Package: BreastSubtypeR (via r-universe)

October 31, 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 1.5.12

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

Depends R (>= 4.5.0)

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

Suggests S4Vectors, htmltools, lifecycle, tidyverse, shiny (>= 1.9.1), bslib (>= 0.8.0), BiocStyle, knitr, rmarkdown, testthat (>= 3.2.0)

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

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

License GPL-3

VignetteBuilder knitr

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

Config/testthat/edition 3

LazyData FALSE

RoxygenNote 8.1.1

Config/pak/sysreqs libicu-dev libpng-dev perl zlib1g-dev

Repository https://bioc.r-universe.dev

Date/Publication 2026-10-31 18:48:15 UTC

RemoteUrl https://github.com/bioc/BreastSubtypeR

RemoteRef HEAD

RemoteSha 6b2a4ab16798493f87713a91cee4674b2ae4df8e

Contents

AIMSmodel . . . 3
BreastSubtypeR . . . 4
BreastSubtypeRobj . . . 5
BS_AIMS . . . 6
BS_cIHC . . . 8
BS_cIHC.itr . . . 11
BS_Multi . . . 16
BS_parker . . . 21
BS_PCAPAM50 . . . 26
BS_ssBC . . . 30
BS_sspbc . . . 34
Gene.ID.ann . . . 36
iBreastSubtypeR . . . 36
InputCoverage . . . 38
Mapping . . . 39
OSLO2EMIT0obj . . . 45
sspbc.models . . . 46
sspbc.models.fullname . . . 47
TCGABRCAobj . . . 47
Vis_boxplot . . . 48
Vis_heatmap . . . 49
Vis_Multi . . . 50
Vis_PCA . . . 51
Vis_pie . . . 52

Index **54**

`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`. The bundled `k = 20` identifies the fitted-model setting; its `all.pairs` vector and fitted naive Bayes tables contain all 100 rule predictors, involving 151 unique genes. Thus `k` is not the total rule denominator. The model assigns Basal-like, HER2-enriched, LumA, LumB and Normal-like subtypes independently for each sample. See `InputCoverage` for descriptive input reporting.

Value

`all.pairs`	Character vector of the 100 AIMS rules (EntrezID comparisons).
`k`	Integer fitted-model setting (20 in the bundled model), not the total rule count.
`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. **AUTO mode** selects methods using implemented receptor-annotation and subgroup-size rules and returns per-method calls without a consensus prediction.

By standardizing input handling, applying method-specific normalization, and providing probe-to-gene mapping, BreastSubtypeR offers a common workflow for comparing published classifiers. A companion Shiny app (**iBreastSubtypeR**) runs in the user's local R session. Outputs are for research use and do not replace validated clinical diagnostics. AUTO selection does not establish complete input adequacy or clinical validity. `InputCoverage()` provides descriptive reporting of processed inputs without changing method eligibility.

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-normalized 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. **Visualization**: Summarize and interpret subtyping results with `Vis_Multi`.

Key Features:

- **Multi-method framework**: Ten published NC-based and SSP-based method IDs under one interface.
- **AUTO mode**: Uses implemented cohort rules based on receptor annotations and subgroup sizes; returns separate method calls.
- **standardized normalization**: Upper-quartile log2-CPM for NC-based methods; FPKM for SSP-based methods.
- **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: Emmanouil G. Sifakis <emmanouil.sifakis@ki.se> (ORCID)

Authors:

- Qiao Yang <yq.kiuo@gmail.com> (ORCID)
- 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_postMean` `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` Original and updated subgroup-specific quantiles used by `ssBC` and `ssBC.v2`; see Details.

`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.

Details

The original `ER` and `TN` methods retain the reference quantiles for subgroup-specific gene centering introduced by Zhao et al. (2015). Fernandez-Martinez et al. (2020) described a HER2/ER-specific application. The five updated columns match the public `SIGMA.txt` reference associated with Fernandez-Martinez et al. (2024), across all 50 PAM50 genes: `ER.v2` uses `ERneg_HER2neg`, `ERpos_HER2neg`, `HER2pos_ERneg` and `HER2pos_ERpos`; `TN.v2` uses `TNBC`.

Source

Perou Lab normalization repository. The updated columns were verified against SIGMA.txt at commit 9ba84eb.

References

Parker JS, Mullins M, Cheang 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, Rødland EA, Tibshirani R, Plevritis S. Molecular subtyping for clinically defined breast cancer subgroups. *Breast Cancer Res*. 2015;17:29. https://doi.org/10.1186/s13058-015-0520-4

Fernandez-Martinez A, Krop IE, Hillman DW, 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(35):4184-4193. https://doi.org/10.1200/JCO.20.01276

Fernandez-Martinez A, Rediti M, Tang G, et al. Tumor Intrinsic Subtypes and Gene Expression Signatures in Early-Stage ERBB2/HER2-Positive Breast Cancer: A Pooled Analysis of CALGB 40601, NeoALTTO, and NSABP B-41 Trials. *JAMA Oncol*. 2024;10(5):603-611. https://doi.org/10.1001/jamaoncol.2023.73

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.

Details

The bundled k = 20 names the fitted-model setting and output column; its all.pairs vector and fitted naive Bayes tables contain 100 rule predictors. k is not the total rule denominator. InputCoverage() reports represented genes and rule evaluability without modifying this model. If no model rule is evaluable for a sample, its class, selected-class probability and all posterior probabilities are NA. Both TRUE and FALSE comparisons count as evaluable; with at least one, native predictions are retained. The posterior probabilities come from the fitted model and are distinct from input coverage, multi-method entropy and Parker Confidence. They are not externally calibrated measures of clinical reliability.

Value

A **list** with the following elements:

- cl: Character matrix of AIMS subtype calls, with one row per sample (named by sample ID) and one column named "20" for the model's rule-pair setting. Calls are "Basal", "Her2", "LumA", "LumB", or "Normal".
- prob: Numeric matrix of posterior probabilities for the assigned calls, with the same dimensions and dimnames as cl.
- all.probs: Named list with element "20" containing a posterior-probability matrix (rows = samples, columns = the five subtypes). Matrix dimensions are retained for a single sample.
- rules.matrix: 0/1 matrix of the 100 AIMS rules used for assignment (rows = rules, columns = samples); 1 indicates the rule evaluated to TRUE.
- data.used: Expression values actually used to evaluate the rules (filtered/ordered subset aligned to the AIMS gene set).
- EntrezID.used: Character vector of Entrez IDs used by AIMS.

