Package 'BreastSubtypeR'

April 7, 2026

Type Package

Title Cohort-aware methods for intrinsic molecular subtyping of breast cancer

Description BreastSubtypeR provides an assumption-aware, multi-method framework for intrinsic molecular subtyping of breast cancer. The package harmonizes several published nearest-centroid (NC) and single-sample predictor (SSP) classifiers, supplies method-specific preprocessing and robust probe-to-gene mapping, and implements a cohort-aware AUTO mode that selectively enables classifiers compatible with the cohort composition. A local Shiny app (iBreastSubtypeR) is included for interactive analyses and to support users without programming experience.

Encoding UTF-8

Version 1.2.0

biocViews RNASeq, Software, GeneExpression, Classification, Preprocessing, Visualization

Depends R (>= 4.5.0)

Imports methods, Biobase, tidyselect, dplyr, ggplot2, magrittr, rlang, stringr, withr, edgeR, 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, utils

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

URL https://doi.org/10.18129/B9.bioc.BreastSubtypeR,https://github.com/yqkiuo/BreastSubtypeR,https://github.com/JohanHartmanGroupBioteam/BreastSubtypeR

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

License GPL-3

VignetteBuilder knitr

Roxygen list(markdown = TRUE, roclets = c(``rd'', ``namespace'', ``collate''))

LazyData FALSE

RoxygenNote 7.3.3

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

git_branch RELEASE_3_22

git_last_commit f1af123

git_last_commit_date 2025-10-29

Repository Bioconductor 3.22

Date/Publication 2026-04-07

Author Qiao Yang [aut, cre] (ORCID: <https://orcid.org/0000-0002-4098-3246>), Emmanouil G. Sifakis [aut] (ORCID: <https://orcid.org/0000-0001-9919-4471>)

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

Contents

`AIMSmodel` *AIMSmodel: Model object for AIMS*

Description

Model definition for AIMS consisting of 100 pairwise rules and a Naive Bayes classifier (via **e1071**) as described by Paquet & Hallett (2015).

Usage

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

Format

An object of class `list` of length 4.

Details

The 100 rules are of the form "EntrezID gene A < EntrezID gene B". A subset of k rules (typically 20) is used within a Naive Bayes classifier to assign subtypes (Basal-like, HER2-enriched, LumA, LumB, Normal-like) on a per-sample basis.

Value

`all.pairs`	Character vector of the 100 AIMS rules (EntrezID comparisons).
`k`	Integer; number of optimal rules (commonly 20).
`one.vs.all.tsp`	Naive Bayes classifier object used with the rules.
`selected.pairs.list`	Rules ranked by discriminative power per subtype.

References

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

Examples

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

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

Description

BreastSubtypeR is an R/Bioconductor package that unifies multiple published intrinsic subtyping (IS) methods for breast cancer into a single, reproducible framework. It supports both nearest-centroid (NC-based) and single-sample predictor (SSP-based) classifiers and introduces an assumption-aware **AUTO mode** that dynamically selects methods compatible with the input cohort.

By standardising input handling, applying method-specific normalisation, and providing optimised probe-to-gene mapping, BreastSubtypeR reduces inconsistencies across platforms and improves reproducibility in translational research. A companion Shiny app (**iBreastSubtypeR**) offers an intuitive GUI for non-programmers while preserving data privacy.

Workflow:

1. **Data Input**: Supply a gene expression dataset as a `SummarizedExperiment`. Supported inputs include raw RNA-seq counts (with gene lengths), log2(FPKM+1) RNA-seq, or log2-normalised microarray/nCounter data.
2. **Gene Mapping**: Prepare expression data with `Mapping`, including Entrez ID-based resolution of duplicates.
3. **Subtyping**: Apply multiple classifiers simultaneously using `BS_Multi`, or enable **AUTO mode** for cohort-aware method selection.

4. **Visualisation**: Summarise and interpret subtyping results with `Vis_Multi`.

