# Package 'BreastSubtypeR'

June 25, 2025

**Type** Package

**Title** Methods for breast cancer intrinsic subtyping

**Description** BreastSubtypeR is an R package that provides a collection of methods for intrinsic molecular subtyping of breast cancer. It includes subtyping methods for nearest centroid-based subtyping (NC-based) and single sample predictor (SSP-based), along with tools for integrating clinical data and visualizing results.

**Encoding** UTF-8

**Version** 1.1.0

**biocViews** RNASeq, Software, GeneExpression, Classification

**Depends** R (>= 4.5.0)

**Imports** methods, Biobase, tidyselect, dplyr, ggplot2, magrittr, rlang, stringr, withr, ComplexHeatmap, impute (>= 1.80.0), data.table (>= 1.16.0), RColorBrewer (>= 1.1-3), circlize (>= 0.4.16), ggrepel (>= 0.9.6), e1071 (>= 1.7-8), SummarizedExperiment

**Suggests** shiny (>= 1.9.1), bslib (>= 0.8.0), knitr, rmarkdown, BiocStyle, testthat

**URL** https://github.com/yqkiuo/BreastSubtypeR

**BugReports** https://github.com/yqkiuo/BreastSubtypeR/issues

**License** GPL (>= 3)

**VignetteBuilder** knitr

**Roxygen** list(markdown = TRUE)

**LazyData** FALSE

**RoxygenNote** 7.3.2

**git_url** https://git.bioconductor.org/packages/BreastSubtypeR

**git_branch** devel

**git_last_commit** 02e2619

**git_last_commit_date** 2025-04-15

**Repository** Bioconductor 3.22

**Date/Publication** 2025-06-25

**Author** Qiao Yang [aut, cre, fnd] (ORCID: <https://orcid.org/0000-0002-4098-3246>)

**Maintainer** Qiao Yang <yq.kiuo@gmail.com>

# Contents



---

`AIMSmodel` *The AIMS model*

---

## Description

This is the model definition for AIMS. It contains the naive bayes classifier composed of the 100 rules described in Paquet et al. "Absolute assignment of breast cancer intrinsic molecular subtype" (under review at JNCI).

## Usage

```
data("AIMSmodel")
```

## Format

An object of class `list` of length 4.

## Details

This is the AIMS model define using 100 simple rules of the form gene A < gene B and combine within a naive bayes classifier within e1071. (Paquet et al. under review JNCI).

Briefly, using a suitably large training dataset(~5000 gene breast cancer gene expression profiles), the approach identifies a small set of simple binary rules (~20) that examine the raw expression measurements for pairs of genes from a single breast cancer patient, and only that patient. The binary rules are of the form "if the expression of gene x is greater than gene y, then tend to assign

subtype z for that patient". Subtypes could be : Basal, Her2, LumA, LumB, or Normal. The collection of binary rules is combined for a single estimation of a patient subtype via a single probabilistic model using naiveBayes in e1071. In this way, since only expression levels of genes with a single patient is considered, the method represents a promising approach to ablate the instability caused by relativistic approaches (Paquet et al. in review at JNCI).

## Value

| | |
|---|---|
| `all.pairs` | The 100 rules in AIMS in the form EntrezID gene A < EntrezID gene B |
| `k` | The selected number of optimal rules. For AIMS we have shown it is 20 |
| `one.vs.all.tsp` | The Naive bayes classifier used in combination with the 100 rules |
| `selected.pairs.list` | The list of rules sorted from the best discriminating rule to the least discriminating rules subdivided by subtype |

## References


- Paquet ER, Hallett MT. *Absolute assignment of breast cancer intrinsic molecular subtype.* J Natl Cancer Inst. 2015;107(1). https://doi.org/10.1093/jnci/dju357


## Examples

```
library(BreastSubtypeR)
data("AIMSmodel")
```

---

`BreastSubtypeR` *BreastSubtypeR: A Unified R Package for Intrinsic Molecular Subtyping in Breast Cancer Research*

---

## Description

**BreastSubtypeR** is an R package designed to unify and streamline intrinsic molecular subtyping methods for breast cancer (BC).

It integrates both nearest-centroid (NC-based) and single-sample predictor (SSP-based) approaches, along with an innovative **AUTO mode** feature (described below).The package utilizes standardized input and output formats, providing a cohesive framework that is fully compatible with other R packages in the gene expression profiling field. Additionally, its core functions are accessible through an **interactive Shiny app**, making it user-friendly for researchers and clinicians with limited R programming experience.

### Workflow:

1. **Data Input**: Load example data or supply your own gene expression dataset as a SummarizedExperiment object.
2. **Gene Mapping**: Prepare your dataset for subtyping using the `Mapping` function.
3. **Subtyping**: Run multiple subtyping methods (or leverage AUTO mode) with the `BS_Multi` function.
4. **Visualization**: Explore and interpret the subtyping results using the `Vis_Multi` function.

