

# Package ‘DeSousa2013’

September 23, 2025

**Type** Package

**Title** Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion

**Version** 1.45.0

**Date** 2013-04-10

**Author** Xin Wang <Xin.Wang@cruk.cam.ac.uk>

**Maintainer** Xin Wang <Xin.Wang@cruk.cam.ac.uk>

**Depends** R (>= 2.15),

**Imports** affy, frma, frmaTools, hgu133plus2.db, hgu133plus2frmavecs, sva, rgl, ConsensusClusterPlus, cluster, siggenes, ROCR, pamr, survival, gplots, AnnotationDbi, Biobase

**Description** This package reproduces the main pipeline to analyze the AMC-AJCCII-90 microarray data set in De Sousa et al. accepted by Nature Medicine in 2013.

**License** Artistic-2.0

**biocViews** CancerData, ColonCancerData, MicroarrayData

**Collate** preprocessing.R consensusClustering.R featureSelection.R classification.R prognosis.R figures.R pipeline.R

**LazyLoad** yes

**git\_url** <https://git.bioconductor.org/packages/DeSousa2013>

**git\_branch** devel

**git\_last\_commit** 1658098

**git\_last\_commit\_date** 2025-04-15

**Repository** Bioconductor 3.22

**Date/Publication** 2025-09-23

## Contents

|                           |   |
|---------------------------|---|
| buildClassifier . . . . . | 2 |
| compGapStats . . . . .    | 3 |
| conClust . . . . .        | 4 |
| CRCPipeLine . . . . .     | 5 |
| data-AMC . . . . .        | 7 |
| filterDiffGenes . . . . . | 8 |

|                          |    |
|--------------------------|----|
| filterSamples . . . . .  | 9  |
| findDiffGenes . . . . .  | 10 |
| geneExpPre . . . . .     | 11 |
| getCentroids . . . . .   | 12 |
| pamClassify . . . . .    | 13 |
| pbs2unigenes . . . . .   | 14 |
| progAMC . . . . .        | 15 |
| selTopVarGenes . . . . . | 16 |

|              |           |
|--------------|-----------|
| <b>Index</b> | <b>18</b> |
|--------------|-----------|

---

|                 |                                                 |
|-----------------|-------------------------------------------------|
| buildClassifier | <i>Build a gene expression based classifier</i> |
|-----------------|-------------------------------------------------|

---

## Description

This function employs PAM to build a gene expression based classifier.

## Usage

```
buildClassifier(sigMat, clus.f, nfold=10, nboot=100)
figPAMCV(err)
```

## Arguments

|        |                                                                                                                                                                                          |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| sigMat | a matrix of median centered expression values of selected genes (by function <a href="#">filterDiffGenes</a> ) for selected cancer samples (by function <a href="#">filterSamples</a> ). |
| clus.f | a numeric vector of cluster labels for selected cancer samples.                                                                                                                          |
| nfold  | an integer value specifying the fold of cross validation.                                                                                                                                |
| nboot  | an integer value specifying the number of bootstraps.                                                                                                                                    |
| err    | a matrix of cross validation error rates for different shrinkage thresholds.                                                                                                             |

## Details

The expression data of the retained most predictive genes were trained by PAM to build a robust classifier. To select the optimal threshold for centroid shrinkage, we performed 10-fold cross-validation over a range of shrinkage thresholds for 1000 iterations, and selected the one yielding a good performance (error rate < 2%) with the least number of genes.

## Value

This function will return a list including signature (a character vector of signature genes), pam.rslt (an list of training results returned by [pamr.train](#)), thresh (the selected shrinkage threshold), err (a matrix of cross validation error rates for different shrinkage thresholds), cents (a numeric matrix of PAM centroids for three subtypes).

## Author(s)

Xin Wang <xw264@cam.ac.uk>

## References

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Tibshirani, Robert and Hastie, Trevor and Narasimhan, Balasubramanian and Chu, Gilbert (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 99(10), 6567-6572.