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
)
## A single sample retains its identifier and output dimensions
single_input <- Mapping(OSLO2EMIT0obj$se_obj[, 1], RawCounts = FALSE)
single <- BS_AIMS(single_input$se_SSP)
single$cl
dim(single$all.probs[["20"]])
stopifnot(identical(rownames(single$cl), colnames(single_input$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, normalized expression matrix with genes (Gene Symbols) as rows and samples as columns.
- **Column metadata** (`colData`): Must include:
  - `"PatientID"`: Unique sample or patient identifier. Values are converted to character before name-based indexing, so factor labels (including unused levels) are matched to expression column names. Missing, empty, duplicated or unmatched identifiers are rejected.
  - `"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": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes.

`NODE` affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric `NODE` values give a warning and a missing additional group; numerical ROR scores remain available. Missing `NODE` (`NA`/`NaN`) gives a missing additional group without a NODE warning. See `Mapping()` for clinical missing-value behavior.

`seed` Integer. Random seed for reproducibility of ER-balancing.

Value

A list with the following elements:

- BS.all: data.frame with PatientID, BS (five-class call) and, when Subtype = TRUE, BS.Subtype (four-class call).
- score.ROR: data.frame with the per-sample centroid correlations, calls, confidence, ESR1/ERBB2 expression and the ROR-S and ROR-P scores and risk groups; with hasClinical = TRUE also the ROR-C and ROR-PC scores and risk groups.
- mdns: data.frame of gene medians used for centering (reference platform medians plus the ER-balanced cohort medians).
- outList: list with the internal nearest-centroid results (predictions, distances, centroids, ...).

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional ROR-PC (Subtype + Clinic + Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker ROR-PC (Subtype + Clinic + Prolif) is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class bsr_experimental_ror_pc_warning once per top-level call when a nonempty returned score.ROR contains both legacy columns. BS_Multi() aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in Mapping(), with hasClinical = FALSE, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign high with nonmissing tumor size.

This is an intentional warning: `options(warn = 2)` turns it into an error. An expert workflow can handle only the `bsr_experimental_ror_pc_warning` class:

- Establish a handler with `withCallingHandlers()`.
- Within that handler, call `invokeRestart("muffleWarning")`.

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCA-PAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research ROR-PC (Subtype + Clinic + Prolif.Subtype) is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged TCGABRCAobj example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (NA) proliferation score. Every exported numerical ROR score is finite or NA: a NaN or infinite result, including overflow in the final score transformation, has a missing group rather than low, med or high. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

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. Each iteration subsamples the ER group that is over-represented relative to an ER+ : ER- target (ratio, default 54/64); er_subsample = "larger" restores the behaviour of BreastSubtypeR 1.5.5 and earlier.

Usage

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

Arguments

se_obj	A SummarizedExperiment object containing: • **Assay data**: A log2-transformed, normalized 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-". PatientID values are converted to character before selecting expression columns in every iteration. Factor labels, including unused levels, are accepted. IDs must be non-missing, non-empty, unique and present in the assay column names. Valid character-ID results, ER balancing, RNG behavior and numerical ROR formulas are unchanged.
iteration	Integer. Number of iterations for the ER-balancing procedure. Default: 100.
ratio	A single finite, strictly positive number. Its meaning depends on er_subsample: • er_subsample = "target" (default): the ER+ : ER- target of the balancing subset, a single positive number. The default 54 / 64 (about 0.84) is the ER+ : ER- distribution of the UNC PAM50 reference set; 1 gives equal group sizes. In each iteration the ER group that is over-represented relative to the target is subsampled without replacement, to ceiling(ratio * n_ERneg) ER+ or ceiling(n_ERpos / ratio) ER- samples, and every sample of the other group is kept. A cohort already at the target (within this rounding) is not subsampled; every iteration then uses all samples. With the default, for example, 100 ER+ and 40 ER- samples give 34 ER+ : 40 ER-, 40 ER+ and 100 ER- samples give 40 ER+ : 48 ER-, and 54 ER+ and 64 ER- samples are used in full. • er_subsample = "larger": the size of the subsample of the numerically larger ER group (ER+ when the groups are equal) relative to the smaller group, which is kept in full: ceiling(ratio * n_smaller) samples are drawn. ratio must not exceed n_larger / n_smaller. This is the behaviour of BreastSubtypeR 1.5.5 and earlier, in which the subset follows the ER+ : ER- target only when ER+ samples are at least as many as ER- samples. Write 54 / 64, not 54:64: in R, 54:64 is the sequence 54, 55, ..., 64 and gives an error. Values above one are permitted in target mode; legacy mode retains

the feasibility limit described above. Extreme positive ratios can retain very few samples from one ER group, making the calibration subset poorly representative. Acceptance of a ratio does not establish calibration adequacy for a particular cohort.

`Subtype` Logical. If `TRUE`, also returns four-class consensus calls (Luminal A, Luminal B, HER2-enriched, Basal-like) in `BS.Subtype`. The five-class consensus calls in `BS` are retained.

`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": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes.

`NODE` affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric `NODE` values give a warning and a missing additional group; numerical ROR scores remain available. Missing `NODE` (`NA`/`NaN`) gives a missing additional group without a NODE warning. See `Mapping()` for clinical missing-value behavior.

`seed` Integer. Random seed for reproducibility.

`er_subsample` Character. Which ER group is subsampled in each iteration (see `ratio`):

- `"target"` (default): the group over-represented relative to the ER+ : ER- target.
- `"larger"`: the numerically larger group, as in BreastSubtypeR 1.5.5 and earlier; it reproduces earlier results exactly.

With the default `ratio = 54 / 64` and otherwise identical inputs, seed and settings, the two modes give the same results when ER+ samples are at least as many as ER- samples. This is not a guarantee for other values of `ratio`: `"larger"` rejects a `ratio` above `n_larger / n_smaller`, which `"target"` accepts.

Value

A list with the following elements:

- `BS.all`: `data.frame` with `PatientID`, `BS` (five-class consensus call across iterations) and, when `Subtype = TRUE`, `BS.Subtype` (four-class consensus call). The consensus is the most frequent call over the iterations; ties are resolved in favour of the alphabetically first subtype (Basal, Her2, LumA, LumB, Normal).
- `score.ROR`: `data.frame` with the per-sample centroid correlations (averaged over the iterations whose call equals the consensus call), calls, confidence, ESR1/ERBB2 expression and the ROR-S and ROR-P scores and risk groups computed from these averages; with `hasClinical = TRUE` also the ROR-C and ROR-PC scores and risk groups.
- `outList`: list with the consensus predictions, the averaged test data and distances, and the centroids.