### Key Functions:

- `Mapping`: Prepares gene expression data for subtyping.
- `BS_Multi`: Executes multiple subtyping methods simultaneously, including an **AUTO** mode for method selection based on cohort characteristics.
- `Vis_Multi`: Generates visualizations to facilitate interpretation of the subtyping outcomes.

## Author(s)

**Maintainer**: Qiao Yang <yq.kiuo@gmail.com> (ORCID) [funder]

## See Also

`Mapping`, `BS_Multi`, `Vis_Multi`

---

`BreastSubtypeRobj` *BreastSubtypeRobj: Data for NC-based Methods*

---

## Description

A list object containing the data required for nearest-centroid (NC)-based molecular subtyping methods. This includes platform medians, centroids, gene signatures, and subgroup quantiles, as well as information from the UNC232 training cohort.

## Usage

```
data("BreastSubtypeRobj")
```

## Format

A list with the following elements:

`medians` A matrix of medians prepared for nine sequencing platforms.

`centroid` The centroids provided by the `parker.original` method.

`genes.sig50` A data frame of 50 genes used in NC-based methods, including their proliferation information.

`ssBC.subgroupQuantile` Subgroup medians prepared by the `ssBC` method.

`genes.signature` A collection of genes used in both NC-based and SSP-based methods.

`UNC232` Data from the UNC232 training cohort.

`platform.UNC232` The sequencing platform used for the UNC232 training cohort.

## References


- Bernard PS, Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. *Supervised risk predictor of breast cancer based on intrinsic subtypes*. Journal of Clinical Oncology. 2009;27(8). https://doi.org/10.1200/JCO.2008.18.1370
- Zhao X, Rodland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Research. 2015;17(1). https://doi.org/10.1186/s13058-015-0520-4

- Fernandez-Martinez A, Krop IE, Hillman DW, Polley MY, Parker JS, Huebner L, et al. *Survival, pathologic response, and genomics in CALGB 40601 (Alliance), a neoadjuvant Phase III trial of paclitaxel-trastuzumab with or without lapatinib in HER2-positive breast cancer.* Journal of Clinical Oncology. 2020. https://doi.org/10.1200/JCO.20.01276
- Picornell AC, Echavarria I, Alvarez E, López-Tarruella S, Jerez Y, Hoadley K, et al. Breast cancer PAM50 signature: Correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. BMC Genomics. 2019;20(1). https://doi.org/10.1186/s12864-019-5849-0


## Examples

```
library(BreastSubtypeR)
data("BreastSubtypeRobj")
```

---

`BS_AIMS` *AIMS Intrinsic Subtyping (BS_AIMS)*

---

## Description

This function predicts breast cancer intrinsic subtypes using the **AIMS (Absolute assignment of Intrinsic Molecular Subtype)** method.

## Usage

```
BS_AIMS(se_obj)
```

## Arguments

| | |
|---|---|
| `se_obj` | A `SummarizedExperiment` object containing:<br>• **Assay data**: A gene expression matrix with genes (EntrezID) as rows and samples as columns. Important: The gene expression values should not be gene-centered. All expression values must be **positive**. |

## Value

Returns a vector of intrinsic subtypes assigned to the samples, as estimated by the AIMS method.

## References


- Paquet ER, Hallett MT. *Absolute assignment of breast cancer intrinsic molecular subtype.* J Natl Cancer Inst. 2015;107(1). https://doi.org/10.1093/jnci/dju357


## Examples

```
# Load required datasets
data("OSLO2EMIT0obj")


# Perform subtyping
res <- BS_AIMS(
    se_obj = OSLO2EMIT0obj$data_input$se_SSP
)
```

BS_cIHC *Conventional IHC Intrinsic Subtyping (BS_cIHC)*

## Description

This function predicts breast cancer intrinsic subtypes using the conventional estrogen receptor (ER)-balancing via immunohistochemistry (cIHC).

## Usage

```
BS_cIHC(se_obj, Subtype = FALSE, hasClinical = FALSE, seed = 118)
```

## Arguments

| | |
|---|---|
| se_obj | A SummarizedExperiment object containing:<br>• **Assay data**: A log-transformed, normalized gene expression matrix with genes (Gene Symbols) as rows and samples as columns.<br>• **Column metadata** (colData): A clinical information table, which must include:<br>– "PatientID": Unique sample or patient identifiers.<br>– "ER": Estrogen receptor (ER) status, recorded as "ER+" or "ER-". |
| Subtype | Logical (TRUE or FALSE). If TRUE, the function predicts four subtypes, **excluding** the Normal-like subtype. |
| hasClinical | Logical (TRUE or FALSE). If TRUE, the function incorporates clinical data from the phenotype (pheno) table. Required columns:<br>• "TSIZE": Tumor size (0 for <= 2cm, 1 for > 2cm).<br>• "NODE": Lymph node status (0 for negative, 1 or higher for positive nodes; this column must be numeric). |
| seed | An integer used to set the random seed for reproducibility. |

## Value

Returns a data frame containing intrinsic subtypes estimated using the conventional IHC (cIHC) method.

