Package 'tidyestimate'

July 22, 2025

Type Package

Title A Tidy Implementation of 'ESTIMATE'

Version 1.1.1

Description The 'ESTIMATE' package infers tumor purity from expression data as a function of immune and stromal infiltrate, but requires writing of intermediate files, is un-pipeable, and performs poorly when presented with modern datasets with current gene symbols. 'tidyestimate' a fast, tidy, modern reimagination of 'ESTIMATE' (2013) <doi:10.1038/ncomms3612>.

License GPL (>= 2)

URL https://github.com/KaiAragaki/tidyestimate

BugReports https://github.com/KaiAragaki/tidyestimate/issues

Depends R (>= 4.1.0)

Imports glue, dplyr, stats, rlang, ggrepel, ggplot2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests rmarkdown, knitr

VignetteBuilder knitr

NeedsCompilation no

Author Kai Aragaki [aut, cre] (ORCID: <https://orcid.org/0000-0002-9458-0426>),
Paul Roebuck [cph] (Copyright holder of ESTIMATE package),
Kosuke Yoshihara [aut] (Author of original ESTIMATE algorithm),
Rahulsimham Vegesna [aut] (Author of original ESTIMATE algorithm),
Hoon Kim [aut] (Author of original ESTIMATE algorithm),
Roel Verhaak [aut] (Author of original ESTIMATE algorithm)

Maintainer Kai Aragaki <aaragak1@jhmi.edu>

Repository CRAN

Date/Publication 2023-08-21 03:50:02 UTC

Contents

`common_genes` *Genes shared between six expression platforms*

Description

As the ESTIMATE model was trained on a specific set of genes, only those within this dataset should be included before running `estimate_scores`.

These are the genes common to 6 platforms:

- Affymetrix HG-U133Plus2.0

- Affymetrix HT-HG-U133A

- Affymetrix Human X3P

- Agilent 4x44K (G4112F)

- Agilent G4502A

- Illumina HiSeq RNA sequence

The Entrez IDs for the original 10412 genes were matched to HGNC symbols using `biomaRt`. Duplicates and blank entries were filtered. As some have now been discovered to be pseudogenes or have been deprecated, 22 genes (at time of writing, June 2021) that were in the ESTIMATE package do not exist here.

As one gene can have multiple synonyms/aliases, and there is only one alias per line, the number of rows in the data frame (26339) does not reflect the number of unique genes in the dataset (10391).

Usage

```
common_genes
```

Format

A data frame with 26339 rows and 3 variables:

entrezgene_id Entrez id of the gene

hgnc_symbol Human Genome Organisation (HUGO) Gene Nomenclature Committee symbol

external_synonym A synonym/alias a given gene may go by or previously went by

Details

The ESTIMATE model was trained on a set of genes shared between six expression profiling platforms. Those genes are listed in this dataset.

Source

https://r-forge.r-project.org/scm/viewvc.php/pkg/estimate/data/common_genes.RData?root=estimate&view=log

---

`estimate_score` *Infer tumor purity using the ESTIMATE algorithm*

Description

Infer tumor purity by using single-sample gene-set-enrichment-analysis with stromal and immune cell signatures.

Usage

```
estimate_score(df, is_affymetrix)
```

Arguments

`df`	a `data.frame` of expression data, where columns are tumors and rows are genes. Gene names must be in the first column, and in the form of HGNC symbols.
`is_affymetrix`	logical. Is the expression data from an Affymetrix array?

Details

ESTIMATE (and this tidy implementation) infers tumor infiltration using two gene sets: a stromal signature, and an immune signature (see `tidyestimate::gene_sets`).

Enrichment scores for each sample are calculated using an implementation of single sample Gene Set Enrichment Analysis (ssGSEA). Briefly, expression is ranked on a per-sample basis, and the density and distribution of gene signature 'hits' is determined. An enrichment of hits at the top of the expression ranking confers a positive score, while an enrichment of hits at the bottom of the expression ranking confers a negative score.

An 'ESTIMATE' score is calculated by adding the stromal and immune scores together.

For Affymetrix arrays, an equation to convert an ESTIMATE score to a prediction of tumor purity has been developed by Yoshihara et al. (see references). It takes the approximate form of:

$$purity = cos(0.61 + 0.00015 * ESTIMATE)$$

Values have been rounded to two significant figures for display purposes.

Value

A `data.frame` with sample names, as well as scores for stromal, immune, and ESTIMATE scores per tumor. If `is_affymetrix = TRUE`, purity scores as well.

Purity scores can be interpreted absolutely: a purity of 0.9 means that tumor is likely 90 available (such as in RNAseq), ESTIMATE scores can only be interpreted relatively: a sample that has a lower ESTIMATE score than another in one study can be regarded as more pure than another, but its absolute purity cannot be inferred, nor can purity across other studies be inferred.

References

Barbie et al. (2009) <doi:10.1038/nature08460>

Yoshihara et al. (2013) <doi:10.1038/ncomms3612>

Examples

```
filter_common_genes(ov, id = "hgnc_symbol", tidy = FALSE, tell_missing = TRUE, find_alias = TRUE) |>
  estimate_score(is_affymetrix = TRUE)
```

`filter_common_genes` *Remove non-common genes from data frame*

Description

As ESTIMATE score calculation is sensitive to the number of genes used, a set of common genes used between six platforms has been established (see `?tidyestimate::common_genes`). This function will filter for only those genes.

