Package 'EDec'

October 12, 2016

Title Cell type specific analysis of complex tissues through Epigenomic Deconvolution

Version 0.9

Description EDec (Epigenomic Deconvolution) is a technique that, starting from methylation and gene expression profiles of bulk tissue samples, infers cell type composition of each input sample as well as DNA methylation and gene transcription profiles of constituent cell types.

Depends R (>= 3.1.2)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports quadprog, gtools, stats, clue, utils

RoxygenNote 5.0.1.9000

Suggests EDecExampleData, RColorBrewer, knitr, rmarkdown, gplots

VignetteBuilder knitr

NeedsCompilation no

Author Vitor Onuchic [aut, cre]

Maintainer Vitor Onuchic <vitor.onuchic@bcm.edu>

R topics documented:

`estimate_meth_qp` *Estimate cell type specific methylation*

Description

Given methylation profiles of complex tissue samples and propotions of contituent cell types in each sample, `estimate_meth_qp` will estimate average methylation profiles of constituent cell types across the set of input samples.

Usage

```
estimate_meth_qp(meth_bulk_samples, cell_type_props)
```

Arguments

`meth_bulk_samples`
: Matrix of methylation profiles of bulk complex tissue samples. Columns correspond to different samples and rows correspond to different loci/probes.

`cell_type_props`
: Matrix of proportions of constituent cell types. Columns correspond to different cell types and rows correspond to different bulk tissue samples.

Details

EDec assumes that the methylation profiles of complex tissue samples correspond to the linear combination of cell type proportions and methylation profiles of each cell type. Given the methylation profiles of a set of complex tissue samples and the proportions of constituent cell types in each sample, this function estimates average methylation profiles of constituent cell types by solving constrained least squares problems through quadratic programming. The constraint is that the methylation profiles of constituent cell types are numbers in the [0,1] interval.

Value

Matrix with estimated average methylation profiles of constituent cell types.

`estimate_props_qp` *Estimate cell type proportions*

Description

This function will estimate the proportions of constituent cell types in each input sample, given methylation profiles of complex tissue samples and methylation profiles of constituent cell types.

Usage

```
estimate_props_qp(meth_bulk_samples, cell_type_specific_meth)
```

Arguments

`meth_bulk_samples`
: Matrix of methylation profiles of bulk complex tissue samples. Columns correspond to different samples and rows correspond to different loci/probes.

`cell_type_specific_meth`
: Matrix of methylation profiles of constituent cell types. Columns correspond to different cell types and rows correspond to different loci/probes.

Details

EDec assumes that the methylation profiles of complex tissue samples correspond to the linear combination of cell type proportions and methylation profiles of each cell type. Given the methylation profiles of a set of complex tissue samples and the methylation profiles of constituent cell types this function estimates cell type proportions in each sample by solving constrained least squares problems through quadratic programming. The constraints are that the proportions of constituent cell types are numbers in the [0,1] interval and that the proportions of all cell types in each sample sum up to one.

Value

Matrix with estimated proportions of constituent cell types in each sample.

`estimate_stability` *Estimate stability of EDec models*

Description

This function runs EDec Stage 1 for a series of random subsets of methylation profiles of bulk tissue samples, with varying numbers of constituent cell types. It then computes the similarity of estimated methylation profiles and proportions of constituent cell types across subsets of data for models with each number of constituent cell types. Stability of the model across subsets of the data is generally a good indicator of which number of cell types is an appropriate choice for that dataset.

Usage

```
estimate_stability(meth_bulk_samples, informative_loci, possible_num_ct,
  subset_prop = 0.8, num_subsets = 5, reps_per_subset = 1,
  max_its = 1000, rss_diff_stop = 1e-08)
```

Arguments

`meth_bulk_samples`
: Matrix with methylation profiles of bulk tissue samples. Rows correspond to loci/probes and columns correspond to different samples.

`informative_loci`
: A vector containing names (strings) of rows corresponding to loci/probes that are informative for distinguishing cell types.