## See Also

[filterSamples](#), [findDiffGenes](#), [filterDiffGenes](#)

## Examples

```
data(diffGenes.f, package="DeSousa2013")
data(silh, package="DeSousa2013")
sigMat <- sdat.f[diffGenes.f, names(clus.f)]
##set a small number of bootstraps, only for testing the function
classifier <- buildClassifier(sigMat, clus.f, nboot=10)
data(classifier, package="DeSousa2013")
figPAMCV(err)
```

---

compGapStats

*Computing Gap statistics to identify the optimal number of subtypes*

---

## Description

Compute Gap statistics to identify the optimal number of subtypes

## Usage

```
compGapStats(ge.CRC, ntops=c(2, 4, 8, 12, 16, 20)*1000, K.max=6, nboot=100)
figGAP(gapsmat, gapsSE)
```

## Arguments

|         |                                                                                                                    |
|---------|--------------------------------------------------------------------------------------------------------------------|
| ge.CRC  | a numeric matrix of expression data of genes expressed in at least one sample.                                     |
| ntops   | an integer vector of top variable genes, measured by MAD (median absolute deviation).                              |
| K.max   | an integer value specifying the maximal number of clusters to compute GAP statistics.                              |
| nboot   | an integer value specifying the number of bootstraps, which is an argument B of function <a href="#">clusGap</a> . |
| gapsmat | a numeric matrix of GAP statistics.                                                                                |
| gapsSE  | standard errors of means of the GAP statistics.                                                                    |

## Details

GAP statistic is a popular method to estimate the number of clusters in a set of data by comparing the change in observed and expected within-cluster dispersion. To identify the optimal number of clusters, GAP statistic can be computed for  $k=1$  to  $K.max$  with  $nboot$  bootstraps for  $ntops$  top variable genes in the AMC data set.

The function `figGAP` is designed to visualize GAP curves.

**Value**

This function will return a list including gapsmat (a numeric matrix of GAP statistics) and gapsSE (standard errors of means of the GAP statistics).

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Tibshirani, Robert and Walther, Guenther and Hastie, Trevor (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

**Examples**

```
data(ge.CRC, package="DeSousa2013")
ge.CRC <- ge.all[selPbs, ]
gaps <- compGapStats(ge.CRC, ntops=c(2, 4)*1000, K.max=6, nboot=10)
figGAP(gaps$gapsmat, gaps$gapsSE)
```

---

conClust

*Consensus clustering*


---

**Description**

This function performs consensus clustering for the AMC-AJCCII-90 data set.

**Usage**

```
conClust(sdat, maxK=12, reps=1000, savepath = ".")
```

**Arguments**

|          |                                                                                             |
|----------|---------------------------------------------------------------------------------------------|
| sdat     | a matrix of median centered expression values of top variable probesets.                    |
| maxK     | an integer value specifying the maximal number of clusters to perform consensus clustering. |
| reps     | an integer value specifying resampling times.                                               |
| savepath | the path to a directory where figures will be saved.                                        |

**Details**

Using the most variable probesets (MAD>0.5), we performed hierarchical clustering with agglomerative average linkage to cluster these samples. Consensus clustering was employed, with 1000 iterations and 0.98 subsampling ratio, to assess the clustering stability.

**Value**

This function will return clus, which is a numeric vector of cluster labels named by colon cancer samples.

**Author(s)**

Xin Wang <ww264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Monti, Stefano and Tamayo, Pablo and Mesirov, Jill and Golub, Todd (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, 52(1), 91-118.

**Examples**

```
library(ConsensusClusterPlus)
data(dat, package="DeSousa2013")
##set a small number of bootstraps only for testing the function
clus <- conClust(sdat, maxK=6, reps=10, savepath=".")
```

---

CRCPipeLine

*Pipeline function for De Sousa 2013*

---

**Description**

This function reproduces the main pipeline to analyze the AMC-AJCCII-90 microarray data set in De Sousa et al. accepted by Nature Medicine in 2013.

**Usage**

```
CRCPipeLine(celpath=".", AMC_sample_head, AMC_CRC_clinical, preprocess=FALSE,
gap.ntops = c(2, 4, 8, 12, 16, 20)*1000, gap.K.max = 6, gap.nboot = 100,
MADth=0.5, conClust.maxK=12, conClust.reps=1000, diffG.pvalth=0.01,
diffG.aucth=0.9, savepath=".")
```

