Package 'cellOrigins'

July 22, 2025

Type Package

Title Finds RNASeq Source Tissues Using In Situ Hybridisation Data

Version 0.1.3

Author David Molnar

Maintainer David Molnar <dmolnar100@icloud.com>

Description

Finds the most likely originating tissue(s) and developmental stage(s) of tissue-specific RNA sequencing data. The package identifies both pure transcriptomes and mixtures of transcriptomes. The most likely identity is found through comparisons of the sequencing data with high-throughput in situ hybridisation patterns. Typical uses are the identification of cancer cell origins, validation of cell culture strain identities, validation of single-cell transcriptomes, and validation of identity and purity of flow-sorting and dissection sequencing products.

License CC BY-NC-SA 4.0

Encoding UTF-8

LazyData true

Imports iterpc

NeedsCompilation no

Repository CRAN

Date/Publication 2020-06-05 09:00:02 UTC

Contents

`cellOrigins-package` *Finding the most likely originating tissue(s) and developmental stage(s) of RNASeq data*

Description

`cellOrigins` compares RNASeq read coverages with in high-throughput RNA in situ hybridisation patterns for transcriptome source identification and verification. The package can identify both pure transcriptomes and mixtures of transcriptomes. Typical uses are the identification of cancer cell origins, validation of cell culture strain identities, validation of single-cell transcriptomes, and validation of identity and purity of flow-sorting and dissection sequencing products.

The comparison of quantitative RNA sequencing coverage with thresholded, qualitative staining patterns is probabilistic. First, given the sequenced transcriptome, a prediction is made how likely each sequenced transcript would lead to a positive signal in a high-throughput in situ hybridisation experiment. The probability of staining increases with the logarithm of the sequencing coverage. This relationship was empirically found through a comparison between *Drosophila* embryo transcriptomes and RNA in situ staining results. Then, using Bayes's theorem all the genes in the simulated and observed hybridisation patterns are compared. The pattern (or linear combination of patterns) with the highest posterior probability is identified as the most likely source.

Batteries included: the package contains a filtered high-confidence expression pattern dataset for *Drosophila melanogaster* embryos (based on BDGP insitu).

Typical use:

I GENERATE INPUT

Input is RNASeq mean FPKM (fragments per kilobase per million reads). Whole-gene FPKM may be used (as output by e.g. cufflinks/cuffquant), however assignment difficulties at overlapping transcripts and transcript isoforms reduce prediction quality. For best results use FPKM values calculated for the targets of the in situ hybridisation probes as described below:

Step 1) Generate masking bed file – this file is included for BDGP insitu in the extdata folder. For other species align probe sequences to the target genome using BLAT (https://genome.ucsc.edu/FAQ/FAQblat.html). Convert the best-scoring alignments to a masking bed file with `psl_to_bed_best_score.pl` (https://gist.github.com/davetan Then sort with `bedtools sort` (http://bedtools.readthedocs.org/).

Step 2) Get coverages. Use Bedtools with the masking bed file to extract the mean sequencing covereage from wig files in the in situ probed regions:

```
bedtools map -a sorted_probes.bed -b sequenced.wig -o max -c 4 >insitu_high_confidence.tsv
```

Use the output tab separated values file as input for the function `seqVsInsitu`.

II SOURCE IDENTIFICATION

`seqVsInsitu` and `iterating_seqVsInsitu` calculate the probability for each in situ expression pattern that it is produced by the same gene expression patterns as the sequencing data. If you believe you have a mixed input, allow combined patterns from several target tissues. This is computationally expensive for more than two tissues. `iterating_seqVsInsitu` is faster thorugh calculating all combinations for n==2 and then using only the top tissues for n==3. The top tissues of n==3 is then are used for n==4 etc.

III INTERPRETATION

`seqVsInsitu` and `iterating_seqVsInsitu` return the terms or term combinations together with a log2 probability score for each. They also produce two diagnostic graphs. If multiple tissues contribute to the sample, the scatterplot should show a number of clusters at low n. As n increases, the clusters should merge into just two clusters at the ideal value of n. The line graph shows the log2 probability distribution.