Key Features:

- **Multi-method framework**: Ten published NC- and SSP-based classifiers, harmonised under one interface.
- **AUTO mode**: Evaluates cohort composition (e.g., ER/HER2 prevalence, subtype purity, subgroup sizes) and disables classifiers with violated assumptions; improves accuracy, Cohen's kappa, and IHC concordance.
- **Standardised normalisation**: Upper-quartile log2-CPM for NC-based methods; FPKM for SSP-based methods.
- **Optimised gene mapping**: Entrez ID-based mapping with conflict resolution.
- **Dual accessibility**: A Bioconductor-compliant R API and a local Shiny app (iBreastSubtypeR).

Author(s)

Maintainer: Qiao Yang <yq.kiuo@gmail.com> (ORCID)

Authors:

- Emmanouil G. Sifakis <emmanouil.sifakis@ki.se> (ORCID)

See Also

`Mapping`, `BS_Multi`, `Vis_Multi`

`BreastSubtypeRobj` *BreastSubtypeRobj: Resources for NC-based methods*

Description

List of reference resources required by nearest-centroid (NC) subtyping methods: platform medians, centroids, signatures, subgroup quantiles, and metadata from the UNC232 training cohort.

Usage

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

Format

A list with:

`medians` Matrix/data frame of platform-specific medians for **11** expression/sequencing platforms, derived as described in Picornell et al. (2019). Platform columns include: `nCounter`, `totalRNA.FFPE.20151111`, `RNAseq.Freeze.20120907`, `RNAseq.V2`, `RNAseq.V1`, `GC.4x44Kcustom`, `Agilent_244K`, `commercial_1x44k_post` `commercial_4x44k_postMeanCollapse_WashU_v2`, `htp1.5_WU_update`, `arrayTrain_postMeanCollapse`.

`centroid` PAM50 centroids used by `parker.original`.

`genes.sig50` Data frame of the 50 PAM50 genes with a proliferation flag.

`ssBC.subgroupQuantile` Subgroup-specific quantiles used by `ssBC`.

`genes.signature` Marker genes used across NC- and SSP-based methods.

`UNC232` Summary data for the UNC232 training cohort.

`platform.UNC232` Platform annotation for UNC232.

References

Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160–1167. 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 Res*. 2015;17(1):29. 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). *J Clin Oncol*. 2020;38(36):4184–4197. 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 TNBC series. *BMC Genomics*. 2019;20(1):452. https://doi.org/10.1186/s12864-019-5849-0

Examples

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

`BS_AIMS` *AIMS Intrinsic Subtyping (BS_AIMS)*

Description

Implements the **AIMS (Absolute Assignment of Intrinsic Molecular Subtype)** method for breast cancer intrinsic subtyping. Unlike nearest-centroid (NC) approaches, AIMS is a single-sample predictor (SSP): it assigns subtypes independently for each sample using within-sample, pairwise gene expression rules. This makes it robust to cohort composition and scaling.

Usage

```
BS_AIMS(se_obj)
```

Arguments

`se_obj` A `SummarizedExperiment` object containing:

- **Assay data**: A gene expression matrix with genes (Entrez IDs) as rows and samples as columns.
  - Expression values must be **positive** (e.g., FPKM or log2(FPKM+1)).
  - Values should not be gene-centered or globally scaled.

Value

A character vector of intrinsic subtype predictions assigned to each sample using the AIMS method.

References

Paquet ER, Hallett MT. *Absolute assignment of breast cancer intrinsic molecular subtype.* Journal of the National Cancer Institute. 2015;107(1):dju357. https://doi.org/10.1093/jnci/dju357

Examples

```
## Example using SummarizedExperiment input
data("OSLO2EMIT0obj")
res <- BS_AIMS(
    se_obj = OSLO2EMIT0obj$data_input$se_SSP
)
```

BS_cIHC *Conventional IHC Intrinsic Subtyping (BS_cIHC)*

Description

Implements the conventional immunohistochemistry-based (cIHC) intrinsic subtyping approach, which balances cohorts by estrogen receptor (ER) status before applying gene-expression–based subtyping. This method is useful for ER-skewed cohorts where assumptions of nearest-centroid classifiers are violated.