## References

• Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. *Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer*. Cell. 2015;163(2). https://doi.org/10.1016/j.cell.2015.09.0

## Examples

```
data("OSLO2EMIT0obj")
res <- BS_cIHC(
    se_obj = OSLO2EMIT0obj$data_input$se_NC,
    Subtype = FALSE,
    hasClinical = FALSE
)
```

BS_cIHC.itr *Iterative conventional IHC Intrinsic Subtyping (BS_cIHC.itr)*

## Description

This function predicts breast cancer intrinsic subtypes using an **iterative** version of conventional estrogen receptor (ER)-balancing via immunohistochemistry (cIHC). It allows customization of the ER+/ER- ratio to refine subtype classification..

## Usage

```
BS_cIHC.itr(
  se_obj,
  iteration = 100,
  ratio = 54/64,
  Subtype = FALSE,
  hasClinical = FALSE,
  seed = 118
)
```

## Arguments

| | |
|---|---|
| `se_obj` | A `SummarizedExperiment` object containing:<br>• **Assay data**: A log-transformed, normalized gene expression matrix with genes (Gene Symbol) in rows and samples in columns.<br>• **Column metadata** (`colData`): Clinical information table.<br>– `"PatientID"`: Unique sample or patient identifiers.<br>– `"ER"`: Estrogen receptor (ER) status recorded as `"ER+"` or `"ER-"`. |
| `iteration` | Integer. The number of iterations for the ER-balanced procedure with the specified ratio. Default: 100. |
| `ratio` | Numeric. Specifies the ER+/ER- ratio for balancing. Options:<br>• `1:1`: Equal balancing.<br>• `54:64`: Default, based on the ER+/ER- ratio in the UNC232 training cohort. |
| `Subtype` | Logical (`TRUE` or `FALSE`). If `TRUE`, the function predicts four subtypes, **excluding** the Normal-like subtype. |
| `hasClinical` | Logical (`TRUE` or `FALSE`). If `TRUE`, the function incorporates clinical data from the phenotype (`pheno`) table. Required columns:<br>• `"TSIZE"`: Tumor size (`0` for <= 2cm, `1` for > 2cm).<br>• `"NODE"`: Lymph node status (`0` for negative, `1` or higher for positive nodes; this column must be numeric). |
| `seed` | An integer used to set the random seed for reproducibility. |

## Value

A list containing:

- Intrinsic subtype predictions.
- Confidence levels for each subtype.
- Percentages of ER+ and ER??? subsets across iterations.

## References


- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*. Nature. 2012;486(7403). https://doi.org/10.1038/nature10983


## Examples

```
data("OSLO2EMIT0obj")
res <- BS_cIHC.itr(
    se_obj = OSLO2EMIT0obj$data_input$se_NC,
    iteration = 10, ## For final analysis, set iteration = 100
    Subtype = FALSE,
    hasClinical = FALSE
)
```

---

`BS_Multi` *Intrinsic Subtyping with Multiple Approaches (BS_Multi)*

---

## Description

This function predicts breast cancer intrinsic subtypes using multiple methods. Users can either specify the subtyping approaches directly or enable automatic selection ("AUTO") based on the ER/HER2 distribution of the test cohort.

## Usage

```
BS_Multi(data_input, methods = "AUTO", Subtype = FALSE, hasClinical = FALSE)
```

## Arguments

| | |
|---|---|
| `data_input` | The output from the `Mapping()` function, containing processed gene expression data prepared for subtyping analysis. |
| `methods` | A character vector specifying the subtyping methods to be used. Available options:<br>• "parker.original": Original PAM50 by Parker et al., 2009 (Parker et al., 2009)<br>• "genefu.scale": PAM50 implementation as in the genefu R package (scaled version) (Gendoo et al., 2016)<br>• "genefu.robust": PAM50 implementation as in the genefu R package (robust version) (Gendoo et al., 2016)<br>• "cIHC": Conventional estrogen receptor (ER)-balancing using immunohistochemistry (cIHC) (Ciriello et al., 2015)<br>• "cIHC.itr": Iterative version of cIHC (Curtis et al., 2012)<br>• "PCAPAM50": PCA-based iterative PAM50 (ER-balancing using ESR1 gene expression) (Raj-Kumar et al., 2019)<br>• "ssBC": Subgroup-specific gene-centering PAM50 (Zhao et al., 2015)<br>• "ssBC.v2": Updated subgroup-specific gene-centering PAM50 with refined quantiles (Fernandez-Martinez et al., 2020) |

- "AIMS": Absolute Intrinsic Molecular Subtyping (AIMS) method (Paquet & Hallett, 2015)
- "sspbc": Single-Sample Predictors for Breast Cancer (AIMS adaptation) (Staaf et al., 2022)
- "AUTO": Automatically selects subtyping methods based on the ER/HER2 distribution of the test cohort.

Notes:

- If "AUTO" is selected, it must be the sole value in the vector.
- If "AUTO" is not selected, at least **two** methods must be specified; otherwise, an error will occur.