`discovery_probability` if RNASeq and in situ hybridisation data from the same tissue are paired, then with increasing FPKM the probability of RNA in situ discovery should increase logarithmically. If the tissue sources do not match, no such relationship should be visible. Using this function, if the tissue combination in the argument is a match, there should by a nearly linearly increasing relationship in the log-plot, with saturation at very high FPKM values only.

Details

Package:	cellOrigins
Type:	Package
Version:	1.0
Date:	2015-03-18
License:	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Author(s)

David Molnar

Maintainer: David Molnar <dmolnar100@icloud.com>

References

Molnar, D 2015, 'Single embryo-single organ transcriptomics of Drosophila embryos', PhD thesis, University of Cambridge.
BDGP insitu: Tomancak, Genome Biol. 2007;8(7):R145.
BDGP insitu homepage: insitu.fruitfly.org/cgi-bin/ex/insitu.pl

Examples

```
## Not run:
pmoracle <- seqVsInsitu(transcriptomeMatrix)
rownames(pmoracle)[1:3]
diagnosticPlots(pmoracle)

## End(Not run)
##loading the BDGP insitu probe coordinates if not
##copied directly from the package extdata folder
system.file("extdata", "BDGP_insitu_probes.bed", package = "cellOrigins")
```

BDGP_insitu_dmel_embryo

Patterns of gene expression in Drosophila melanogaster embryos

Description

High-confidence dataset of embryonic *Drosophila melanogaster* RNA expression patterns at 6 developmental stages. This dataset was generated by filtering the "BDGP insitu" high-throughput RNA in situ hybridisation data set (Tomancak, Genome Biol. 2007;8(7):R145) for high-confidence results. Only genes useful for tissue identification were retained, and they thus represent gene expression fingerprints of organs.

Usage

```
data("BDGP_insitu_dmel_embryo")
```

Format

The format is: num [1:2395, 1:337] 1 0 0 0 1 1 0 1 1 1 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:2395] "LD11379" "LD11394" "LD12611" "LD12613" ... ..$: chr [1:337] "1|maternal" "2|pole cell" "3|pole cell" "4|germ cell" ...

Details

The expression data are collated in a matrix. The columns in the matrix are labeled stage|domain (e.g. "6|midgut"). The expression domains are denoted using the BDGP insitu controlled anatomical vocabulary.The rows are labeled with transcripts/probe names according to the BDGP insitu data set. The hybridisation probe genomic coordinates (Drosophila melanogaster genome release 5) are supplied as an additional file in this package.

The data set characterises the expression of 2395 RNA species. This is the differentially expressed, high-confidence subset of BDGP insitu. The starting point for dataset preparation was the published SQL database dump with annotations (http://insitu.fruitfly.org/insitu-mysql-dump/insitu.sql.gz). All in situ hybridisations for wild type *Drosophila melanogaster* embryos were extracted from this source. The reporter construct annotations were not used.

Only high-confidence expression patterns were retained. The gene expression in the BDGP insitu database was annotated by human curators from microscopic images. Depending on the quality of images and staining some expression patterns were easier to discern than others. The curators expressed their confidence in their expression call together with the annotation data of each gene. The filtering criteria for including a probe's exression pattern were that

1. the final call of the annotators was 'acceptable',
2. there was no remark about staining intensity (pointing to substandard quality),
3. the microscopic image was not excluded by quality control,
4. the annotation was displayed on the database's website,
5. the probe/staining was not flagged for repeating or for giving up, and

6. the final word of the annotators (a free text field) did not contain negative remarks like "weak", "nonspecific", "muddy", "poor", "dull", "spillover" or "suspicious" staining; lack of staining penetration; a call to repeat the staining; signs of doubt (e.g. "might", "perhaps", "maybe", "could", "not sure", "not confirmed", "unconvincing", "conflicting", "can't say", "failure", "wrong", "junk"); on camera problems; artefacts or transposons.
7. there was no annotation with "no staining" to avoid false negatives.