Usage

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

Arguments

se_obj	A SummarizedExperiment object containing: • **Assay data**: A log2-transformed, normalised expression matrix with genes (Gene Symbols) as rows and samples as columns. • **Column metadata** (colData): Must include: – "PatientID": Unique sample or patient identifier. – "ER": Estrogen receptor status, coded as "ER+" or "ER-".
Subtype	Logical. If TRUE, returns only the four main subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.
hasClinical	Logical. If TRUE, incorporates additional clinical variables from colData(se_obj). Required columns: • "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm). • "NODE": Lymph node status (0 = negative; $\geq$ 1 = positive). Must be numeric.
seed	Integer. Random seed for reproducibility of ER-balancing.

Value

A data.frame containing intrinsic subtype assignments estimated using the conventional IHC (cIHC) approach.

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):506–519. https://doi.org/10.1016/j.cell.2015.09.033

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

Implements an **iterative** version of the conventional IHC-based intrinsic subtyping approach. This method repeatedly balances samples by estrogen receptor (ER) status across multiple iterations, allowing refinement of subtype calls in ER-skewed cohorts. Users can customise the ER+/ER– ratio to match specific cohort assumptions (e.g., training distribution).

Usage

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

Arguments

`se_obj`	A `SummarizedExperiment` object containing: • **Assay data**: A log2-transformed, normalised expression matrix with genes (Gene Symbols) as rows and samples as columns. • **Column metadata** (`colData`): Must include: – `"PatientID"`: Unique sample or patient identifier. – `"ER"`: Estrogen receptor status, coded as `"ER+"` or `"ER-"`.
`iteration`	Integer. Number of iterations for the ER-balancing procedure. Default: 100.
`ratio`	Numeric. Target ER+/ER– ratio for balancing. Options: • `1:1`: Equal balancing. • `54:64`: Default; reflects the ER+/ER– ratio in the UNC232 training cohort.
`Subtype`	Logical. If `TRUE`, returns only the four main subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.
`hasClinical`	Logical. If `TRUE`, incorporates additional clinical variables from `colData(se_obj)`. Required columns: • "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm). • "NODE": Lymph node status (0 = negative; $\geq$ 1 = positive). Must be numeric.
`seed`	Integer. Random seed for reproducibility.

Value

A list containing:

- `subtypes`: Intrinsic subtype predictions across iterations.
- `confidence`: Confidence estimates for each assigned subtype.
- `ER_balance`: Proportions of ER+ and ER– subsets observed 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):346–352. 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, use iteration = 100
    Subtype = FALSE,
    hasClinical = FALSE
)
```

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

Description

Executes multiple intrinsic molecular subtyping methods in parallel. Users can either specify a set of classifiers directly, or enable the **AUTO mode**, which dynamically selects methods based on cohort composition (e.g., ER/HER2 distribution, subtype purity, subgroup size). AUTO reduces misclassification in skewed or subtype-specific cohorts by disabling methods whose assumptions are violated, but does not perform consensus voting—subtypes are still returned per method.

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.
`methods`	Character vector specifying the subtyping methods to run. Available options include: • `"parker.original"`: Original PAM50 (Parker et al., 2009). • `"genefu.scale"`: PAM50 (scaled version; Gendoo et al., 2016). • `"genefu.robust"`: PAM50 (robust version; Gendoo et al., 2016). • `"cIHC"`: Conventional ER-balancing with immunohistochemistry (Ciriello et al., 2015).

- `"cIHC.itr"`: Iterative ER-balancing (Curtis et al., 2012).
- `"PCAPAM50"`: PCA-based PAM50 using ESR1 balancing (Raj-Kumar et al., 2019).
- `"ssBC"`: Subgroup-specific gene-centering (Zhao et al., 2015).
- `"ssBC.v2"`: Updated subgroup-specific centering (Fernandez-Martinez et al., 2020).
- `"AIMS"`: Absolute Intrinsic Molecular Subtyping (Paquet & Hallett, 2015).
- `"sspbc"`: SSPBC, a large-cohort SSP trained on SCAN-B (Staaf et al., 2022).
- `"AUTO"`: Cohort-aware selection of compatible methods (must be the only entry).

Notes:

- If `"AUTO"` is selected, it must be the sole value in `methods`.
- Otherwise, at least **two** methods must be specified.

`Subtype` Logical. If `TRUE`, returns four subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.

`hasClinical` Logical. If `TRUE`, incorporates clinical data from `colData(se_obj)`. Required columns:

- "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm).
- "NODE": Lymph node status (0 = negative; $\geq$ 1 = positive).

Value

A list containing per-method subtype assignments for each sample.

References

Yang Q, Hartman J, Sifakis EG. *BreastSubtypeR: A Unified R/Bioconductor Package for Intrinsic Molecular Subtyping in Breast Cancer Research.* NAR Genomics and Bioinformatics. 2025. https://doi.org/10.1093/nargab/lqaf131. Selected as Editor's Choice.

Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol. 2009;27(8):1160-1167. 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):1097-1099. 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):506-519. https://doi.org/10.1016/j.cell.2015.09.033

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):346-352. https://doi.org/10.1038/nature10983

Zhao X, Rodland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Res. 2015;17(1):29. 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.* J Clin Oncol. 2020;38(36):4184-4197. 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):dju357. 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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

```
## Example: run multiple methods
data("OSLO2EMIT0obj")
methods <- c("parker.original", "genefu.scale", "genefu.robust")
res.test <- BS_Multi(
    data_input = OSLO2EMIT0obj$data_input,
    methods = methods,
    Subtype = FALSE,
    hasClinical = FALSE
)
```

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

Description

Implements the original PAM50 nearest-centroid classifier as described by Parker et al. (2009), along with supported calibration strategies and variations. This function assigns intrinsic breast cancer subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like, and optionally Normal-like).

Usage

```
BS_parker(
  se_obj,
  calibration = "None",
  internal = NA,
  external = NA,
  medians = NA,
  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`): Optional sample- or patient-level information.
`calibration`	Character. One of: • `"None"`: no centering/scaling. • `"Internal"`: center by a method derived from the current cohort (see `internal`). • `"External"`: center by medians from an external cohort (see `external`).

`internal` Internal calibration method used when `calibration = "Internal"`. Accepts:

- `NA` or `"medianCtr"` (identical): gene-wise median centering (as in Parker et al.).
- `"meanCtr"`: gene-wise z-scoring (mean 0, sd 1; as implemented in `genefu.scale`).
- `"qCtr"`: robust centering (quantile rescale with mq = 0.05; as in `genefu.robust`). Defaults to `NA` (median centering).

`external` Character string specifying the external calibration source.

- To use training cohort medians, provide the platform/column name.
- To supply user-defined medians, set `external = "Given.mdns"` and pass values via `medians`.

`medians` A matrix or data.frame of user-provided medians (required if `external = "Given.mdns"`).

- First column: 50 PAM50 genes.
- Second column: Corresponding median expression values.

`Subtype` Logical. If `TRUE`, assigns only the four main intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.

`hasClinical` Logical. If `TRUE`, incorporates clinical variables from `colData(se_obj)`. Required columns:

- "TSIZE": Tumor size (0 = ≤ 2 cm; 1 = > 2 cm).
- "NODE": Lymph node status (0 = negative; ≥ 1 = positive). Must be numeric.

Value

A list containing PAM50 intrinsic subtype calls using the Parker classifier and selected calibration strategy.