- BS.itr.keep: character matrix of the five-class calls per sample (rows) and iteration (columns).
- BS.itr.keep.Subtype: the four-class analogue, present when Subtype = TRUE.

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional ROR-PC (Subtype + Clinic + Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker ROR-PC (Subtype + Clinic + Prolif) is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class bsr_experimental_ror_pc_warning once per top-level call when a nonempty returned score.ROR contains both legacy columns. BS_Multi() aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in Mapping(), with hasClinical = FALSE, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign high with nonmissing tumor size.

This is an intentional warning: options(warn = 2) turns it into an error. An expert workflow can handle only the bsr_experimental_ror_pc_warning class:

- Establish a handler with withCallingHandlers().
- Within that handler, call invokeRestart("muffleWarning").

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCA-PAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce

missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

References

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. *The genomic and transcriptomic architecture of 2,000 breast tumors 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 one call. Users can either specify a set of classifiers directly, or enable the **AUTO mode**, which selects methods using implemented receptor-annotation and subgroup-size rules. AUTO returns per-method calls without a consensus prediction. Selection does not establish complete input adequacy or clinical validity. `InputCoverage()` reports processed-input coverage separately and does not change eligibility, calibration or predictions.

Usage

```
BS_Multi(
  data_input,
  methods = "AUTO",
  Subtype = FALSE,
  hasClinical = FALSE,
  er_subsample = c("target", "larger")
)
```

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, 2024); see `BS_ssBC()` for quantile provenance.
- `"AIMS"`: Absolute Intrinsic Molecular Subtyping (Paquet & Hallett, 2015).
- `"sspbc"`: SSPBC, a large-cohort SSP trained on SCAN-B (Staaf et al., 2022).
- `"AUTO"`: Selection using implemented cohort rules (must be the only entry).

Notes:

- If `"AUTO"` is selected, it must be the sole value in `methods`.
- Otherwise, at least **two distinct** methods must be specified.
- Duplicate entries are rejected rather than run repeatedly.

`Subtype` Logical. If `TRUE`, also returns a four-class summary in `res_subtypes.Subtype` (Luminal A, Luminal B, HER2-enriched, Basal-like). The five-class `res_subtypes` table is retained. AIMS has no four-class model: its Normal-like calls become `NA` in the four-class summary, while its per-method result retains the original five-class calls.

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

- "TSIZE": Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm).
- "NODE": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes.

`NODE` affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric `NODE` values give a warning and a missing additional group; numerical ROR scores remain available. Missing `NODE` (`NA`/`NaN`) gives a missing additional group without a NODE warning. See `Mapping()` for clinical missing-value behavior.

er_subsample — Character, passed to BS_cIHC.itr() and used by no other method: "target" (default) subsamples the ER group that is over-represented relative to the 54/64 ER+ : ER- target; "larger" subsamples the numerically larger group, as in BreastSubtypeR 1.5.5 and earlier, to reproduce earlier results.

Details

With the default AUTO routing thresholds, an eligible mixed cohort is ER-balanced only when 0.39 < ER-positive fraction < 0.69; equality at either boundary is outside that interval. The balanced and unbalanced mixed-cohort panels contain ten and eight methods, respectively, and both include cIHC.itr when the ER-group size gates are met. TN-only and HER2+ cohort routes are evaluated separately. See the package vignette for cohort-specific eligibility and minimum subgroup sizes.

AUTO subcohort results for ssBC and ssBC.v2 are restored by character PatientID labels, including factor labels with unused levels. Restoration requires complete, unique sample identities. Ineligible samples retain missing calls; subgroup eligibility and prediction algorithms are unchanged.

For ssBC.v2, Fernandez-Martinez et al. (2020) described the earlier HER2/ER-specific application. The updated quantiles match the public SIGMA.txt reference in the Perou Lab normalization repository, associated with Fernandez-Martinez et al. (2024). ER.v2 uses the four ER/HER2 columns and TN.v2 uses TNBC; see BS_ssBC() for their names. The original ssBC methods retain the Zhao et al. (2015) reference quantiles.

Value

A list containing per-method subtype assignments for each sample. res_subtypes has samples as rows and one column per executed method plus an entropy column: the raw Shannon entropy in bits of the calls that the executed methods returned for that sample. It is 0 when every contributing method agrees and NA when no method returned a call. The entropy is computed over the contributing methods only, so a low value obtained from few contributing methods reflects few available calls rather than strong agreement; the per-method columns show which methods contributed. When Subtype = TRUE, res_subtypes.Subtype has the same orientation and its entropy is calculated from the available four-class calls.

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional ROR-PC (Subtype + Clinic + Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker `ROR-PC (Subtype + Clinic + Prolif)` is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class `bsr_experimental_ror_pc_warning` once per top-level call when a nonempty returned `score.ROR` contains both legacy columns. `BS_Multi()` aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in `Mapping()`, with `hasClinical = FALSE`, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign `high` with nonmissing tumor size.

This is an intentional warning: `options(warn = 2)` turns it into an error. An expert workflow can handle only the `bsr_experimental_ror_pc_warning` class:

- Establish a handler with `withCallingHandlers()`.
- Within that handler, call `invokeRestart("muffleWarning")`.

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCA-PAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

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, Cheang 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 tumors reveals novel subgroups.* Nature. 2012;486(7403):346-352. https://doi.org/10.1038/nature10983

Zhao X, Rødland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Res. 2015;17:29. https://doi.org/10.1186/s13058-015-0520-4

Fernandez-Martinez A, Krop IE, Hillman DW, 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(35):4184-4193. https://doi.org/10.1200/JCO.20.01276

Fernandez-Martinez A, Rediti M, Tang G, et al. *Tumor Intrinsic Subtypes and Gene Expression Signatures in Early-Stage ERBB2/HER2-Positive Breast Cancer: A Pooled Analysis of CALGB 40601, NeoALTTO, and NSABP B-41 Trials.* JAMA Oncol. 2024;10(5):603-611. https://doi.org/10.1001/jamaoncol.2023.7304

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, Hakkinen 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;8:94. 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 = $\leq$ 2 cm; 1 = $>$ 2 cm). • "NODE": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes.

NODE affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric NODE values give a warning and a missing additional group; numerical ROR scores remain available. Missing NODE (NA/NaN) gives a missing additional group without a NODE warning. For NODE = 0 or 1–3, an undefined numerical NA or NaN additional ROR-PC score has an actual missing risk category, not the text "NaN". The numerical score is unchanged. The independent NODE >= 4 high-risk rule still applies with nonmissing tumor size, even if the score is missing. See Mapping() for clinical missing-value behavior.