Genes with known ubiquitous expression (including faint-ubiquitous) at any stage were excluded.

Genes for which there was no published probe sequence (approximately 300) were excluded. Most of the RNA in situ hybridisation probes originated from the Drosophila Gold Collection (http://www.fruitfly.org/EST/gold_co and the Drosophila Gene Collection (http://www.fruitfly.org/DGC/index.html).

Annotated gene expression in each anatomical unit was propagated to all its anatomical subunits. For example "5|Malpighian tubule primordium" expression was propagated to "5|Malpighian tubule main body primordium" and "5|Malpighian tubule tip cell primordium". Only this made both the presence and the absence of staining meaningful. In the original data set gene expression was usually only annotated to the largest unit of expression, but not to its subunits. For instance if there was expression in the whole foregut, there was by necessity also expression in its pharynx subunit. However, in such a case expression in the pharynx was not commonly denoted in the original data set. Consequently some anatomic units had very few expressed genes associated. These genes were those that were exclusively expressed in those anatomical units and in no superior units.

Source

Tomancak, Genome Biol. 2007;8(7):R145

Examples

```
data(BDGP_insitu_dmel_embryo)
```

`diagnosticPlots` *Diagnostic plots to explore seqVsInsitu results*

Description

Accepts the result of `seqVsInsitu` and `iterating_seqVsInsitu` and produces diagnostic plots. If the sequencing data fits to one or more terms or combinations of terms, then the scatterplot will cluster into foci. As the number of combined terms is increased the foci merge into fewer groups. A diagonal in the scatterplot is a sign of error.

Usage

```
diagnosticPlots(seqVsInsitu_results)
```

Arguments

`seqVsInsitu_results`
: Value of `seqVsInsitu` or `iterating_seqVsInsitu`.

Value

None.

Examples

```
fpath <- system.file("extdata", "vncMedianCoverage.tsv", package="cellOrigins")
vncExpression <- read.delim(file = fpath, header=FALSE, as.is=TRUE)

expression <- vncExpression$V2
names(expression) <- vncExpression$V1

result <- seqVsInsitu(expression, depth=1)
diagnosticPlots(result)

## Not run:
oracleResponse <- iterating_seqVsInsitu(expression, 3)
diagnosticPlots(oracleResponse)

## End(Not run)
```

`discovery.log` *Calculates discovery probability by RNA in situ hybridisation given a sequencing signal*

Description

A set of functions with different assumptions on the probability of RNA in situ staining, given a sequencing coverage.

Usage

```
discovery.log(seq, saturate = 60, bias = 0.01)
discovery.linear(seq, saturate = 60, bias = 0.01)
discovery.identic(seq, saturate=Inf, bias=0)
```

Arguments

`seq`	A vector of sequencing FPKMs.
`saturate`	FPKM value from which on maximum discovery probability (=0.99) is assumed (i.e. almost certain true positives). Value of 60 is default, may need adjustment to sequencing coverage.
`bias`	Positive staining probability of 0 FPKM transcripts (i.e. false positives). Must be >0. Default is 0.01, an empirically determined value.

Details

1. **discovery.log** Uses a logarithmic saturation function for discovery probabilities. This relationship was empirically determined from sequencing and hybridisation data.
2. **discovery.linear** Linear saturation function for discovery probabilities.
3. **discovery.identic** Passes input through. Useful for comparing RNASeq Vs. RNASeq data. Also for cases when the discovery probability for each transcript has been already determined in some other way.

Value

A vector of probabilities. Element names are preserved.