`possible_num_ct`
: A vector of containing the possible numbers of cell types to be used in EDec Stage 1

`subset_prop` Proportion of samples from the full dataset to be included in each subset of the data.

`num_subsets` Number of random subsets of the data on which EDec Stage 1 with different numbers of cell types will be tested.

`reps_per_subset` How many times to run EDec Stage 1 with each number of cell types in each subset of the data.

`max_its` Maximum number of iterations after which the EDec Stage 1 algorithm will stop.

`rss_diff_stop` Maximum difference between the residual sum of squares of the model in two consecutive iterations for the EDec Stage 1 algorithm to converge.

Details

A specified number of subsets (`num_subsets`) of the samples with methylation profiles will be generated by randomly selecting a fraction (`subset_prop`) of the columns of `meth_bulk_samples`. For each of those subsets of samples, EDec Stage 1 will be run using all possible number of cell types (`possible_num_ct`). Since different runs of EDec Stage 1 with the same parameters can give different results, there is also the option of running EDec Stage 1 multiple times (`reps_per_subset`) with each number of cell types in each subset of the data, and keeping the best fitting model. Once all runs of EDec Stage 1 are complete, the estimated methylation profiles and proportions of constituent cell types for each given number of constituent cell types will be compared across data subsets. Such comparisons will be made by computing the Pearson correlation between methylation profiles or proportion estimates for the same cell type in each pair of data subsets. To determine which methylation profiles or proportion estimates correspond to the same cell type in two runs of EDec, this function will compute the correlation between every pair of estimated methylation profiles, and find the permutation of the correlation matrix that is most similar to the identity matrix.

Value

A list with the following components:

`most_stable_num_ct` The number of cell types giving the most stable models across the data subsets. Minimum Pearson correlation between either methylation or proportion estimates across all data subsets is used to determine most stable model.

`methylation_estimates` A list containing matrices of average methylation profiles of constituent cell types for each data subset and number of cell types.

`proportion_estimates` A list containing matrices of proportions of constituent cell types in each input sample for each data subset and number of cell types.

`stability_metric_meth` A matrix containing 0 to 100th quantiles, with 5 methylation profiles of constituent cell types across subsets of the data for models with each possible number of cell types. Rows represent different number of cell types. Columns represent different quantiles.

`stability_metric_props` A matrix containing 0 to 100th quantiles, with 5 proportions of constituent cell types across subsets of the data for models with each possible number of cell types. Rows represent different number of cell types. Columns represent different quantiles.

`stability_metric_comb` A matrix containing 0 to 100th quantiles, with 5 proportions of constituent cell types and between methylation profiles of constituent cell types across subsets of the data for models with each possible number of cell types. Rows represent different number of cell types. Columns represent different quantiles.

find_best_perm *Find best permutation of correlation matrix*

Description

This function finds the permutation of a correlation matrix that is most similar to the identity matrix.

Usage

```
find_best_perm(cor_matrix)
```

Arguments

cor_matrix	A matrix of pairwise correlations between a set of numeric vectors

Value

A list containing the following components:

pmatrix The permuted correlation matrix.

pvec The permutation vector, such that pmatrix = cor_matrix[pvec,].

perform_t_tests_all_classes_each_pair *Comparison of each pair of classes*

Description

For each row of a matrix, this function performs T-tests comparing the values of that row for samples in one class against the values in that row samples in another class. Such comparisons are done for each pair of reference classes.

Usage

```
perform_t_tests_all_classes_each_pair(data_matrix, class_vector)
```

Arguments

data_matrix	Matrix containing the data for all samples. Columns correspond to different samples.
class_vector	A vector of numbers or strings representing the classes associated with each column in data_matrix.

Value

A list containing:

p.values A matrix containing the p-values for each comparison. Each column corresponds to the comparison between a pair of classes.

diff.means A matrix containing the difference between group means for each comparison. Each column corresponds to the comparison between a pair of classes.

perform_t_tests_all_classes_one_vs_rest

Comparison of one class against all others

Description

For each row of a matrix, this function performs T-tests comparing the values of that row for samples in one class against the values in that row for all other samples. Such comparisons are done for each reference class in turn.

