] Rhdf5lib_1.14.0             labeling_0.4.2
```

```
##  [77] bit_4.0.4                  tidyselect_1.1.1
##  [79] GSEABase_1.54.0            GGally_2.1.1
##  [81] plyr_1.8.6                 magrittr_2.0.1
##  [83] R6_2.5.0                   generics_0.1.0
##  [85] DelayedArray_0.18.0        DBI_1.1.1
##  [87] withr_2.4.2                pillar_1.6.1
##  [89] haven_2.4.1                foreign_0.8-81
##  [91] forestplot_1.10.1          KEGGREST_1.32.0
##  [93] abind_1.4-5                RCurl_1.98-1.3
##  [95] tibble_3.1.2               crayon_1.4.1
##  [97] car_3.0-10                 utf8_1.2.1
##  [99] readxl_1.3.1               data.table_1.14.0
## [101] blob_1.2.1                 forcats_0.5.1
## [103] digest_0.6.27              xtable_1.8-4
## [105] tidyr_1.1.3                R.utils_2.10.1
## [107] munsell_0.5.0
```

# 8 References


[1] Hatzis, C. et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer *JAMA*, **305**, 1873–81.

[2] Tan, T.Z. et al. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO molecular medicine*, **6**, 1279–93.

[3] Yoshihara, K. et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, **4**, 2612.

[4] Aran, D. et al. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, **18**, 220.

[5] Malta, T.M. et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, **173**, 338–354.

[6] Carter, S.L. et al. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature genetics*, **38**, 1043.

[7] Hanzelmann, S. et al. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, **14**, 7.

[8] Liberzon, A. et al. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, **1**, 417–425.

[9] Liberzon, A. et al. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–40.