Details

The Confidence column is round(1 - p, 2), where p is the Spearman correlation-test p-value for the processed sample and its assigned centroid over matched, complete gene pairs. It is NA when a valid test cannot be computed, including insufficient or constant values. This is a test-based summary, not a calibrated subtype probability or clinical reliability. It differs from InputCoverage(), multi-method entropy and SSP model probabilities. No score or classifier is changed by this interpretation.

Value

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

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional ROR-PC (Subtype + Clinic + Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker ROR-PC (Subtype + Clinic + Prolif) is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien

implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class `bsr_experimental_ror_pc_warning` once per top-level call when a nonempty returned `score.ROR` contains both legacy columns. `BS_Multi()` aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in `Mapping()`, with `hasClinical = FALSE`, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign `high` with nonmissing tumor size.

This is an intentional warning: `options(warn = 2)` turns it into an error. An expert workflow can handle only the `bsr_experimental_ror_pc_warning` class:

- Establish a handler with `withCallingHandlers()`.
- Within that handler, call `invokeRestart("muffleWarning")`.

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCA-PAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses

only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

References

- Parker JS, Mullins M, Cheang 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, normalized expression matrix with genes (Gene Symbols) as rows and samples as columns. • **Column metadata** (colData): Must include: – "PatientID": Unique sample or patient identifier. Values are converted to character and matched to assay column names, including factor labels rather than their integer codes. Missing, empty, duplicated or unmatched identifiers give an error before classification. – "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": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes. NODE affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric NODE values give a warning and a missing additional group; numerical ROR scores remain available. Missing NODE (NA/NaN) gives a missing additional group without a NODE warning. See Mapping() for clinical missing-value behavior.
seed	Integer. Random seed for reproducibility.

Details

The PC1-guided step uses the IHC-defined ER status as the reference labels for the cutoff search and for building the ER-balanced gene-centering set. `BS_PCAPAM50()` derives those labels from the `ER` column, so a sample whose ER status is missing or not coded as `"ER+"`/`"ER-"` has no evaluable label. Such samples are excluded from the PC1 axis check, from the misclassification-minimizing cutoff search and from the ER-balanced median set, and a message reports how many were excluded; they are still classified and appear in the returned calls. Cohorts with a complete ER column are unaffected. The reference implementation (Raj-Kumar et al., CRAN package PCAPAM50) expects a complete, user-supplied IHC subtype column in which ER-positive subtypes start with "L" and all other labels are ER-negative; it does not define behaviour for missing labels.

Before second-stage sampling, complete intermediate rows are ordered by character `PatientID`, matching the first-step merge in PCAPAM50 1.0.4 under the caller's collation. This can change selected samples, centering medians, subtype calls and derived scores relative to earlier package behavior. The first-step calculations and labels are unchanged. The caller's locale is not changed; universal input-permutation or cross-locale equivalence is not guaranteed. This ordering correction is independent of the shared original ROR standardization described under Proliferation panel coverage.

After both PCA-PAM50 steps complete, all sample-indexed outputs follow the input sample order: rows of `BS.all` and `score.ROR`, prediction vectors and distance-matrix rows in `outList`, and columns of `outList$testData`. Gene medians and centroids are not reordered. This final reordering leaves internal calibration and sampling, and per-sample values for the same input, seed and environment, unchanged. It is separate from the intermediate ordering used before second-stage sampling.

Value

A list with the following elements:

- `BS.all`: `data.frame` with `PatientID`, `BS` (five-class call) and, when `Subtype = TRUE`, `BS.Subtype` (four-class call).
- `score.ROR`: `data.frame` with the per-sample centroid correlations, calls, confidence, ESR1/ERBB2 expression and the ROR-S and ROR-P scores and risk groups; with `hasClinical = TRUE` also the ROR-C and ROR-PC scores and risk groups.
- `mdns.fl`: `data.frame` of gene medians used for centering (reference platform medians plus the PCA-PAM50 refined cohort medians).
- `outList`: list with the internal nearest-centroid results (`predictions`, `distances`, `centroids`, ...).

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional `ROR-PC (Subtype + Clinic + Prolif.Subtype)` score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/

s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker `ROR-PC (Subtype + Clinic + Prolif)` is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class `bsr_experimental_ror_pc_warning` once per top-level call when a nonempty returned `score.ROR` contains both legacy columns. `BS_Multi()` aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in `Mapping()`, with `hasClinical = FALSE`, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign `high` with nonmissing tumor size.

This is an intentional warning: `options(warn = 2)` turns it into an error. An expert workflow can handle only the `bsr_experimental_ror_pc_warning` class:

- Establish a handler with `withCallingHandlers()`.
- Within that handler, call `invokeRestart("muffleWarning")`.

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCAPAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

References

Raj-Kumar PK, Liu J, Hooke JA, Kovatich AJ, Kvecher L, Shriver CD, Hu H. *PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B.* Scientific Reports. 2019;9:7956. 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, normalized 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 quantiles matching the public `SIGMA.txt` reference associated with Fernandez-Martinez et al. (2024), following the HER2/ER-specific application described by Fernandez-Martinez et al. (2020). See Details for the reference resource and column mapping.

`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": Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes.

`NODE` affects only the additional subtype-specific risk group (Experimental 18-gene ROR-PC group), using categories 0, 1–3 and 4 or more. It does not enter any numerical ROR score. This group requires finite nonnegative integer-valued counts. Negative, fractional or infinite numeric `NODE` values give a warning and a missing additional group; numerical ROR scores remain available. Missing `NODE` (`NA`/`NaN`) gives a missing additional group without a NODE warning. See `Mapping()` for clinical missing-value behavior.

Details

Zhao et al. (2015) introduced subgroup-specific gene centering. Fernandez-Martinez et al. (2020) described its HER2/ER-specific application. The updated quantiles match the public `SIGMA.txt` reference in the Perou Lab normalization repository, associated with Fernandez-Martinez et al. (2024). For the 50 PAM50 genes, `ER.v2` uses `ERneg_HER2neg`, `ERpos_HER2neg`, `HER2pos_ERneg` and `HER2pos_ERpos`; `TN.v2` uses `TNBC`. The original `ER` and `TN` reference quantiles remain separate. See BreastSubtypeRobj for the pinned numerical reference.