**Arguments**

|                  |                                                                                                                                                                            |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| celpath          | the path to a directory where .CEL files of the AMC-AJCCII-90 set are placed.                                                                                              |
| AMC_sample_head  | a data frame include clinical information and mapping information between microarray ids and sample ids (details in <a href="#">AMC_sample_head</a> ).                     |
| AMC_CRC_clinical | a data frame including include clinical information (details in <a href="#">AMC_CRC_clinical</a> ).                                                                        |
| preprocess       | a logical value specifying whether preprocessing of microarrays should be performed. If FALSE (default), preprocessed expression data ge.CRC will be loaded automatically. |
| gap.ntops        | an integer vector of top variable genes, measured by MAD (median absolute deviation).                                                                                      |
| gap.K.max        | an integer value specifying the maximal number of clusters to compute GAP statistics.                                                                                      |
| gap.nboot        | an integer value specifying the number of bootstraps, which is an argument B of function <a href="#">clusGap</a> .                                                         |

|               |                                                                                                 |
|---------------|-------------------------------------------------------------------------------------------------|
| MADth         | an numeric value specifying the cutoff of MAD (median absolute deviation).                      |
| conClust.maxK | an integer value specifying the maximal number of clusters to perform consensus clustering.     |
| conClust.reps | an integer value specifying resampling times.                                                   |
| diffG.pvalth  | a numeric value specifying the fdr cutoff to select differential genes between subtypes by SAM. |
| diffG.aucth   | a numeric value specifying the AUC cutoff.                                                      |
| savepath      | the path to a directory where figures will be saved.                                            |

### Details

The function wraps up the preprocessing, feature selection, classification and subtype clinical characterization steps altogether in a signal function. It reproduces the main results and figures of De Sousa 2013, Nature Medicine.

### Value

This function will save figure of gap statistics, Silhouette information, PAM cross-validation error rates, classification heatmap as well as Kaplan Meier plot to savepath.

### Author(s)

Xin Wang <xw264@cam.ac.uk>

### References

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

### See Also

[geneExpPre](#), [compGapStats](#), [selTopVarGenes](#), [pbs2unigenes](#), [conClust](#), [filterSamples](#), [findDiffGenes](#), [filterDiffGenes](#), [buildClassifier](#), [pamClassify](#), [progAMC](#)

### Examples

```
## Not run:
##This function may take a long time (hours) to finish.
##Please use R-2.15 and corresponding frma package to run the function.
##Please contact the author if there is any confusion
data(AMC, package="DeSousa2013")
CRCPipeLine(celpath=".", AMC_sample_head, AMC_CRC_clinical, savepath=".")

## End(Not run)
```

data-AMC

*Data used in the package DeSousa2013***Description**

See ‘details’ for the description of each data files included in this package.

**Usage**

```
##see example for details
```

**Details**

AMC: AMC\_CRC\_clinical and AMC\_sample\_head include clinical information and mapping information between microarray ids and sample ids, respectively.

ge.CRC: ge.all is a numeric matrix including expression levels of all probesets of the 90 colon cancer samples. selPbs is a character vector of probeset ids that are present in any sample.

gaps: gapsmat is a numeric matrix of GAP statistics. gapsSE is a numeric matrix of standard errors of means of the GAP statistics.

dat: sdat is a matrix of median centered expression values of top variable probesets.

uniGenes: uniGenes is a character vector of unique gene symbols named by probesets.

conClust: clus is a numeric vector of cluster labels named by colon cancer samples.

silh: sdat.f is a matrix of median centered expression values of top variable genes for selected samples. clus.f is a numeric vector of cluster labels for selected cancer samples. silh is an object of class [silhouette](#).

diffGenes: diffGenes is a character vector of differential genes.

diffGenes.f: diffGenes.f is a character vector of most predictive genes.

classifier: signature is a character vector of signature genes. pam.rslt is an list of training results returned by [pamr.train](#). thresh is the selected shrinkage threshold. err is a matrix of cross validation error rates for different shrinkage thresholds. cents is a numeric matrix of PAM centroids for three subtypes.

predAMC: sdat.sig is a numeric matrix of expression values of signature genes for the 90 samples. pred is a numeric matrix of posterior probabilities for samples to be classified to subtypes. clu.pred is a numeric vector of classification labels named by colon cancer samples. nam.ord is a character vector of samples ordered by their classification probabilities for visualization. gclu.f is the result of hierarchical clustering on the expression of signature genes for visualization.

survival: surv is the result of [survfit](#) containing survival curves of the AMC data set. survstats is the result of [survdif](#). data4surv is the data to perform survival analysis.