`Subtype` Logical (`TRUE` or `FALSE`). If `TRUE`, the function predicts four subtypes, **excluding** the Normal-like subtype.

`hasClinical` Logical (`TRUE` or `FALSE`). If `TRUE`, the function incorporates clinical data from the phenotype (`pheno`) table. Required columns:

- `"TSIZE"`: Tumor size (`0` for <= 2cm, `1` for > 2cm).
- `"NODE"`: Lymph node status (`0` for negative, `1` or higher for positive nodes; this column must be numeric).

## Value

Returns a list of intrinsic subtypes estimated by the selected methods.

## References


- Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. *Supervised risk predictor of breast cancer based on intrinsic subtypes*. Journal of Clinical Oncology. 2009;27(8). https://doi.org/10.1200/JCO.2008.18.1370
- Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. *Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer*. Bioinformatics. 2016;32(7). https://doi.org/10.1093/bioinformatics/btv693
- Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. *Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer*. Cell. 2015;163(2). https://doi.org/10.1016/j.cell.2015.09.0
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*. Nature. 2012;486(7403). https://doi.org/10.1038/nature10983
- Zhao X, Rodland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Research. 2015;17(1). https://doi.org/10.1186/s13058-015-0520-4
- Fernandez-Martinez A, Krop IE, Hillman DW, Polley MY, Parker JS, Huebner L, et al. *Survival, pathologic response, and genomics in CALGB 40601 (Alliance), a neoadjuvant Phase III trial of paclitaxel-trastuzumab with or without lapatinib in HER2-positive breast cancer.* Journal of Clinical Oncology. 2020. https://doi.org/10.1200/JCO.20.01276


- Zhao X, Rodland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Research. 2015;17(1). https://doi.org/10.1186/s13058-015-0520-4
- Fernandez-Martinez A, Krop IE, Hillman DW, Polley MY, Parker JS, Huebner L, et al. *Survival, pathologic response, and genomics in CALGB 40601 (Alliance), a neoadjuvant Phase III trial of paclitaxel-trastuzumab with or without lapatinib in HER2-positive breast cancer.* Journal of Clinical Oncology. 2020. https://doi.org/10.1200/JCO.20.01276

- Paquet ER, Hallett MT. *Absolute assignment of breast cancer intrinsic molecular subtype.* J Natl Cancer Inst. 2015;107(1). https://doi.org/10.1093/jnci/dju357
- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer.* NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3


## Examples

```
# Load required dataset
data("OSLO2EMIT0obj")

# Define methods to use for consensus subtyping
methods <- c("parker.original", "genefu.scale", "genefu.robust")

# Perform subtyping
res.test <- BS_Multi(
    data_input = OSLO2EMIT0obj$data_input,
    methods = methods,
    Subtype = FALSE,
    hasClinical = FALSE
)
```

---

`BS_parker` *Original Parker Intrinsic Subtyping (BS_parker)*

---

## Description

This function predicts breast cancer intrinsic subtypes using the original Parker et al. (2019) method, along with variations of the original approach.

## Usage

```
BS_parker(
  se_obj,
  calibration = "None",
  internal = NA,
  external = NA,
  medians = NA,
  Subtype = FALSE,
  hasClinical = FALSE
)
```

## Arguments

| | |
|---|---|
| `se_obj` | A `SummarizedExperiment` object containing:<br>• **Assay data**: A log-transformed, normalized gene expression matrix with genes (Gene Symbols) as rows and samples as columns.<br>• **Column metadata** (`colData`): Optional clinical information. |
| `calibration` | Specifies the calibration method to apply. Options include: |

- "None": No calibration is applied.
- "Internal": Uses internal calibration strategies (see internal argument).
- "External": Uses external medians (see external argument).

internal — Specifies the internal calibration strategy when calibration = "Internal". Options include:

- "medianCtr" (default): Median-centered calibration.
- "meanCtr": Mean-centered calibration (aligned with genefu.scale).
- "qCtr": Quantile-based calibration (aligned with genefu.robust).

external — Specifies the platform name (i.e., column name) for external medians derived from the training cohort.

- To use user-provided medians, set "external = "Given.mdns" and provide values via the medians argument.

medians — A matrix or table of user-provided median values, required if external = "Given.mdns".

- The first column should contain 50 genes.
- The second column should contain the corresponding median values.

Subtype — Logical (TRUE or FALSE). If TRUE, the function predicts four subtypes, **excluding** the Normal-like subtype.

hasClinical — Logical (TRUE or FALSE). If TRUE, the function incorporates clinical data from the phenotype (pheno) table. Required columns:

- "TSIZE": Tumor size (0 for <= 2cm, 1 for > 2cm).
- "NODE": Lymph node status (0 for negative, 1 or higher for positive nodes; this column must be numeric).

## Value

Returns a list containing intrinsic subtypes assigned using the Parker-based method, or its variations.