Value

A list with the following elements:

- `BS.all`: `data.frame` with `PatientID`, `BS` (five-class call) and, when `Subtype = TRUE`, `BS.Subtype` (four-class call). Samples whose subgroup (`s`) status is missing receive `NA`.
- `score.ROR`: `data.frame` with the per-sample centroid correlations, calls, confidence, ESR1/ERBB2 expression and the ROR-S and ROR-P scores and risk groups; with `hasClinical = TRUE` also the ROR-C and ROR-PC scores and risk groups.
- `mdns`: `data.frame` of the precomputed subgroup-specific quantiles used for gene centering.
- `outList`: list with the internal nearest-centroid results (`predictions`, `distances`, `centroids`, ...).

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional `ROR-PC (Subtype + Clinic`

+ Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker ROR-PC (Subtype + Clinic + Prolif) is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class bsr_experimental_ror_pc_warning once per top-level call when a nonempty returned score.ROR contains both legacy columns. BS_Multi() aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in Mapping(), with hasClinical = FALSE, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign high with nonmissing tumor size.

This is an intentional warning: options(warn = 2) turns it into an error. An expert workflow can handle only the bsr_experimental_ror_pc_warning class:

- Establish a handler with withCallingHandlers().
- Within that handler, call invokeRestart("muffleWarning").

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and ROR-PC (Subtype + Clinic + Prolif) grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of low, med or high. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before 100 * (score + 0.35) / 0.85, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This

limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCAPAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

References

Zhao X, Rødland EA, Tibshirani R, Plevritis S. *Molecular subtyping for clinically defined breast cancer subgroups.* Breast Cancer Research. 2015;17:29. https://doi.org/10.1186/s13058-015-0520-4

Fernandez-Martinez A, Krop IE, Hillman DW, 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(35):4184-4193. https://doi.org/10.1200/JCO.20.01276

Fernandez-Martinez A, Rediti M, Tang G, et al. *Tumor Intrinsic Subtypes and Gene Expression Signatures in Early-Stage ERBB2/HER2-Positive Breast Cancer: A Pooled Analysis of CALGB 40601, NeoALTTO, and NSABP B-41 Trials.* JAMA Oncology. 2024;10(5):603-611. https://doi.org/10.1001/jamaoncol.2023.7304

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).

Details

If no model rule is evaluable for a sample, its class, selected-class probability and all posterior probabilities are NA. Both TRUE and FALSE comparisons count as evaluable; with at least one, native predictions are retained. Posterior probabilities are conditional on the fitted naive Bayes model and its native handling of the available rule inputs. They differ from InputCoverage(), multi-method entropy and Parker Confidence; they do not establish externally validated predictive or clinical reliability.

Value

A **list** with the following elements:

- cl: Molecular class identified by the sspbc models for each sample (one of "Basal", "Her2", "LumA", "LumB", with or without "Normal").
- prob: Numeric vector of posterior probabilities corresponding to the assigned subtype in cl (one value per sample).
- all.probs: Matrix of posterior probability values for all samples and all subtypes (rows = samples, columns = subtypes).
- rules.matrix: Binary (0/1) matrix of the pairwise gene-expression rules (*gene A* < *gene B*) used for assignment (rows = rules, columns = samples); 1 indicates the rule evaluated to TRUE for that sample.
- data.used: Expression values actually used to evaluate the simple rules.
- EntrezID.used: Character vector of EntrezGene IDs used for rule evaluation.

References

Staaf J, Hakkinen 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;8:94. 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, Hakkinen 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;8:94. 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 checks and loads dependency namespaces; `shiny::runApp()` starts the local application.

Usage

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

Arguments

attach	Character vector of package namespaces to load before launch. Defaults to c("shiny","bslib"). Set to character(0) to skip this explicit dependency check; the application still requires its UI dependencies. The argument name is retained for compatibility.
attach_tidyverse	Logical; if TRUE and tidyverse is installed, it will have its namespace loaded quietly (purely optional).
max_upload_mb	Numeric; Shiny upload size limit (in MB). Default 1000.

Details

Load the bundled example or upload expression, clinical and annotation tables, then select **Preprocess & map** before running subtyping. Replacing an input, reloading the example or changing the input data type clears dependent mapping, predictions, plots and exports. Relevant prediction-setting changes clear predictions and plots while preserving valid mapped inputs. A failed prediction cannot leave old results available.

Download controls are shown only for a supported, current prediction. AUTO, AIMS and SSPBC provide calls only; supported single NC methods also provide full metrics. Server-side guards reject stale or unsupported downloads even when an earlier download URL is requested directly.

For cIHC.itr, the ratio is the ER-positive : ER-negative target, defaults to `54 / 64`, and must be finite and strictly positive. Values above one are allowed. The Shiny iteration control defaults to 100, has minimum 10 and step 10, and accepts whole numbers from 10 to 2147483647, including values such as 11. These UI validation rules do not change the R API. Invalid prediction settings produce an actionable error; correct them and rerun without reloading the page or remapping valid inputs.

Value

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

See Also

`BS_AIMS()`, `BS_cIHC.itr()`, `BS_Multi()`

Examples

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

`InputCoverage` *Describe model-specific coverage of processed expression inputs*

Description

Describe model-specific coverage of processed expression inputs

Usage

```
InputCoverage(data_input)
```

Arguments

`data_input`	The named list returned by `Mapping()`, containing `se_NC` and `se_SSP` `SummarizedExperiment` objects.

Details

This is descriptive reporting for four input models, independently of which methods AUTO selects or a user runs. It does not run a classifier, change predictions, assess calibration or prediction reliability, or impose coverage thresholds or method-exclusion policies.

NC-based coverage concerns the shared 50-gene PAM50 subtype reference. It does not assess calibration feasibility or coverage of every ROR component.

Gene representation and rule evaluability are different properties. Gene and rule denominators come from the bundled references and fitted model rule sets. Gene presence is structural: a represented gene with zero or missing expression is distinct from an absent gene. Genes are matched by their classifier-specific identifiers, including the native SSPBC Entrez-prefix translation. Duplicate identifiers are counted once; SSP duplicate expression rows follow the native per-sample selection.

A rule is evaluable when the native strict `gene A < gene B` comparison yields either `TRUE` or `FALSE`. Equal values, including zero versus zero, give an evaluable `FALSE`. Missing genes and `NA`/`NaN` comparisons are unavailable; infinities follow the native comparison semantics. Counts may therefore differ across samples. No low-expression cutoff is used.

These diagnostics describe the processed matrices in `data_input`. `Mapping()` may already have performed imputation, so the report cannot distinguish originally observed values from imputed values. It neither preprocesses the expression again nor changes inputs, models or RNG state.