**Author(s)**

Xin Wang <wx264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

**Examples**

```
data(AMC, package="DeSousa2013")
```

---

|                 |                                                                 |
|-----------------|-----------------------------------------------------------------|
| filterDiffGenes | <i>Select the most predictive genes from differential genes</i> |
|-----------------|-----------------------------------------------------------------|

---

**Description**

This function evaluates AUC (area under ROC curve) to select the most predictive genes from differential genes.

**Usage**

```
filterDiffGenes(sdat.f, clus.f, diffGenes, aucth=0.9)
```

**Arguments**

|           |                                                                                           |
|-----------|-------------------------------------------------------------------------------------------|
| sdat.f    | a matrix of median centered expression values of top variable genes for selected samples. |
| clus.f    | a numeric vector of cluster labels for selected cancer samples.                           |
| diffGenes | a character vector of differential genes.                                                 |
| aucth     | a numeric value specifying the AUC cutoff.                                                |

**Details**

After obtaining differential genes between subtypes, we calculated AUC (area under ROC curve, using package ROCR) to assess each gene's ability to separate one subtype from the others.

**Value**

This function will return `diffGenes.f`, which is a character vector of most predictive genes.

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940-3941.

**See Also**

[findDiffGenes](#)



**Examples**

```
data(uniGenes)
data(silh)
data(diffGenes)
##select randomly part of the whole differential genes only for testing the function
diffGenes <- diffGenes[sample(1:length(diffGenes), 100)]
diffGenes.f <- filterDiffGenes(sdat.f, clus.f, diffGenes, aucth=0.9)
```

filterSamples

*Filter colon cancer samples by Silhouette width***Description**

This function computes Silhouette widths for the 90 colon cancer samples.

**Usage**

```
filterSamples(sdat, uniGenes, clus)
figSilh(silh)
```

**Arguments**

|          |                                                                          |
|----------|--------------------------------------------------------------------------|
| sdat     | a matrix of median centered expression values of top variable probesets. |
| uniGenes | a character vector of unique gene symbols named by probesets.            |
| clus     | a numeric vector of cluster labels named by colon cancer samples.        |
| silh     | an object of class <a href="#">silhouette</a> .                          |

**Details**

Silhouette width was computed to identify the most representative samples within each cluster. Samples with positive silhouette width were retained to build the PAM classifier.

**Value**

This function will return a list including `sdat.f` (a matrix of median centered expression values of top variable genes for selected samples), `clus.f` (a numeric vector of cluster labels for selected cancer samples) and `silh` (an object of class [silhouette](#)).

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Rousseeuw, Peter J (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of computational and applied mathematics, 20, 53-65.

**See Also**

[selTopVarGenes](#)

**Examples**

```
data(uniGenes, package="DeSousa2013")
data(dat, package="DeSousa2013")
data(conClust, package="DeSousa2013")
samp.f <- filterSamples(sdat, uniGenes, clus)
figSilh(samp.f$silh)
```

---

findDiffGenes

---

*Search for differential genes between subtypes*


---

**Description**

This function employs SAM to search for differential genes between subtypes

**Usage**

```
findDiffGenes(sdat.f, clus.f, pvalth=0.01)
```

**Arguments**

|        |                                                                                                 |
|--------|-------------------------------------------------------------------------------------------------|
| sdat.f | a matrix of median centered expression values of top variable genes for selected samples.       |
| clus.f | a numeric vector of cluster labels for selected cancer samples.                                 |
| pvalth | a numeric value specifying the fdr cutoff to select differential genes between subtypes by SAM. |

**Details**

In this function, we find differential genes between each two of three subtypes and take the unique genes of combined differential genes altogether.

**Value**

This function will return diffGenes, which is a character vector of differential genes.

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Tusher, Virginia Goss and Tibshirani, Robert and Chu, Gilbert (2001). Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98(9), 5116-5121.