## References


- Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. *Supervised risk predictor of breast cancer based on intrinsic subtypes*. Journal of Clinical Oncology. 2009;27(8). https://doi.org/10.1200/JCO.2008.18.1370
- Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. *Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer*. Bioinformatics. 2016;32(7). https://doi.org/10.1093/bioinformatics/btv693


## Examples

```
data("OSLO2EMIT0obj")
res <- BS_parker(
    se_obj = OSLO2EMIT0obj$data_input$se_NC,
    calibration = "Internal",
    internal = "medianCtr",
    Subtype = FALSE,
    hasClinical = FALSE
)
```

BS_PCAPAM50 *PCA-PAM50 Intrinsic Subtyping (BS_PCAPAM50)*

## Description

This function predicts breast cancer intrinsic subtypes using the PCA-PAM50 method. This approach integrates **Principal Component Analysis (PCA)** to perform estrogen receptor (ER) balancing based on ESR1 gene expression.

## Usage

```
BS_PCAPAM50(se_obj, Subtype = FALSE, hasClinical = FALSE, seed = 118)
```

## Arguments

| | |
|---|---|
| se_obj | A SummarizedExperiment object containing:<br>• **Assay data**: A log-transformed, normalized gene expression matrix with genes (Gene Symbols) as rows and samples as columns.<br>• **Column metadata** (colData): Clinical information table.<br>– "PatientID": Unique sample or patient identifiers.<br>– "ER": Estrogen receptor (ER) status, recorded as "ER+" or "ER-". |
| Subtype | Logical (TRUE or FALSE). If TRUE, the function predicts four subtypes, **excluding** the Normal-like subtype. |
| hasClinical | Logical (TRUE or FALSE). If TRUE, the function incorporates clinical data from the phenotype (pheno) table. Required columns:<br>• "TSIZE": Tumor size (0 for <= 2cm, 1 for > 2cm).<br>• "NODE": Lymph node status (0 for negative, 1 or higher for positive nodes; this column must be numeric). |
| seed | An integer used to set the random seed for reproducibility. |

## Value

Returns a vector of intrinsic subtypes assigned to the samples, as estimated by the PCA-PAM50 method.

## References


- Raj-Kumar PK, Liu J, Hooke JA, Kovatich AJ, Kvecher L, Shriver CD, et al. *PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping, reclassifying a subset of luminal A tumors as luminal B.* Sci Rep. 2019;9(1). https://doi.org/10.1038/s41598-019-44339-4


## Examples

```
data("OSLO2EMIT0obj")
res <- BS_PCAPAM50(
    se_obj = OSLO2EMIT0obj$data_input$se_NC,
    Subtype = FALSE,
    hasClinical = FALSE
)
```

`BS_ssBC` *Subgroup-specific gene-centering Intrinsic Subtyping (BS_ssBC)*

## Description

This function predicts breast cancer intrinsic subtypes using the **subgroup-specific (ssBC)** method. The ssBC method applies a subgroup-specific gene-centering approach to cohorts with a skewed distribution of clinicopathological characteristics compared to the original training cohort (e.g., an ER+ selected cohort).

## Usage

```
BS_ssBC(se_obj, s, Subtype = FALSE, hasClinical = FALSE)
```

## Arguments

`se_obj` A `SummarizedExperiment` object containing:

- **Assay data**: A log-transformed, normalized gene expression matrix with genes (Gene Symbols) as rows and samples as columns.
- **Column metadata** (`colData`): A clinical information table. If `hasClinical = TRUE`, this table must include:
  - `"PatientID"`: Unique identifiers for patients or samples.
  - Additional columns depending on the `s` parameter:
    - `"ER"`: Estrogen receptor status (`"ER+"` or `"ER-"`) if `s = "ER"`.
    - `"HER2"`: HER2 status (`"HER2+"` or `"HER2-"`) if `s = "ER.v2"`.
    - `"TN"`: Triple-negative status (`"TN"` or `"nonTN"`) if `s = "TN"` or `s = "TN.v2"`, indicating a triple-negative cohort.

`s` Character. Specifies which subgroup-specific quantiles to use:

- `"ER"` and `"TN"`: Original subgroup-specific quantiles published in *Breast Cancer Research* (2015).
- `"ER.v2"` and `"TN.v2"`: Updated subgroup-specific quantiles published in *Journal of Clinical Oncology* (2020).

`Subtype` Logical (`TRUE` or `FALSE`). If `TRUE`, the function predicts four subtypes, **excluding** the Normal-like subtype.

`hasClinical` Logical (`TRUE` or `FALSE`). If `TRUE`, the function incorporates clinical data from the phenotype (`pheno`) table. Required columns:

- `"TSIZE"`: Tumor size (`0` for <= 2cm, `1` for > 2cm).
- `"NODE"`: Lymph node status (`0` for negative, `1` or higher for positive nodes; this column must be numeric).

## Value

Returns a vector of intrinsic subtypes assigned to the samples, as estimated by the ssBC method.