Value

A data.frame with rows in this order: `NC-based (PAM50)`, `AIMS`, `SSPBC (ssp.subtype)`, and `SSPBC (ssp.pam50)`. The `model` column is character; all count columns are integer:

genes_present, genes_total, genes_absent Unique model genes represented in the processed input, included in the bundled model, and absent from the input, respectively.

rules_evaluable_min, rules_evaluable_max Smallest and largest number of evaluable gene-pair rules across samples.

rules_total Number of gene-pair rules in the bundled model.

All three rule columns are NA for NC-based (PAM50). Malformed or unavailable processed inputs produce an error; an input with no gene rows reports zero represented genes and zero evaluable rules.

Examples

```
data("OSLO2EMIT0obj")
data_input <- Mapping(OSLO2EMIT0obj$se_obj[, 1:3], verbose = FALSE)
InputCoverage(data_input)
```

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 = FALSE,
  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**: sample-level metadata in colData(). Specifically:
  - PatientID (**required**).
  - **Receptor fields required by specific methods** (and by BS_Multi() / AUTO):
    - ER - required for ssBC, cIHC / cIHC.itr, PCAPAM50, and AUTO.

    * HER2 - required for ssBC with s="ER.v2" and for AUTO.
    * TN - per-sample annotation required for ssBC with s="TN" / "TN.v2"; AUTO uses TNBC-specific cohort handling only when all evaluable TN annotations indicate TN.
  - **Clinical covariates for ROR (NC-based methods only)**:
    * TSIZE: Tumor size (0 = $\leq$ 2 cm; 1 = $>$ 2 cm). Must be numeric (0/1).
    * NODE: Number of positive lymph nodes. Supply the actual nonnegative integer count as a numeric value when available. Binary 0/1 is accepted but cannot distinguish 1–3 from 4 or more positive nodes. The additional ROR-PC group requires finite nonnegative integer-valued counts; see Details.
  - **Preferred canonical receptor coding**:
    * ER: ER+, ER-
    * HER2: HER2+, HER2- (note: **HER2 "2+" is not coerced**)
    * TN: TN, nonTN
  - **Automatic normalization**: Mapping() maps common variants to canonical forms (e.g., Pos/Neg, Positive/Negative, 1/0, True/False). Ambiguous HER2="2+" is **not** remapped.
  - These receptor fields are **not required** for purely SSP methods (AIMS, sspbc) or for PAM50 when used without AUTO/ROR; they are recommended so AUTO can select suitable methods.

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, i.e. several rows mapping to the same Entrez ID. "mean", "median", "iqr" and "stdev" follow collapseIDs() of the original PAM50 bioclassifier code (Parker et al., 2009); "max" is an addition of this package. Options:
  - "mean": per-sample mean across the duplicate probes (collapses the probes into one row; long-oligo arrays, e.g., Agilent/Illumina).
  - "median": per-sample median across the duplicate probes (collapses the probes into one row).
  - "iqr": keeps the single probe with the highest interquartile range (short-oligo arrays, e.g., Affymetrix).
  - "stdev": keeps the single probe with the highest standard deviation.
  - "max": keeps the single probe with the largest row sum across samples (often used for RNA-seq). A group with one row is retained without ranking, including any remaining NA or NaN measurements. If every probe of a multi-probe group has an undefined row sum (NA, NaN or Inf - Inf), Mapping() stops with an error naming the Entrez ID instead of selecting a probe.

impute
: Logical; default FALSE preserves missing measurements after duplicate resolution. For nonraw input, TRUE requests optional gene-level KNN with fixed k =

10, using the full eligible supplied donor matrix. Only originally missing measurements in represented genes selected for the existing NC/SSP outputs are estimated. Complete required outputs bypass KNN.

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.

Missingness and optional imputation. Absent genes are never synthesized. Duplicate methods retain their existing rules: mean/median use available probes, while max requires a defined total for a duplicate probe group. Donors are prepared only when requested output measurements remain missing. Unusable nonprotected background genes (failed duplicate resolution, infinite observations or entirely missing profiles) are excluded with diagnostics; required genes are protected. Otherwise usable partially missing donors are retained, including profiles with more than 50% missingness.

KNN compares each target with every other eligible original gene profile. Distance is the mean squared difference over shared finite observations, accumulated in input sample order. Exactly ten distinct other profiles with finite distances and positive overlap are selected; exact ties use original donor row order. Each missing coordinate uses the available original values among those ten neighbors (one to ten observed contributors per coordinate). No estimated value is reused, and there is no adaptive k, mean fallback or zero filling. NA and NaN are both missing, with separate internal counts. Observed and nontarget measurements and caller RNG state are preserved. When optional KNN runs, retained input infinities are unsupported; disabled/complete-output bypasses add no infinity filter. Nonfinite calculated distances are diagnosed and excluded individually.

Unsupported or partly unresolved represented targets stop the call with a structured error identifying target genes and reasons. Full estimator diagnostics remain available on the condition for inspection, without adding fields to successful Mapping outputs. Newly estimated values must also remain finite on their existing NC/SSP output scales. Integer diagnostic/work limits are checked before estimation; workload is never silently reduced.

This revised implementation differs from the original publication's pipeline. Incomplete-input estimates and downstream predictions may change, including calls that explicitly set impute = TRUE.

Raw counts and exact output scales. Raw NA/NaN counts are rejected with either impute setting. KNN is never used on raw counts or raw-derived outputs; impute = TRUE errors if missing raw-derived expression remains. Full-count normalization, library sizes and gene-length handling are unchanged: raw NC uses upper-quartile log2-CPM with the existing prior count, and raw SSP uses linear FPKM. Nonraw NC retains collapsed log expression; SSP uses 2^x, so log2(FPKM+1) input yields FPKM+1 without subtraction of one.

Clinical ROR covariates. With hasClinical = TRUE, NC method wrappers require both TSIZE and NODE columns and check that NODE is numeric. ROR-S and ROR-P scores and groups use

neither covariate. ROR-C and both ROR-PC scores use TSIZE, but no numerical ROR score uses NODE. Only the additional risk group ROR-PC Group (Subtype + Clinic + Prolif.Subtype) uses NODE categories 0, 1–3 and 4 or more. With nonmissing TSIZE, 4 or more positive nodes receive high risk; the other categories use different score cutoffs.

NODE is the number of positive lymph nodes. For this additional group, the NC methods accept only finite nonnegative integer-valued counts, stored as doubles or integers. Negative, fractional or infinite numeric values give one NODE warning per call and a missing additional group, while all numerical ROR scores and the other groups remain available. NODE is never rounded, truncated or recoded. Mapping() itself performs no NODE validation.