**See Also**

[filterDiffGenes](#)

## Examples

```
data(uniGenes, package="DeSousa2013")
data(conClust, package="DeSousa2013")
data(silh, package="DeSousa2013")
##select randomly part of the whole data set only for testing the function
diffGenes <- findDiffGenes(sdat.f[sample(1:nrow(sdat.f), 500), ], clus.f, pvalth=0.01)
```

---

geneExpPre

*Preprocessing of the AMC-AJCCII-90 microarray data*

---

## Description

Preprocessing of the AMC-AJCCII-90 microarray data

## Usage

```
geneExpPre(celpath, AMC_sample_head)
```

## Arguments

**celpath** the path to a directory where .CEL files of the AMC-AJCCII-90 set are placed.

**AMC\_sample\_head** a data frame include clinical information and mapping information between microarray ids and sample ids (details in [AMC\\_sample\\_head](#)).

## Details

This function reproduces the preprocessing of the AMC-AJCCII-90 colon cancer microarrays, together with 13 adenomas and 6 normal samples. The microarrays of the 90 cancer samples and normal samples were generated in one batch, and the adenoma samples in a different batch. These two batches were first normalized separately by [frma](#), and then corrected for batch effect using [ComBat](#). The [barcode](#) function is used to call present probesets in the two batches, respectively. Finally, probesets present in any one of the cancer, adenoma or normal samples are selected for the following analysis.

## Value

This function will return a list including `ge.all` (a numeric matrix including expression levels of all probesets of the 90 colon cancer samples) and `selPbs` (a character vector of probeset ids that are present in any sample).

## Author(s)

Xin Wang <[xw264@cam.ac.uk](mailto:xw264@cam.ac.uk)>

## References

- De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted
- McCall, Matthew N and Bolstad, Benjamin M and Irizarry, Rafael A (2010). Frozen robust multi-array analysis (fRMA). *Biostatistics*, 11(2), 242-253.
- McCall, Matthew N and Uppal, Karan and Jaffee, Harris A and Zilliox, Michael J and Irizarry, Rafael A (2011). The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl 1), D1011-D1015.
- Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* 8(1):118-127.

## Examples

```
## Not run:
##Please make sure that all colon cancer, adenomas as well as normal samples
##have been collected from GEO before running this function.
##This function may take a long time (> half an hour) to finish.
##Please use R-2.15 and corresponding frma package to run the function.
##Different versions of dependent microarray data preprocessing packages and annotation
##packages may give slightly different results.
##Please contact the author if there is any confusion
data("AMC")
ge.pre <- geneExpPre(celpath=".", AMC_sample_head)

## End(Not run)
```

---

|              |                                               |
|--------------|-----------------------------------------------|
| getCentroids | <i>Retrieve centroids of a PAM classifier</i> |
|--------------|-----------------------------------------------|

---

## Description

This function retrieves centroids of a PAM classifier

## Usage

```
getCentroids(fit, data, threshold)
```

## Arguments

|           |                                                          |
|-----------|----------------------------------------------------------|
| fit       | the fitting result of a PAM classifier to training data. |
| data      | the training data used for building the PAM classifier   |
| threshold | the threshold for PAM classification                     |

## Details

This is an internal function called by [buildClassifier](#) to retrieve the centroids of PAM classifier built using the AMC-AJCCII-90 data set.

## Value

This function will return a numeric matrix of centroids of the PAM classifier.

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

Tibshirani, Robert and Hastie, Trevor and Narasimhan, Balasubramanian and Chu, Gilbert (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 99(10), 6567-6572.

**See Also**

[getCentroids](#)

---

|             |                                    |
|-------------|------------------------------------|
| pamClassify | <i>Classifying the AMC samples</i> |
|-------------|------------------------------------|

---

**Description**

These functions classify and visualize the 90 AMC colon cancer samples.

**Usage**

```
pamClassify(datsel, signature, pam.rslt, thresh, postRth=1)
figClassify(AMC_CRC_clinical, pred, clu.pred, sdat.sig, gclu.f, nam.ord)
```

**Arguments**

|                  |                                                                                                       |
|------------------|-------------------------------------------------------------------------------------------------------|
| datsel           | a numeric value specifying the cutoff of posterior odds in favor of optimal subtype to the other two. |
| signature        | a character vector of signature genes.                                                                |
| pam.rslt         | an list of training results returned by <a href="#">pamr.train</a> .                                  |
| thresh           | the selected shrinkage threshold.                                                                     |
| postRth          | a numeric value specifying the cutoff of posterior odds in favor of optimal subtype to the other two. |
| AMC_CRC_clinical | a data frame including include clinical information (details in <a href="#">AMC_CRC_clinical</a> ).   |
| pred             | a numeric matrix of posterior probabilities for samples to be classified to subtypes.                 |
| clu.pred         | a numeric vector of classification labels named by colon cancer samples.                              |
| sdat.sig         | a numeric matrix of expression values of signature genes for the 90 samples.                          |
| gclu.f           | the result of hierarchical clustering on the expression of signature genes for visualization.         |
| nam.ord          | a character vector of samples ordered by their classification probabilities for visualization.        |