See Also

seqVsInsitu

Examples

```
plot(0:80, discovery.log(0:80),
  ylim=c(0,1.1), type="l",
  xlab="FPKM", ylab="p(discovery insitu hybridization)")

plot(0:80, discovery.linear(0:80),
  ylim=c(0,1.1), type="l",
  xlab="FPKM", ylab="p(discovery insitu hybridization)")
```

---

discovery_probability *In situ discovery probability as a function of FPKM*

Description

Groups transcripts by expression strength and calculates for each such group the percentage of genes that gave a positive staining signal in the in situ hybridisation.

If the sequenced material matches the in situ hybridisation tissue, then weakly expressed genes in the sequenced material should be rearely in the in situ staining set of genes. Strongly expressed genes should correspondingly often also stain during hybridisation. Overall, if the match is not spurious, there should be a logarithmic dose-response relationship between sequencing read coverage and staining probability. In a plot of discovery probability against log(coverage) this shows as an approximately straight line (see example).

Usage

```
discovery_probability(seq_signature, terms, cut.points,
    insitu=cellOrigins::BDGP_insitu_dmel_embryo)
```

Arguments

seq_signature	A named vector containing FPKM RNAseq data. Each element name must correspond to the names used in the insitu argument. NAs are permitted.
terms	A vector of anatomical terms which together are assumed to be the origin of the RNAseq data.
cut.points	A vector of cut points for grouping of values. E.g. 0:3 denotes the bins 0<=x<1, 1<=x<2, 2<=x<3.
insitu	Matrix with in situ hybridisation data. Rows are transcript names (same names as used for seq_signature) and coloumns are anatomical terms (possibly combined with developmental stages). 1 denotes staining of a particular transcript in a particular tissue, 0 denotes no staining. Defaults to BDGP_insitu_dmel_embryo, a staining dataset for *Drosophila melanogaster* embryos.

Value

A matrix with a row for each bin and three coloumns. The first coloumn is the probability of discovery, the second the number of transcripts in the expression bin that were discovered by in situ hybridisation. The third coloumn is the total number of transcripts in the bin.

See Also

iterating_seqVsInsitu, BDGP_insitu_dmel_embryo, discovery.log, discovery.linear, discovery.identic, prior.temporal_proximity_is_good, prior.all_equal, diagnosticPlots.

Examples

```
fpath <- system.file("extdata", "vncMedianCoverage.tsv", package="cellOrigins")
vncExpression <- read.delim(file = fpath, header=FALSE, as.is=TRUE)

expression <- vncExpression$V2
names(expression) <- vncExpression$V1

p <- discovery_probability(expression,
  "6|ventral nerve cord", c(0, 2^(0:10)))

plot(x=-1:9, y=p[,1], type="l",
  xlab="log2(FPKM)", ylab="p(discovery in situ)")
```

---

iterating_seqVsInsitu *Faster comparisons between mixed tissue-specific RNA sequencing data and high-throughput RNA in situ hybridisation*

Description

The same functionality as seqVsInsitu but computationally less expensive if combinations of anatomical terms are tested.

The number of term combinations to test increases rapidly in seqVsInsitu. For example with 350 anatomical terms there are 61425 combinations of 2 terms and 7207200 combinations of 3 terms. This makes the exhaustive search of seqVsInsitu costly with depth>2.

iterating_seqVsInsitu reduces the computational cost by initially testing the combinations of only a few terms. Then in each iteration the cardinality of the combinations is increased by one, but only the top anatomical terms of the previous iteration are used to reduce the number of tested combinations.

Usage

```
iterating_seqVsInsitu(seq_signature, upto_depth, use_topN = 50,
  start_depth = 2, insitu = cellOrigins::BDGP_insitu_dmel_embryo,
  insitu_discovery_function = discovery.log, saturate = 500,
  prior = prior.temporal_proximity_is_good)
```