## References


- Zhao X, Rodland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Research. 2015;17(1). https://doi.org/10.1186/s13058-015-0520-4
- Fernandez-Martinez A, Krop IE, Hillman DW, Polley MY, Parker JS, Huebner L, et al. *Survival, pathologic response, and genomics in CALGB 40601 (Alliance), a neoadjuvant Phase III trial of paclitaxel-trastuzumab with or without lapatinib in HER2-positive breast cancer.* Journal of Clinical Oncology. 2020. https://doi.org/10.1200/JCO.20.01276


## Examples

```
## ssBC.v2
data("OSLO2EMIT0obj")
res <- BS_ssBC(
    se_obj = OSLO2EMIT0obj$data_input$se_NC,
    s = "ER.v2",
    Subtype = FALSE,
    hasClinical = FALSE
)
```

---

`BS_sspbc` *Intrinsic Subtyping using SSPBC (BS_sspbc)*

---

## Description

This function predicts breast cancer intrinsic subtypes using SSPBC (Single Sample Predictor for Breast Cancer). SSPBC is based on a refined version of the original AIMS methodology, utilizing a large, uniformly accrued population-based cohort (SCAN-B) for training. This method supports RNA sequencing data and provides flexibility in selecting the prediction model.

## Usage

```
BS_sspbc(se_obj, ssp.name = "ssp.pam50")
```

## Arguments

| | |
|---|---|
| `se_obj` | A `SummarizedExperiment` object containing:<br>• **Assay data**: A gene expression matrix with genes (EntrezID) as rows and samples as columns. Important: The gene expression values should not be gene-centered. All expression values must be **positive**. |
| `ssp.name` | Specifies the model to use. Options are:<br>• "ssp.pam50": For PAM50-based predictions.<br>• "ssp.subtype": For predicting Prosigna-like subtypes (four subtypes, **excluding** the Normal-like subtype). |

## Value

Returns a vector of intrinsic subtypes assigned to the samples, as estimated by the SSPBC method.

## References


- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer*. NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3


## Examples

```
# Load required dataset
data("OSLO2EMIT0obj")

# Perform subtyping with the SSPBC method
res <- BS_sspbc(
    se_obj = OSLO2EMIT0obj$data_input$se_SSP,
    ssp.name = "ssp.pam50"
)
```

---

`Gene.ID.ann` *Gene annotation table.*

---

## Description

Annotation table for the GENCODE Human Release 27 genes (Gene.ID) included in the StringTie target when summarizing gene expression. The annotation are from from GENCODE Human Release 27 metadata files and include HGNC, EntrezGene and RefSeq.

## Usage

```
data("Gene.ID.ann")
```

## Format

An object of class `data.frame` with 19675 rows and 6 columns.

## Details

Annotation table used by the applySSP to translate gene identifiers as needed before classification with provided ssp models.

## Value

`Gene.ID.ann` Annotation table for GENCODE Human Release 27 genes.

## References


- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer*. NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3

## Examples

```
library(BreastSubtypeR)
## Load the Gene.ID.ann
data(Gene.ID.ann)
```

---

`iBreastSubtypeR` *iBreastSubtypeR*

---

## Description

Starts an interactive BreastSubtypeR shiny web app.

BreastSubtypeR integrates intrinsic molecular subtyping methods for breast cancer, including nearest-centroid (NC-based) and single-sample predictor (SSP-based) approaches. It employs standardized input and output formats, providing a unified framework that is highly compatible with other R packages in the gene expression profiling field.

The `iBreastSubtypeR()` function launches an interactive Shiny web application. This app enables users to configure the arguments of subtyping functions and execute subtyping on their local computer. For detailed descriptions of the arguments, including their default and alternative values, please refer to the manual pages of the respective functions.

Step 1:

The input data can be loaded from the user's workspace or by selecting a CSV/text file for the expression data, a CSV/text file for clinical information, and a CSV/text file for feature annotations.

After loading the necessary files, users can click the "Map Now" button once and wait for the notification. If the Mapping() function runs successfully, a message will appear stating, "You may now proceed to Step 2."

Step 2:

Users can select the desired subtyping method and adjust the relevant parameters to conduct their analysis. Once the analysis is complete, a message will indicate, "Analysis is complete." Two visualizations will be displayed, and you will have the option to download the results as a text file. If you wish to continue your analysis, you can directly run another method without needing to repeat Step 1.

## Usage

```
iBreastSubtypeR()
```

## Value

A table with subtyping and ROR score

## Examples

```
library(BreastSubtypeR)

# This will open your browser with the BreastSubtypeR shiny web app

iBreastSubtypeR()
```

`Mapping` *Gene ID Mapping*

## Description

A function to map gene identifiers and preprocess gene expression data for downstream analyses.