## Details

The function `pamClassify` classifies the 90 AMC colon cancer samples using the gene expression based classifier built by PAM. The function `figClassify` generates classification results with a heatmap of median centred expression values across samples, a track indicating classification probabilities, as well as a track indicating relapse of patients.

## Value

This function `pamClassify` will return a list including `sdat.sig` (a numeric matrix of expression values of signature genes for the 90 samples), `pred` (a numeric matrix of posterior probabilities for samples to be classified to subtypes), `nam.ord` (a character vector of samples ordered by their classification probabilities for visualization), `gclu.f` (result of hierarchical clustering on the expression of signature genes for visualization).

## Author(s)

Xin Wang <xw264@cam.ac.uk>

## References

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

## See Also

[buildClassifier](#)

## Examples

```
data(AMC, package="DeSousa2013")
data(dat, package="DeSousa2013")
data(uniGenes, package="DeSousa2013")
data(diffGenes.f, package="DeSousa2013")
data(classifier, package="DeSousa2013")
datsel <- sdat[names(uniGenes), ]
rownames(datsel) <- uniGenes
datsel <- datsel[diffGenes.f, ]
pamcl <- pamClassify(datsel, signature, pam.rslt, thresh, postRth=1)
figClassify(AMC_CRC_clinical, pamcl$pred, pamcl$clu.pred, pamcl$sdat.sig,
pamcl$gclu.f, pamcl$nam.ord)
```

---

pbs2unigenes

*Generate a mapping file between probesets and unique gene symbols*

---

## Description

This function takes the expression data of top variable (MAD) genes to generate a mapping file between probesets and unique gene symbols.

## Usage

```
pbs2unigenes(ge.CRC, sdat)
```

**Arguments**

`ge.CRC` a numeric matrix of expression data of genes expressed in at least one sample.  
`sdat` a matrix of median centered expression values of top variable probesets.

**Details**

To facilitate the use of the classifier on other platforms, we collapsed the expression levels for probesets to genes. For each gene the probeset with highest overall expression was selected.

**Value**

This function will save `uniGenes` (a character vector of unique gene symbols named by probesets) to a file called `uniGenes.RData` in the current working directory.

**Author(s)**

Xin Wang <[xw264@cam.ac.uk](mailto:xw264@cam.ac.uk)>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

**See Also**

[selTopVarGenes](#)

**Examples**

```
library(hgu133plus2.db)
data(ge.CRC, package="DeSousa2013")
data(dat, package="DeSousa2013")
ge.CRC <- ge.all[selPbs, ]
uniGenes <- pbs2unigenes(ge.CRC, sdat)
```

---

progAMC

*Prognosis of the three subtypes in the AMC data set*

---

**Description**

These functions performs survival analysis for the AMC data set, and generate a kaplan-meier plot to diagnose the prognosis of three subtypes.

**Usage**

```
progAMC(AMC_CRC_clinical, AMC_sample_head, clu.pred)
figKM(surv, survstats)
```

**Arguments**

|                  |                                                                                                                               |
|------------------|-------------------------------------------------------------------------------------------------------------------------------|
| AMC_CRC_clinical | a data frame including include clinical information (details in <a href="#">AMC_CRC_clinical</a> ).                           |
| AMC_sample_head  | a data frame include mapping information between microarray ids and sample ids (details in <a href="#">AMC_sample_head</a> ). |
| clu.pred         | a numeric vector of classification labels named by colon cancer samples.                                                      |
| surv             | the result of <a href="#">survfit</a> containing survival curves of the AMC data set.                                         |
| survstats        | the result of <a href="#">survdiff</a> for the AMC data set.                                                                  |

**Details**

The function progAMC performs survival analysis for the AMC data set and compares the prognosis of the three subtypes. The function figKM helps visualize the prognosis in a kaplan-meier plot.