In the R API, an NA or NaN value in NODE makes only this additional group missing, without a NODE warning. An NA value in TSIZE makes all clinical ROR scores and groups missing; ROR-S and ROR-P remain available. These are missing values within required columns, not permission to omit the columns. The additional subtype-specific ROR-PC group uses inclusive upper boundaries: NODE = 0 has low <= 40, intermediate > 40 and <= 60, high > 60; NODE = 1–3 has low <= 15, intermediate > 15 and <= 40, high > 40. Scores are not rounded before grouping. This resolves issue #154; the original Parker risk groups are unchanged. For NODE = 0 or 1–3, an undefined numerical NA or NaN additional ROR-PC score has an actual missing category. The independent NODE >= 4 high-risk rule still applies with nonmissing TSIZE.

Phenodata normalization. If receptor fields are present, they are coerced to canonical encodings (ER -> {ER+, ER-}, HER2 -> {HER2+, HER2-}, TN -> {TN, nonTN}). Ambiguous values (e.g., HER2 == "2+") are left unchanged and emit a warning.

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).

Research ROR interpretation

The research ROR outputs use Spearman centroid correlations, following the research implementations (Parker et al., 2009, https://doi.org/10.1200/JCO.2008.18.1370; Buus et al., 2021, https://doi.org/10.1038/s41523-021-00216-w). The additional ROR-PC (Subtype + Clinic + Prolif.Subtype) score uses a PAM50 subset excluding BIRC5, CCNB1, GRB7 and MYBL2, an 18-gene proliferation component and binary tumor size. Commercial Prosigna uses Pearson correlations within its validated assay workflow (Wallden et al., 2015, https://doi.org/10.1186/s12920-015-0129-6). Shared coefficients do not establish equivalence of preprocessing, calibration or clinical validation. These scores and groups are research outputs, not clinical Prosigna results. No correlation calculation is changed in version 1.5.8.

Experimental 18-gene ROR-PC and **Experimental 18-gene ROR-PC group** identify the existing score and group columns, respectively:

```
ROR-PC (Subtype + Clinic + Prolif.Subtype)
ROR-PC Group (Subtype + Clinic + Prolif.Subtype)
```

Both legacy column names, their types and values are retained for compatibility. The original 11-gene Parker `ROR-PC (Subtype + Clinic + Prolif)` is a different output.

Experimental 18-gene ROR-PC: this research calculation differs from the published Lien implementation. Its prognostic performance and risk-category thresholds have not been independently validated. It is not a clinical Prosigna result and must not guide patient care. The published Lien implementation (2021, https://doi.org/10.3390/cancers13236118) standardizes each proliferation gene across patients; this package's additional score retains its calibrated, unstandardized 18-gene mean. Reproducing Lien's published results therefore does not validate this package variant. The categorical labels are experimental threshold assignments, not validated clinical risk strata.

Relevant public NC wrappers signal a warning of class `bsr_experimental_ror_pc_warning` once per top-level call when a nonempty returned `score.ROR` contains both legacy columns. `BS_Multi()` aggregates this notice once per call when a returned component contains them, rather than once per method, sample or iteration. There is no such warning on package loading, in `Mapping()`, with `hasClinical = FALSE`, or for AIMS/SSPBC-only results. An all-NA score column does not establish that its companion category is absent: the existing NODE >= 4 rule can still assign `high` with nonmissing tumor size.

This is an intentional warning: `options(warn = 2)` turns it into an error. An expert workflow can handle only the `bsr_experimental_ror_pc_warning` class:

- Establish a handler with `withCallingHandlers()`.
- Within that handler, call `invokeRestart("muffleWarning")`.

Unrelated warnings, including invalid NODE counts, should remain visible. Shiny identifies relevant selected Full metrics with the experimental labels and notice. Its TSV export preserves the existing tabular schema and provenance-comment layout while recording this interpretation.

Known original ROR boundary limitation

The original Parker-derived ROR-S, ROR-P, ROR-C and `ROR-PC (Subtype + Clinic + Prolif)` grouping retains strict comparisons. At exact equality to either unscaled cutoff, a group can contain the numeric score as text instead of `low`, `med` or `high`. The unscaled cutoff pairs are ROR-S: -0.15/0.10; ROR-P: -0.25/0.10; ROR-C: -0.10/0.20; and original ROR-PC: -0.20/0.20. These apply before `100 * (score + 0.35) / 0.85`, not directly to displayed scaled scores. Numeric-text values are invalid risk categories; do not interpret them as an additional category or infer an assignment by rounding. Version 1.5.8 temporarily defers this known categorical-output defect pending scientific review of the intended equality convention. Scores, coefficients and thresholds are preserved. This limitation is separate from the corrected NODE-dependent Experimental 18-gene ROR-PC group boundaries described in `Mapping()`, which remain unchanged.

Proliferation panel coverage

Two distinct proliferation calculations are used and are never interchanged. For the original Parker-derived ROR-P and `ROR-PC (Subtype + Clinic + Prolif)`, a separate copy of the calibrated expression matrix is standardized with column-wise `scale()`: each sample is centered at its mean across the represented PAM50 genes and divided by its sample standard deviation (denominator `n - 1` for `n` observed genes). The original 11-gene proliferation panel is then averaged on this standardized scale with `mean(..., na.rm = TRUE)`. The existing five-class centroid correlations, coefficients, cutoffs and final `100 * (score + 0.35) / 0.85` transformation are retained.

This PAM50-only standardization is a chosen, reproducible convention matching the complete 50-gene Perou Lab executable example distributed with the Parker bioclassifier (also bundled in PCA-PAM50 1.0.4). It is not a universal requirement inferred from the publication. The distributed script scales all rows present at its standardization step, before the centroid intersection; its behavior on arbitrary expanded gene sets is therefore not equivalent to this convention. BreastSubtypeR uses only the represented PAM50 genes in `outList$testData`, after its existing gene selection, duplicate resolution and method-specific calibration. No background genes are added. For `cIHC.itr`, standardization is applied to the final averaged matrix used for ROR scoring, not to each iteration. The classifier input retained in `outList$testData`, subtype predictions, correlations, Confidence and RNG behavior are unchanged.

Earlier implementations averaged the calibrated but unstandardized original proliferation values. Original ROR-P and ROR-PC scores and risk groups can consequently change substantially, even for complete panels. Historical saved ROR outputs are not automatically equivalent to newly computed outputs; comparisons require the implementation version, gene coverage, calibration and input provenance. Bundled examples and historical research results are not automatically recalculated. Reference agreement on complete PAM50 input does not establish clinical validation or validate every calibration-specific extension of the shared ROR formulas.