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 = NA, # NA is equal to "medianCtr"
    Subtype = FALSE,
    hasClinical = FALSE
)
```

BS_PCAPAM50 *PCA-PAM50 Intrinsic Subtyping (BS_PCAPAM50)*

Description

Implements the PCA-PAM50 method, which integrates **Principal Component Analysis (PCA)** of ESR1 expression to adjust for estrogen receptor (ER) imbalance prior to applying the PAM50 nearest-centroid classifier. This approach improves subtype consistency, particularly in ER-skewed cohorts.

Usage

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

Arguments

se_obj	A SummarizedExperiment object containing: • **Assay data**: A log2-transformed, normalised expression matrix with genes (Gene Symbols) as rows and samples as columns. • **Column metadata** (colData): Must include: – "PatientID": Unique sample or patient identifier. – "ER": Estrogen receptor status, coded as "ER+" or "ER-".
Subtype	Logical. If TRUE, returns only the four main subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.
hasClinical	Logical. If TRUE, incorporates additional clinical variables from colData(se_obj). Required columns: • "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm). • "NODE": Lymph node status (0 = negative; $\geq$ 1 = positive). Must be numeric.
seed	Integer. Random seed for reproducibility.

Value

A character vector of intrinsic subtype predictions assigned to each sample using 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.* Scientific Reports. 2019;9(1):1–12. 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

Implements the **subgroup-specific gene-centering (ssBC)** method for breast cancer intrinsic subtyping. The ssBC approach applies precomputed, subgroup-specific centering values to adjust PAM50 nearest-centroid classification when the study cohort is skewed relative to the original training cohort (e.g., ER-selected, HER2-enriched, or triple-negative cohorts).

Usage

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

Arguments

se_obj	A SummarizedExperiment object containing: • **Assay data**: A log2-transformed, normalised expression matrix with genes (Gene Symbols) as rows and samples as columns. • **Column metadata** (colData): If hasClinical = TRUE, must include: – "PatientID": Unique patient/sample identifier. – Depending on the chosen 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 "TN.v2".
s	Character. Specifies which subgroup-specific quantiles to use: • "ER", "TN": Original subgroup-specific quantiles (*Breast Cancer Research*, 2015). • "ER.v2", "TN.v2": Updated subgroup-specific quantiles (*Journal of Clinical Oncology*, 2020).
Subtype	Logical. If TRUE, returns only the four main subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like), excluding Normal-like.
hasClinical	Logical. If TRUE, incorporates additional clinical variables from colData(se_obj). Required columns: • "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm). • "NODE": Lymph node status (0 = negative; $\geq$ 1 = positive). Must be numeric.

Value

A character vector of intrinsic subtype predictions assigned to each sample using 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):29. 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;38(36):4184–4197. https://doi.org/10.1200/JCO.20.01276

Examples

```
## Example: Updated subgroup-specific quantiles (ER.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

Implements **SSPBC (Single Sample Predictor for Breast Cancer)**, a refinement of the original AIMS methodology trained on the large, population-based SCAN-B RNA-seq cohort. SSPBC provides robust single-sample predictions, independent of cohort composition, and supports multiple model variants for different applications.

Usage

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

Arguments

`se_obj`	A `SummarizedExperiment` object containing: • **Assay data**: A gene expression matrix with genes (Entrez IDs) as rows and samples as columns. – Expression values must be **positive** (e.g., FPKM or log2(FPKM+1)). – Values should not be gene-centered or globally scaled.
`ssp.name`	Character. Specifies the SSPBC model to use: • `"ssp.pam50"`: Predicts PAM50-based intrinsic subtypes. • `"ssp.subtype"`: Predicts Prosigna-like subtypes (four subtypes, excluding Normal-like).