**Value**

This function progAMC will save surv (result of [survfit](#) containing survival curves of the AMC data set), survstats (result of [survdiff](#)), data4surv (the data to perform survival analysis) to a file called survival.RData in the current working directory.

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

**Examples**

```
library(survival)
data(AMC, source="DeSousa2013")
data(predAMC, source="DeSousa2013")
prog <- progAMC(AMC_CRC_clinical, AMC_sample_head, clu.pred)
surv <- prog$surv
survstats <- prog$survstats
figKM(surv, survstats)
```

---

selTopVarGenes

*Select top variable probesets*


---

**Description**

This function selects top variable probesets across the 90 cancer samples by median absolute deviations.

**Usage**

```
selTopVarGenes(ge.CRC, MADth=0.5)
```



**Arguments**

|                     |                                                                                |
|---------------------|--------------------------------------------------------------------------------|
| <code>ge.CRC</code> | a numeric matrix of expression data of genes expressed in at least one sample. |
| <code>MADth</code>  | an numeric value specifying the cutoff of MAD (median absolute deviation).     |

**Value**

This function will return `sdat`, which is a matrix of median centered expression values of top variable probesets.

**Author(s)**

Xin Wang <xw264@cam.ac.uk>

**References**

De Sousa E Melo, F. and Wang, X. and Jansen, M. et al. Poor prognosis colon cancer is defined by a molecularly distinct subtype and precursor lesion. accepted

**Examples**

```
data(ge.CRC, package="DeSousa2013")
ge.CRC <- ge.all[selPbs, ]
sdat <- selTopVarGenes(ge.CRC, MADth=0.5)
```

# Index

AMC\_CRC\_clinical, [5](#), [13](#), [16](#)  
AMC\_CRC\_clinical (data-AMC), [7](#)  
AMC\_sample\_head, [5](#), [11](#), [16](#)  
AMC\_sample\_head (data-AMC), [7](#)  
  
barcode, [11](#)  
buildClassifier, [2](#), [6](#), [12](#), [14](#)  
  
cents (data-AMC), [7](#)  
clu.pred (data-AMC), [7](#)  
clus (data-AMC), [7](#)  
clusGap, [3](#), [5](#)  
ComBat, [11](#)  
compGapStats, [3](#), [6](#)  
conClust, [4](#), [6](#)  
CRCPipeLine, [5](#)  
  
data-AMC, [7](#)  
data4surv (data-AMC), [7](#)  
DeSousa2013 (CRCPipeLine), [5](#)  
diffGenes (data-AMC), [7](#)  
  
err (data-AMC), [7](#)  
  
figClassify (pamClassify), [13](#)  
figGAP (compGapStats), [3](#)  
figKM (progAMC), [15](#)  
figPAMCV (buildClassifier), [2](#)  
figSilh (filterSamples), [9](#)  
filterDiffGenes, [2](#), [3](#), [6](#), [8](#), [10](#)  
filterSamples, [2](#), [3](#), [6](#), [9](#)  
findDiffGenes, [3](#), [6](#), [8](#), [10](#)  
frma, [11](#)  
  
gapsmat (data-AMC), [7](#)  
gapsSE (data-AMC), [7](#)  
gclu.f (data-AMC), [7](#)  
ge.all (data-AMC), [7](#)  
geneExpPre, [6](#), [11](#)  
getCentroids, [12](#), [13](#)  
  
nam.ord (data-AMC), [7](#)  
  
pam.rs1t (data-AMC), [7](#)  
pamClassify, [6](#), [13](#)  
  
pamr.train, [2](#), [7](#), [13](#)  
pbs2unigenes, [6](#), [14](#)  
pred (data-AMC), [7](#)  
progAMC, [6](#), [15](#)  
  
sdat (data-AMC), [7](#)  
selPbs (data-AMC), [7](#)  
selTopVarGenes, [6](#), [9](#), [15](#), [16](#)  
signature (data-AMC), [7](#)  
silh (data-AMC), [7](#)  
silhouette, [7](#), [9](#)  
surv (data-AMC), [7](#)  
survdif, [7](#), [16](#)  
survfit, [7](#), [16](#)  
survstats (data-AMC), [7](#)  
  
thresh (data-AMC), [7](#)  
  
uniGenes (data-AMC), [7](#)