The additional research `ROR-PC (Subtype + Clinic + Prolif.Subtype)` is a separate definition: it continues to average the represented 18-gene panel on the calibrated, unstandardized expression scale, with the four-class correlations that exclude BIRC5, CCNB1, GRB7 and MYBL2, its own coefficients and the NODE-dependent groups described above. The panels share nine genes but are evaluated independently. This additional score and its groups, ROR-S and ROR-C are unchanged by the original proliferation correction.

Absent genes are never synthesized and no new panel coverage threshold is imposed. One represented proliferation gene can contribute when sample-wise PAM50 standardization is defined. Fewer than two observed finite PAM50 values, zero variance, or a nonfinite sample scale produce missing original proliferation-dependent scores and groups; no fallback score is substituted. Missing non-proliferation PAM50 genes also change the standardization mean and scale. Incomplete-panel scores are a compatibility limitation, not a validated partial-panel result; the packaged `TCGABRCAobj` example represents 7 of 11 and 8 of 18 proliferation genes. For either panel, no observed panel value or a nonfinite panel estimate gives a missing (`NA`) proliferation score. Every exported numerical ROR score is finite or `NA`: a `NaN` or infinite result, including overflow in the final score transformation, has a missing group rather than `low`, `med` or `high`. The independent NODE >= 4 high-risk rule for the additional score still applies with nonmissing tumor size. The deferred exact-boundary behavior of the original groups is unchanged.

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 OSLO2-EMIT0 log2(FPKM + 1) expression
    # from library-protocol-adjusted FPKM (see ?OSLO2EMIT0obj)
    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 expression values for 356 gene features from 102 OSLO2-EMIT0 samples, calculated as `log2(FPKM + 1)` from library-protocol-adjusted FPKM. Clinical metadata are stored in `colData`, and feature annotations in `rowData`.

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

`res` Example results from `BS_Multi()` run in *AUTO* mode.

Source

`OSLO2EMIT0.103.genematrix_noNeg.Rdata`, Mendeley Data, version 3, from the library-protocol-adjusted FPKM collection. doi:10.17632/yzxtxn4nmd.3 The source matrix contains 19,675 features and 103 sample identifiers. The packaged subset retains 356 features and 102 samples in source-relative order; source sample `OSLO2EMIT0.073` is absent. The historical reason for that omission, the feature-selection rationale and the exact generating script are not established. The expression transformation was verified by matching source feature and sample identifiers, not inferred from the filename.

References

Vallon-Christersson J (2023). RNA Sequencing-Based Single Sample Predictors of Molecular Subtype and Risk of Recurrence for Clinical Assessment of Early-Stage Breast Cancer. Mendeley Data, V3. doi:10.17632/yzxtxn4nmd.3

Staaf J, Hakkinen 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;8:94. 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, Hakkinen 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;8:94. 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, Hakkinen 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;8:94. 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 tumors. *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.

Details

Each plotted value is the maximum across one row of correlations. Rows are paired positionally with out$Subtype; keep the two inputs in the same sample order. Neutral boxes retain the usual quartiles, median, whiskers and outliers. The maximum correlation is not a probability.

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.

Details

Genes are scaled with t(scale(t(x[, out$PatientID]))). The blue, white and red expression scale is anchored at the observed minimum, zero and maximum gene z-score, respectively; the scale is not clipped or made symmetric. Subtype annotations retain the standard subtype colors. Missing subtype calls are grey and keep their expression measurements.

Gene and within-slice sample clustering use Spearman distance and average linkage. The existing dendrogram reordering and ComplexHeatmap clustering of column-slice means are retained. Both gene and sample dendrograms are displayed. Horizontal legends occupy a reserved footer, without changing the returned class. For 50 genes, a device around 10 by 7.4 inches is recommended; use RStudio's Zoom view for a small plot pane, PDF for vector output or PNG at 300 dpi.

Value

A ComplexHeatmap::Heatmap object representing gene expression across the subtype-specific sample slices.

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.

Details

Returns a `ComplexHeatmap::Heatmap` with separate subtype, method-family and (when needed) missing-call keys in a horizontal footer. The method rows, sample sorting rules, subtype colors and grey missing cells are preserved. The entropy annotation displays the supplied raw Shannon entropy in bits: disagreement among available method calls, not calibrated confidence. No entropy values are recomputed or normalized by this plot. A device around 10 by 4.6 inches leaves room for labels and legends.

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.

Details

The scatter plot uses a shared sans-serif theme and a bottom subtype legend. PCA centering, scaling, scores, joins, explained variance and reference lines are unchanged; no confidence ellipses or tests are added.

Value

A ggplot object representing the PCA plot, colored by subtype. If Eigen = TRUE, returns only the scree plot.

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.

Details

The legend reports subtype counts and percentages. As with table(), missing subtype calls are excluded from the assigned-call denominator; the caption reports both counts. Percentages are rounded to two decimals before drawing the slices, preserving the existing slice geometry and subtype order. Missing calls are not added as an extra slice.

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

∗ **datasets**
  AIMSmodel, 3
  BreastSubtypeRobj, 5
  Gene.ID.ann, 36
  OSLO2EMIT0obj, 45
  sspbc.models, 46
  sspbc.models.fullname, 47
  TCGABRCAobj, 47

AIMSmodel, 3

BreastSubtypeR, 4
BreastSubtypeR-package *(*BreastSubtypeR*)*, 4
BreastSubtypeRobj, 5, *31*
BS_AIMS, 6
BS_AIMS(), *37*
BS_cIHC, 8
BS_cIHC.itr, 11
BS_cIHC.itr(), *37*
BS_Multi, *4, 5*, 16, *51*
BS_Multi(), *9*, *14*, *19*, *24*, *28*, *32*, *37*, *43*
BS_parker, 21
BS_PCAPAM50, 26
BS_ssBC, 30
BS_ssBC(), *17, 18*
BS_sspbc, 34

Gene.ID.ann, 36

iBreastSubtypeR, 36
InputCoverage, *3*, 38
InputCoverage(), *4*, *7*, *16*, *23*, *35*

Mapping, *4, 5*, 39
Mapping(), *8–10*, *13–15*, *17*, *19*, *23, 24*, *26*, *28*, *31–33*, *38*, *43*

OSLO2EMIT0obj, 45

sspbc.models, 46
sspbc.models.fullname, 47

TCGABRCAobj, 47

Vis_boxplot, 48
Vis_heatmap, 49
Vis_Multi, *4, 5*, 50
Vis_PCA, 51
Vis_pie, 52