## Usage

```
Mapping(
  se_obj,
  method = c("max", "mean", "median", "iqr", "stdev"),
  impute = TRUE,
  verbose = TRUE
)
```

## Arguments

| | |
|---|---|
| `se_obj` | A `SummarizedExperiment` object containing:<br>• **Assay data**: A log2-transformed, normalized gene expression matrix, where rows correspond to probes (e.g., ProbeID, TranscriptID, or Gene Symbol) and columns correspond to samples. This should be stored in the `assay()` slot.<br>• **Row metadata**: A data frame with probe annotations, including at least the following columns:<br>– `"probe"`: Unique identifiers for the probes (e.g., ProbeID, TranscriptID or Gene Symbol).<br>– `"ENTREZID"`: Entrez Gene IDs corresponding to the probes.<br>• **Column metadata** (optional): Sample metadata stored in the `colData()` slot. |
| `method` | A string specifying the method for resolving duplicate probes in microarray or RNA-seq data. Options include:<br>• `"iqr"`: Selects the probe with the highest interquartile range (IQR), typically used for short-oligo arrays (e.g., Affymetrix).<br>• `"mean"`: Chooses the probe with the highest average expression, commonly used for long-oligo arrays (e.g., Agilent, Illumina).<br>• `"max"`: Retains the probe with the highest expression value, often used for RNA-seq data.<br>• `"stdev"`: Selects the probe with the highest standard deviation.<br>• `"median"`: Chooses the probe with the highest median expression value. |
| `impute` | Logical (`TRUE` or `FALSE`). If `TRUE`, performs K-Nearest Neighbors (KNN) imputation to handle missing data (`NA` values). |
| `verbose` | Logical (`TRUE` or `FALSE`). If `TRUE`, displays progress messages during execution. |

## Details

If gene symbols are used as row identifiers in the gene expression matrix, an additional `SYMBOL` column must be added to the feature table and renamed as`probe`.

## Value

Returns a list containing two preprocessed gene expression datasets:

- "x_NC": A SummarizedExperiment object containing: 1) The log2-transformed gene expression matrix for nearest-centroid (NC)-based methods; 2) clinical metadata.
- "x_SSP": A SummarizedExperiment object containing: 1) The exponential-transformed gene expression matrix for single-sample predictor (SSP)-based methods; 2) clinical metadata.

## Examples

```
data("OSLO2EMIT0obj")
data_input <- Mapping(
    se_obj = OSLO2EMIT0obj$se_obj,
    method = "max",
    impute = TRUE,
    verbose = FALSE
)
```

---

# OSLO2EMIT0obj *OSLO2EMIT0obj: Example Dataset for the OSLO2EMIT0 Cohort*

---

## Description

This example dataset is based on the OSLO2EMIT0 cohort described in Staaf et al., 2022. It contains subsetted data for gene expression, clinical information, feature annotations, and example outputs from the Mapping and BS_Multi functions.

## Usage

```
data("OSLO2EMIT0obj")
```

## Format

A list containing the following elements:

se_obj A SummarizedExperiment object containing a subset of the log2-transformed, normalized gene expression matrix, clinical information, and gene feature annotations for the OSLO2EMIT0 cohort.

data_input Example output from the Mapping function.

res Example output from the BS_Multi function with *AUTO* mode.

## References


- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer*. NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3


## Examples

```
library(BreastSubtypeR)
data("OSLO2EMIT0obj")
```

## sspbc.models *The sspbc models (short name) for the 11 developed predictors*

### Description

List with the 11 SSP models from Staaf J. et al. medRxiv 2021.12.03.21267116.

### Usage

```
data("sspbc.models")
```

### Format

An object of class `list` of length 11.

### Details

List elements are named with short name for respective ssp model. The ssp models in list sspbc.models are the same as in list sspbc.models.fullname.

### Value

`sspbc.models` The collection of 11 ssp models used by sspbc.

#' @references


- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer*. NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3


### Examples

```
library(BreastSubtypeR)
## Load the sspbc.models
data("sspbc.models")
```

## sspbc.models.fullname *The sspbc models (full name) for the 11 developed predictors*

### Description

List with the 11 SSP models from Staaf J. et al. medRxiv 2021.12.03.21267116.

### Usage

```
data("sspbc.models.fullname")
```

### Format

An object of class `list` of length 11.

## Details

List elements are named with full name for respective ssp model. Note that ssp models in list sspbc.models.fullname are the same as in list sspbc.models.

## Value

`sspbc.models.fullname`
: The collection of 11 ssp models used by sspbc.

## References


- Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. *RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer*. NPJ Breast Cancer. 2022;8(1). https://doi.org/10.1038/s41523-022-00465-3


## Examples

```
library(BreastSubtypeR)
## Load the sspbc.models
data(sspbc.models.fullname)
```

---

# Vis_boxplot *Boxplot of Correlation per Subtype*

---

## Description

This function generates a boxplot to visualize the correlation distribution between different subtypes of breast cancer, based on the provided correlation table and subtype information.

## Usage

```
Vis_boxplot(out, correlations)
```

## Arguments

| | |
|---|---|
| `out` | A data frame containing the columns `"PatientID"` and `"Subtype"`. The `"PatientID"` column should have unique identifiers for each patient, and the `"Subtype"` column should specify the assigned subtype for each patient. |
| `correlations` | A data frame or matrix containing the correlation values computed from NC-based methods. |

## Value

A `ggplot` object representing the boxplot visualization of the correlation distributions across the different subtypes.