Usage

```
perform_t_tests_all_classes_one_vs_rest(data_matrix, class_vector)
```

Arguments

data_matrix	Matrix containing the data for all samples. Columns correspond to different samples.
class_vector	A vector of numbers or strings representing the classes associated with each column in data_matrix.

Value

A list containing:

p.values A matrix containing the p-values for each comparison. Each column corresponds to the comparison between a class and all others.

diff.means A matrix containing the difference between group means for each comparison. Each column corresponds to the comparison between a class and all others.

perform_t_tests_all_rows

T-tests for comparing all rows of two matrices

Description

perform_t_tests_all_rows performs two sample T-tests comparing the values in each row of one matrix against the values of the matching row in a second matrix.

Usage

```
perform_t_tests_all_rows(data_group_1, data_group_2)
```

Arguments

data_group_1	A matrix of numbers.
data_group_2	A matrix of numbers.

Value

A matrix where each row corresponds to comparisons between the corresponding rows in `data_group_1` and `data_group_2`. The first column contains the p-values, the second column contains the means of data_group_1 for each row, and the third column contains the means of data_group_2 for each row.

`run_edec_stage_0` *Select informative loci (EDec stage 0)*

Description

`run_edec_stage_0` selects loci/probes that display cell type specific patterns of methylation, and are likely to be informative for deconvolution. This selection is based on comparison of reference methylation profiles representing different cell types.

Usage

```
run_edec_stage_0(reference_meth, reference_classes, max_p_value, num_markers,
  version = "one.vs.rest")
```

Arguments

`reference_meth`	A matrix of reference methylation profiles. Rows correspond to different loci/probes. Columns correspond to different reference samples.
`reference_classes`	A vector of numbers or strings representing the classes associated with each reference methylation profile in `reference_meth`.
`max_p_value`	The maximum p-value, from T-tests comparing reference classes, for a locus to be considered as part of the chosen set of marker loci.
`num_markers`	Number of marker loci to be extracted.
`version`	Either "one.vs.rest" or "each.pair". Specifies whether the marker selection procedure should compare each reference class against all others, or if independent comparisons should be made for each pair of reference classes.

Details

There are two versions of this method. The user can choose which one to use by setting the `version` argument to either "one.vs.rest" (default) or "each.pair". In the first version, the method will perform two-sample T-tests to compare the levels of methylation on each locus/probe between references of one class against all other references. These comparisons will be done for each reference class in turn. The method will then select the most hyper- and hypo-methylated loci/probes that showed p-value less than or equal to `max_p_value`. The same number of markers will be selected from each comparison, adding up to `num_markers`. The difference between the "one.vs.rest" version and the "each.pair" version of the method is that, instead of comparing one class of references against all others as is done in "one.vs.rest" version, the "each.pair" version will perform comparisons between each pair of reference classes.

Value

A vector with the names of the loci/probes selected as markers.

Examples

```
if (requireNamespace("EDecExampleData",quietly=TRUE)) {
  informative_loci <-
    run_edec_stage_0(reference_meth = EDecExampleData::reference_meth,
                     reference_classes = EDecExampleData::reference_meth_class,
                     max_p_value = 1e-5,
                     num_markers = 500,
                     version = "one.vs.rest")
} else {
  print("To run this example, please install EDecExampleData package from
    github by running devtools::install_github('BRL-BCM/EDecExampleData')")
}
```

`run_edec_stage_1` *Run EDec stage 1 algorithm.*

Description

`run_edec_stage_1` takes as input the methylation profiles of complex tissue samples, the set of loci with high variability in methylation across cell types, and the number of constituent cell types. It then estimates the average methylation profiles of constituent cell types, and the proportions of constituent cell types in each input sample.