Value

A character vector of intrinsic subtype predictions for each sample, 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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

```
## Example using SSPBC with the PAM50 model
data("OSLO2EMIT0obj")
res <- BS_sspbc(
    se_obj = OSLO2EMIT0obj$data_input$se_SSP,
    ssp.name = "ssp.pam50"
)
```

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

Description

Annotation table for GENCODE Human Release 27 genes (`Gene.ID`) used by StringTie summarisation. Includes HGNC, EntrezGene, and RefSeq identifiers derived from GENCODE v27 metadata.

Usage

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

Format

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

Details

Used by internal SSP application functions to translate identifiers prior to classification with SSP models.

Value

`Gene.ID.ann` Data frame of annotations for GENCODE v27 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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

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

iBreastSubtypeR *Launch the iBreastSubtypeR Shiny app*

Description

Starts the Shiny UI bundled with the BreastSubtypeR package. The launcher can (optionally) attach Shiny/Bslib so UI/server can use unqualified functions like `tags`, `icon`, `fileInput`, etc.

Usage

```
iBreastSubtypeR(
  attach = c("shiny", "bslib"),
  attach_tidyverse = FALSE,
  max_upload_mb = 1000
)
```

Arguments

`attach`	Character vector of packages to attach before launch. Defaults to c("shiny","bslib"). Set to character(0) to skip attaching.
`attach_tidyverse`	Logical; if TRUE and tidyverse is installed, it will be attached quietly for the session (purely optional).
`max_upload_mb`	Numeric; Shiny upload size limit (in MB). Default 1000.

Value

The value returned by `shiny::runApp()` (usually `invisible(NULL)`).

Examples

```
if (interactive()) {
    iBreastSubtypeR()
    iBreastSubtypeR(attach = character(0))
}
```

Mapping *Gene ID Mapping*

Description

Preprocesses and maps gene expression input to prepare for intrinsic subtyping workflows (NC- and SSP-based).

Usage

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

Arguments

se_obj	A SummarizedExperiment object containing: • **Assay data**: – If RawCounts = FALSE: assay() must contain log2-normalized expression (e.g., pre-normalized microarray/nCounter, or log2(FPKM+1) RNAseq). – If RawCounts = TRUE: assay() contains raw RNA-seq counts (see RawCounts). • **Row metadata** (required): – "probe": feature identifiers (e.g., gene symbols or probe IDs) – "ENTREZID": corresponding Entrez Gene IDs. – If row names are gene symbols, provide an additional SYMBOL column, renamed as probe. • **Column metadata** (optional): sample-level metadata in colData().
RawCounts	Logical. If TRUE, indicates that assay() holds raw RNA-seq counts. In this case, rowData() must also provide gene lengths (column "Length", in base pairs), used for: • NC-based methods: log2-CPM (upper-quartile normalization). • SSP-based methods: linear FPKM (not log-transformed).
method	Strategy for resolving duplicate probes/genes. Options: • "iqr": probe with highest interquartile range (short-oligo arrays, e.g., Affymetrix). • "mean": probe with highest mean expression (long-oligo arrays, e.g., Agilent/Illumina). • "max": probe with highest expression value (often used for RNA-seq). • "stdev": probe with highest standard deviation. • "median": probe with highest median expression.
impute	Logical. If TRUE, applies KNN-based imputation to missing values.
verbose	Logical. If TRUE, prints progress messages during execution.

Details

Mapping() supports multiple input types:

- **Raw RNA-seq counts** (with gene lengths): normalized to CPM (NC) or FPKM (SSP).
- **Precomputed log2(FPKM+1)**: used directly for NC; back-transformed for SSP.
- **log2-normalized microarray/nCounter data**: used directly for NC; back-transformed for SSP.

This design allows users to supply a single expression format, while BreastSubtypeR automatically applies method-specific preprocessing.

Value

A named list with:

se_NC `SummarizedExperiment` holding log2-transformed data prepared for NC-based methods (assay name: `counts`).

se_SSP `SummarizedExperiment` holding linear-scale data prepared for SSP-based methods (assay name: `counts`).