Usage

```
filter_common_genes(
  df,
  id = c("entrezgene_id", "hgnc_symbol"),
  tidy = FALSE,
  tell_missing = TRUE,
  find_alias = FALSE
)
```

Arguments

`df`	a `data.frame` of RNA expression values, with columns corresponding to samples, and rows corresponding to genes. Either rownames or the first column can contain gene IDs (see `tidy`)
`id`	either `"entrezgene_id"` or `"hgnc_symbol"`, whichever `df` contains.
`tidy`	logical. If rownames contain gene identifier, set `FALSE`. If first column contains gene identifier, set `TRUE`

`tell_missing`	logical. If `TRUE`, prints message of genes in common gene set that are not in supplied data frame.
`find_alias`	logical. If `TRUE` and `id = "hgnc_symbol"`, will attempt to find if genes missing from `common_genes` are going under an alias. See details for more information.

Details

The `find_aliases` argument will attempt to find aliases for HGNC symbols in `tidyestimate::common_genes` but missing from the provided dataset. This will only run if `find_aliases = TRUE` and `id = "hgnc_symbol"`.

This algorithm is very conservative: It will only make a match if the gene from the common genes has only one alias that matches with only one gene from the provided dataset, *and* the gene from the provided dataset with which it matches only matches with a single gene from the list of common genes. (Note that a single gene may have many aliases). Once a match has been made, the gene in the provided dataset is updated to the gene name in the common gene list.

While this method is fairly accurate, is is also a heuristic. Therefore, it is disabled by default. Users should check which genes are becoming reassigned to ensure accuracy.

The method of generation of these aliases can be found at `?tidyestimate::common_genes`

Value

A `tibble`, with gene identifiers as the first column

Examples

```
filter_common_genes(ov, id = "hgnc_symbol", tidy = FALSE, tell_missing = TRUE, find_alias = FALSE)
```

---

gene_sets *Gene sets to infer tumor stromal and immune infiltration*

Description

Two gene sets, each 141 genes in length, created to infer stromal and immune infiltration

Usage

```
gene_sets
```

Format

A data frame with 141 row and 2 variables:

stromal_signature Geneset of HGNC symbols used to infer tumor stromal cell infiltration

immune_signature Geneset of HGNC symbols used to infer tumor immune cell infiltration

Source

https://r-forge.r-project.org/scm/viewvc.php/pkg/estimate/data/SI_geneset.RData?root=estimate&view=log

ov *Ovarian cancer tumor RNA expression*

Description

A matrix containing RNA expression of 10 ovarian cancer tumors, measured using the Affymetrix U133Plus2.0 platform. These data have been rounded to the 4th decimal place to reduce file size.

Usage

```
ov
```

Format

A matrix with 17256 rows and 10 columns, where each column represents a tumor, and each row represents a gene. Genes are represented by HGNC symbols in the rownames.

Source

https://r-forge.r-project.org/scm/viewvc.php/pkg/estimate/inst/extdata/sample_input.txt?root=estimate&view=log

plot_purity *Plot Affymetrix purity scores against ESTIMATE study purity scores*

Description

Plot Affymetrix purity scores against ESTIMATE study purity scores

Usage

```
plot_purity(scores, is_affymetrix)
```

Arguments

scores	a data.frame, usually one output from estimate_score
is_affymetrix	logical. Are these data from an Affymetrix experiment? Must be TRUE - this is essentially a verification from the user

Value

a ggplot

Examples

```
filter_common_genes(ov, id = "hgnc_symbol", tidy = FALSE, tell_missing = TRUE, find_alias = TRUE) |>
  estimate_score(is_affymetrix = TRUE) |>
  plot_purity(is_affymetrix = TRUE)
```

---

purity_data_affy *Affymetrix data used to train ESTIMATE algorithm*

Description

A data frame containing the ABSOLUTE-measured and ESTIMATE-predicted purity values of 995 tumors. Additionally, stromal and immune scores as calculated by ESTIMATE. All tumors were profiled on Affymetrix arrays, and were used to generate the Affymetrix algorithm.

Usage

```
purity_data_affy
```

Format

A data frame with 995 rows and 7 variables:

purity_observed The purity of a tumor given by ABSOLUTE, ranging from 0 (least pure) to 1 (most pure)

stromal Stromal infiltration score, as measured by ESTIMATE

immune Immune infiltration score, as measured by ESTIMATE

estimate ESTIMATE score, calculated by the sum of immune and stromal scores

purity_predicted Tumor purity inferred using the ESTIMATE algorithm

ci_95_low Lower bound of a 95% confidence interval of predicted purity scores

ci_95_high Upper bound of a 95% confidence interval of predicted purity scores

Source

https://r-forge.r-project.org/scm/viewvc.php/pkg/estimate/data/PurityDataAffy.RData?root=estimate&view=log

`tidyestimate` *tidyestimate: A modern implementation of the ESTIMATE algorithm*

Description

The tidyestimate is a lightweight, fast, pipe-friendly re-imagination of the ESTIMATE package. tidyestimate is used to infer tumor purity from expression data.

Authors

Author (tidyestimate):

* Kai Aragaki ([ORCID](http://orcid.org/0000-0002-9458-0426)) (author, maintainer)

Authors (ESTIMATE):

* Kosuke Yoshihara kyoshihara@mdanderson.org (author) * P. Roebuck proebuck@mdanderson.org (author, copyright holder)

Reference

https://www.nature.com/articles/ncomms3612

Index