Usage

```
run_edec_stage_1(meth_bulk_samples, informative_loci, num_cell_types,
  max_its = 2000, rss_diff_stop = 1e-10)
```

Arguments

`meth_bulk_samples`	Matrix with methylation profiles of bulk tissue samples. Rows correspond to loci/probes and columns correspond to different samples.
`informative_loci`	A vector containing names (strings) of rows corresponding to loci/probes that are informative for distinguishing cell types.
`num_cell_types`	Number of cell types to use in deconvolution.
`max_its`	Maximum number of iterations after which the algorithm will stop.
`rss_diff_stop`	Maximum difference between the residual sum of squares of the model in two consecutive iterations for the algorithm to converge.

Details

The first stage of EDec performs constrained matrix factorization to find cell type specific methylation profiles and constituent cell type proportions that minimize the Euclidian distance between their linear combination and the original matrix of tissue methylation profiles. The minimization algorithm involves an iterative procedure that, in each round, alternates between estimating constituent cell type proportions (using `estimate_props_qp` function) and methylation profiles (using

estimate_meth_qp function) by solving constrained least squares problems through quadratic programming. The minimization problem is made tractable by the constraints that methylation measurements (beta values) and cell type proportions are numbers in the [0,1] interval, and that cell type proportions within a sample add up to one. These constraints restrict the space of possible solutions, thus making it possible for the local iterative search to reproducibly find a global minimum and an accurate solution. One key requirement for EDec is that cell type proportions vary across samples. A second requirement is that there must be significant differences across constituent cell type methylation profiles. The latter requirement can be met by providing EDec with loci expected to vary in methylation levels across constituent cell types.

Value

A list with the following components:

`methylation` A matrix with average methylation profiles of constituent cell types. Rows represent different loci/probes and columns represent different cell types.

`proportions` A matrix with proportions of constituent cell types in each input sample. Rows represent different samples. Columns represent different cell types.

`iterations` Number of iterations the method went through before reaching convergence or maximum number of iterations.

`explained.variance` Proportion of variance in input methylation profiles over informative loci explained by the final model.

`res.sum.squares` Residual sum of squares for the final model over the set of informative loci.

`aic` Akaike Information Criterion for the final model over the set of informative loci.

`rss.per.iteration` Vector of residual sum of squares for the models generated in each iteration of the algorithm.

`run_edec_stage_2` *Run EDec stage 2 algorithm*

Description

This function implements the second stage of the EDec method. It takes as input the gene expression profiles of complex tissue samples, and the proportions of constituent cell types in each sample. It then estimates average and standard errors of cell type specific gene expression profiles.

Usage

```
run_edec_stage_2(gene_exp_bulk_samples, cell_type_props)
```

Arguments

`gene_exp_bulk_samples`
: Matrix of methylation profiles of bulk complex tissue samples. Columns correspond to different samples and rows correspond to different loci/probes.

`cell_type_props`
: Matrix of proportions of constituent cell types. Columns correspond to different cell types and rows correspond to different bulk tissue samples.

Details

EDec assumes that the gene expression profiles of complex tissue samples correspond to the linear combination of cell type proportions and gene expression profiles of each cell type. Given the gene expression profiles of a set of complex tissue samples and the proportions of constituent cell types in each sample, this function estimates average gene expression profiles of constituent cell types by solving constrained least squares problems through quadratic programming. The constraint is that the gene expression profiles of constituent cell types are numbers greater than or equal to zero.

Value

A list with the following components:

`means` A matrix with the estimated average gene expression profiles of constituent cell types. Rows correspond to different genes. Columns correspond to different cell types.

`std.errors` A matrix with estimated standard errors for each cell type specific gene expression estimate. Rows correspond to different genes. Columns correspond to different cell types.

`degrees.of.freedom` Number of degrees of freedom for estimates of cell type specific gene expression

`explained.variances` Vector with the proportion of variance in input expression of each gene across samples explained by the final model.

`residuals` Matrix with the difference between the original gene expression values and the linear combination between proportions of constituent cell types and gene expression profiles of constituent cell types.

Index