## Examples

```
data("OSLO2EMIT0obj")
res <- OSLO2EMIT0obj$res

# Prepare data: Subtype information and correlation matrix
out <- data.frame(
    PatientID = res$results$genefu.robust$BS.all$PatientID,
    Subtype = res$results$genefu.robust$BS.all$BS
)

correlations <- res$results$genefu.robust$outList$distances

# Generate the boxplot
p <- Vis_boxplot(out, correlations)
plot(p)
```

---

`Vis_heatmap` *Heatmap Visualization of Gene Expression by Subtype*

---

## Description

This function generates a heatmap to visualize gene expression patterns across breast cancer subtypes, based on the provided gene expression matrix and subtype information.

## Usage

```
Vis_heatmap(x, out)
```

## Arguments

| | |
|---|---|
| `x` | A gene expression matrix, where genes are rows and samples are columns. The data should be log2 transformed. |
| `out` | A data frame containing two columns: `"PatientID"` and `"Subtype"`. The `"PatientID"` column should contain unique patient identifiers, and the `"Subtype"` column should specify the assigned subtype for each patient. |

## Value

A `ggplot` or `heatmap` object (depending on implementation) representing the heatmap of gene expression across different subtypes.

## Examples

```
library(SummarizedExperiment)
data("OSLO2EMIT0obj")
res <- OSLO2EMIT0obj$res

# Prepare data: Gene expression matrix and subtype information
x <- assay(OSLO2EMIT0obj$data_input$se_NC)
out <- data.frame(
    PatientID = res$results$genefu.robust$BS.all$PatientID,
    Subtype = res$results$genefu.robust$BS.all$BS
```

```
)

# Generate the heatmap
p <- Vis_heatmap(x, out)
plot(p)
```

---

`Vis_Multi` *Multi-Method Subtype Heatmap Visualization*

---

## Description

This function generates a heatmap to visualize breast cancer subtypes classified by multiple subtyping methods. It helps users compare how different methods assign subtypes to the same set of samples.

## Usage

```
Vis_Multi(data)
```

## Arguments

| | |
|---|---|
| `data` | Output of the `BS_Multi` function. |

## Value

Returns a heatmap visualizing the subtype classifications across multiple methods.

## Examples

```
data("OSLO2EMIT0obj")

# Assuming `OSLO2EMIT0obj$res$res_subtypes` contains multi-method subtype results
p <- Vis_Multi(OSLO2EMIT0obj$res$res_subtypes)
plot(p)
```

---

`Vis_PCA` *PCA Plot Visualization of Gene Expression by Subtype*

---

## Description

This function generates a PCA plot to visualize the principal components of gene expression data, colored by the assigned subtypes. Optionally, it can display a scree plot of eigenvalues to evaluate the explained variance.

## Usage

```
Vis_PCA(x, out, Eigen = FALSE)
```

### Arguments

| | |
|---|---|
| x | A gene expression matrix, where genes are rows and samples are columns. The data should be log2 transformed. |
| out | A data frame containing two columns: "PatientID" and "Subtype". The "PatientID" column should contain unique patient identifiers, and the "Subtype" column should specify the assigned subtype for each patient. |
| Eigen | Logical. If TRUE, the function will display a scree plot showing the eigenvalues of the principal components. |

### Value

A ggplot object representing the PCA plot, colored by subtype. If Eigen is set to TRUE, a scree plot of the eigenvalues is also included.

### Examples

```
library(SummarizedExperiment)
data("OSLO2EMIT0obj")
res <- OSLO2EMIT0obj$res

# Prepare data: Gene expression matrix and subtype information
x <- assay(OSLO2EMIT0obj$data_input$se_NC)
out <- data.frame(
    PatientID = res$results$genefu.robust$BS.all$PatientID,
    Subtype = res$results$genefu.robust$BS.all$BS
)

# Generate the PCA plot
p <- Vis_PCA(x = x, out = out)
plot(p)

# Generate PCA plot with scree plot of eigenvalues
p_with_eigen <- Vis_PCA(x = x, out = out, Eigen = TRUE)
plot(p_with_eigen)
```

---

## Vis_pie *Pie Chart Visualization of Subtype Distribution*

---

### Description

This function generates a pie chart to visualize the distribution of breast cancer subtypes in a cohort, based on the provided Subtype data.

### Usage

```
Vis_pie(out)
```

### Arguments

| | |
|---|---|
| out | A data frame containing two columns: "PatientID" and "Subtype". The "PatientID" column should contain unique patient identifiers, and the "Subtype" column should specify the assigned subtype for each patient. |

## Value

A ggplot object representing a pie chart showing the proportion of each subtype in the dataset.

## Examples

```
data("OSLO2EMIT0obj")
res <- OSLO2EMIT0obj$res

# Prepare data: Subtype information
out <- data.frame(
    PatientID = res$results$genefu.robust$BS.all$PatientID,
    Subtype = res$results$genefu.robust$BS.all$BS
)

# Generate the pie chart
p <- Vis_pie(out = out)
plot(p)
```

# Index