References

Yang Q, Hartman J, Sifakis EG. *BreastSubtypeR: A Unified R/Bioconductor Package for Intrinsic Molecular Subtyping in Breast Cancer Research.* NAR Genomics and Bioinformatics. 2025. https://doi.org/10.1093/nargab/lqaf131. Selected as Editor's Choice.

Examples

```
if (requireNamespace("SummarizedExperiment", quietly = TRUE)) {
    # Using example raw RNA-seq counts (with gene lengths)
    data("TCGABRCAobj")
    se_obj_counts <- TCGABRCAobj$se_obj[, 1:3] # tiny subset to keep checks fast
    res <- Mapping(se_obj_counts, RawCounts = TRUE)

    # Using example pre-normalized log2(FPKM+0.1)
    data("OSLO2EMIT0obj")
    se_obj_fpkm <- OSLO2EMIT0obj$se_obj[, 1:3] # tiny subset to keep checks fast
    res <- Mapping(se_obj_fpkm, RawCounts = FALSE)
}
```

`OSLO2EMIT0obj` *OSLO2EMIT0obj: Example dataset (OSLO2-EMIT0 cohort subset)*

Description

Example object derived from the OSLO2-EMIT0 cohort (Staaf et al., 2022). Includes a subset of normalized expression data, clinical metadata, feature annotations, and example outputs from `Mapping()` and `BS_Multi()`.

Usage

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

Format

A list with:

`se_obj` A `SummarizedExperiment` containing a subset of the log2-transformed, normalised expression matrix (log2(FPKM+0.1)) with `colData` clinical metadata and row-level feature annotations.

`data_input` Example output structure produced by `Mapping()`.

`res` Example results from `BS_Multi()` run in *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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

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

`sspbc.models` *sspbc.models: Short names for 11 SSPBC predictors*

Description

List of 11 single-sample predictor (SSP) models from Staaf et al. (2022), indexed by short names used by `sspbc`.

Usage

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

Format

An object of class `list` of length 11.

Details

Names correspond to short model identifiers. The contents are identical to `sspbc.models.fullname`, which uses full model names.

Value

`sspbc.models` Named list 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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

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

sspbc.models.fullname *sspbc.models.fullname: Full names for 11 SSPBC predictors*

Description

List of the same 11 SSP models (Staaf et al., 2022) indexed by full model names.

Usage

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

Format

An object of class `list` of length 11.

Details

Identical content to `sspbc.models` but with full model names as list keys.

Value

`sspbc.models.fullname`
: Named list 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):27. https://doi.org/10.1038/s41523-022-00465-3

Examples

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

TCGABRCAobj *TCGABRCAobj: Example dataset (TCGA-BRCA subset)*

Description

Example object derived from TCGA-BRCA. Includes a subset of normalized metadata

- raw counts (as a SummarizedExperiment), and example outputs from `Mapping()` and `BS_Multi()` to facilitate runnable examples.

Usage

```
data("TCGABRCAobj")
```

Format

A list with:

`se_obj` A `SummarizedExperiment` containing the integer raw-count matrix (top 5,000 variable genes), `rowData` with `probe`, `SYMBOL`, `ENTREZID`, `Length`, and `colData` with `PatientID`, `ER`, `PR`, `HER2`.

`data_input` Example `Mapping()` output created from `se_obj`.

`res` Example `BS_Multi()` results (e.g., run in *AUTO* mode).

Source

The Cancer Genome Atlas (TCGA) BRCA via GDC; counts summarized with `recount3`; clinical data retrieved with `TCGAbiolinks`.

References

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70. https://doi.org/10.1038/nature11412

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71. https://doi.org/10.1093/nar/gkv1507

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 2017;35(4):319–321. https://doi.org/10.1038/nbt.3838

Examples

```
library(BreastSubtypeR)
data("TCGABRCAobj")
names(TCGABRCAobj)
# str(TCGABRCAobj$se_obj); head(colData(TCGABRCAobj$se_obj))
```

`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