Arguments

seq_signature	A named vector containing FPKM RNAseq data. Each element name must correspond to the names used in the insitu argument. NAs are permitted.
upto_depth	Number of terms to combine in the final iteration.
use_topN	How many of the top results from the previous iteration to use to find the terms for the current iteration.
start_depth	Number of terms to combine in the first iteration. All combinations of all terms are tested at this step.
insitu	Matrix with RNA in situ hybridisation data. Rows are transcript names (queried by probes: same names as used for seq_signature) and coloumns are anatomical terms (possibly combined with developmental stages). If a probe stains in a particular tissue, the value is 1, otherwise 0. Defaults to BDGP_insitu_dmel_embryo, a staining dataset for fruit fly embryos.
insitu_discovery_function	A function that converts FPKM values to the probability of discovery by RNA in situ hybridisation. Values must be]0..1[, 0 and 1 are not permitted. Defaults to discovery.log, an approximation of empirically determined discovery probabilities. Other available functions are discovery.linear and discovery.identic.
saturate	Will be passed on to the insitu_discovery_function. The data set dependent maximum value at which the discovery probability should saturate. Defaults to 500 (FPKM).
prior	A function that evaluates to the log2 prior probability of each anatomic term or combination of terms. Defaults to prior.temporal_proximity_is_good, which works well with BDGP_insitu_dmel_embryo. prior.all_equal assumes equal probability of all terms.

Value

Returns a named list that contains a matrix for each iteration like those produced by seqVsInsitu.

See Also

seqVsInsitu

Examples

```
## Not run:
fpath <- system.file("extdata", "vncMedianCoverage.tsv", package="cellOrigins")
vncExpression <- read.delim(file = fpath, header=FALSE, as.is=TRUE)

expression <- vncExpression$V2
names(expression) <- vncExpression$V1

oracleResponse <- iterating_seqVsInsitu(expression, 3)
head(oracleResponse[[1]])
head(oracleResponse[[2]])
diagnosticPlots(oracleResponse)

## End(Not run)
```

---

prior.temporal_proximity_is_good

Assign a prior probability to a combination of anatomical terms

---

Description

Accepts one or more anatomical terms and assigns to them a prior probability in the Bayesian sense. `prior.all_equal` assumes all terms and combinations to be equally probable. `prior.temporal_proximity_is_good` is meant mainly for use with BDGP_insitu_dmel_embryo if working with single or staged embryos. With this function the prior probability increases if the developmental stages in the tested terms are close together. The magnitude of the prior is scaled to the number of tested genes.

Usage

```
prior.temporal_proximity_is_good(term_pairs, insitu_signature)
```

Arguments

`term_pairs`	A vector with anatomical terms that are tested in combination.
`insitu_signature`	The RNA in situ hybridisation data set as produced by fusion of the expression patterns in `term_pairs`, and as it will be used for calculating the posterior probability in seqVsInsitu.

`seqVsInsitu` *Determine the most likely source(s) of a tissue-specific RNAseq dataset*

Description

Compares tissue-specific RNA sequencing coverage with high-throughput RNA in situ hybridisation patterns of gene expression. All pattern combinations are tested in an exhaustive search.

Usage

```
seqVsInsitu(seq_signature, depth = 2, insitu = cellOrigins::BDGP_insitu_dmel_embryo,
  insitu_discovery_function = discovery.log, saturate = 500,
  prior = prior.temporal_proximity_is_good)
```

Arguments

`seq_signature`	A named vector containing FPKM RNAseq data. Each element name must correspond to the names used in the `insitu` argument. NAs are permitted.
`depth`	Number of RNA in situ expression patterns to combine to identify mixed populations. If 1, the expression patterns as given are used. Otherwise all combinations of `depth` expression patterns are tried. Each term combined with itself is also tested i.e. pure populations will still be identified if depth>1. Defaults to 2. `seqVsInsitu` Depths > 2 can be slow. `iterating_seqVsInsitu` is much faster in these cases.
`insitu`	Matrix with RNA in situ hybridisation results. Rows are transcript names (same names as used for `seq_signature`) and coloumns are anatomical terms (possibly combined with developmental stages). 1 denotes staining of a particular transcript in a particular tissue, 0 denotes no staining. Defaults to `BDGP_insitu_dmel_embryo`, a staining dataset for *Drosophila melanogaster* embryos.
`insitu_discovery_function`	A function that converts FPKM values to the probability of discovery by RNA in situ hybridisation. Probabilities must be]0..1[, the values 0 and 1 are not permitted. Defaults to `discovery.log`, an approximation of empirically determined discovery probabilities. Other available functions are `discovery.linear` and `discovery.identic`.
`saturate`	Will be passed on to the `insitu_discovery_function`. The data set dependent maximum value at which discovery probability should saturate. Defaults to 500 (FPKM).
`prior`	A function that returns the log2 prior probability of each anatomic term or combination of terms. Defaults to `prior.temporal_proximity_is_good`, which works well with `BDGP_insitu_dmel_embryo`. `prior.all_equal` assumes that all terms are equally probable.

Details

First, the function calculates for each sequenced transcript how likely it is that it would produce an RNA in situ signal, given its expresion strength. Using these staining probabilities and Bayes's rule the function then calculates the probability score for each of the given RNA in situ hybridisation patterns that it was produced by the same gene expression pattern as the sequenced transcriptome.

If `depth>1` then the function identifies the origins of not pure sequenced material. For that it merges multiple RNA in situ hybridisation patterns for comparison with the sequenced data. This simulates the outcome of cell populations mixing.

`seq_signature` is best generated by taking the mean coverage of the regions which are actually tested with the RNA in situ hybridisation probes. This circumvents problems from misannotation, overlapping transcripts and faulty quantitation of individual transcripts from sequencing data. A protocol for generating such datasets is given in the package reference.

Value

A matrix with a row for each anatomical term (or combination of terms) and at least four columns. The terms are sorted by the posterior value and the top term is the most likely source of the RNAseq transcriptome.

`posterior` A log2 posterior probability score. The highest value is given to the most likely tissue of origin. The value is only meaningful in comparison with other values within the same result set.

`prior` Prior probability of the anatomical term(s), as given by the function `prior`.

`likelihood.from.absence.insitu`
Probability score from all the genes where RNA in situ hybridisation did not report staining.

`likelihood.from.presence.insitu`
Probability score from all the genes where in situ hybridisation reported staining.

`remaining coloumns`
Number of additional expressed genes added to the in situ signature with each term in the tested combination. Sometimes additional terms add only very few or no new genes at all. Such tissue contributions are meaningless artefacts.

The posterior column is the sum of the other three named columns. The scores are proportional to the (unknown) probabilities of identity.

See Also

`iterating_seqVsInsitu`, `BDGP_insitu_dmel_embryo`, `discovery.log`, `discovery.linear`, `discovery.identic`, `prior.temporal_proximity_is_good`, `prior.all_equal`, `diagnosticPlots`.

Examples

```
fpath <- system.file("extdata", "vncMedianCoverage.tsv", package="cellOrigins")
vncExpression <- read.delim(file = fpath, header=FALSE, as.is=TRUE)

expression <- vncExpression$V2
names(expression) <- vncExpression$V1
```

```
result <- seqVsInsitu(expression, depth=1)
```

---

vncMedianCoverage.tsv *Drosophila melanogaster embryo ventral nerve cord RNASeq coverage*

Description

Median RNAseq read coverages from 3 dissected embryonic (stage 11) fruit fly ventral nerve cords. The sequencing coverages are measured within the probing intervals of high-confidence BDGP insitu probes, as described in cellOrigins-package.

Format

The format is: probe name, coverage, chromosome, probe beginn, probe end, strand.

Source

Molnar, D 2015, 'Single embryo-single organ transcriptomics of Drosophila embryos', PhD thesis, University of Cambridge.

Examples

```
  fpath <- system.file("extdata", "vncMedianCoverage.tsv", package="cellOrigins")
vncExpression <- read.delim(file = fpath, header=FALSE, as.is=TRUE)
```

Index