Package ‘ActiveDriver’

August 23, 2017

Version 1.0.0

License GPL (>= 2)

Description

A mutation analysis tool that discovers cancer driver genes with frequent mutations in protein signalling sites such as post-translational modifications (phosphorylation, ubiquitination, etc). The Poisson generalised linear regression model identifies genes where cancer mutations in signalling sites are more frequent than expected from the sequence of the entire gene. Integration of mutations with signalling information helps find new driver genes and propose candidate mechanisms to known drivers. Reference: Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Juri Reimand and Gary D Bader. Molecular Systems Biology (2013) 9:637 <doi:10.1038/msb.2012.68>.

Title Finding Cancer Driver Proteins with Enriched Mutations in Post-Translational Modification Sites

Depends R (>= 3.0)

Imports stats, parallel, MASS

Collate 'ActiveDriver.R'

RoxygenNote 6.0.1.9000

NeedsCompilation no

Author Juri Reimand [aut, cre]

Maintainer Juri Reimand <juri.reimand@utoronto.ca>

Repository CRAN

Date/Publication 2017-08-23 20:55:51 UTC

R topics documented:

ActiveDriver *Identification of active protein sites (post-translational modification sites, signalling domains, etc) with specific and significant mutations.*

Description

Identification of active protein sites (post-translational modification sites, signalling domains, etc) with specific and significant mutations.

Usage

```
ActiveDriver(sequences, seq_disorder, mutations, active_sites, flank = 7,
  mid_flank = 2, mc.cores = 1, simplified = FALSE,
  return_records = FALSE, skip_mismatch = TRUE,
  regression_type = "poisson", enriched_only = TRUE)
```

Arguments

`sequences` character vector of protein sequences, names are protein IDs.

`seq_disorder` character vector of disorder in protein sequences, names are protein IDs and values are strings 1/0 for disordered/ordered protein residues.

`mutations` data frame of mutations, with [gene, sample_id, position, wt_residue, mut_residue] as columns.

`active_sites` data frame of active sites, with [gene, position, residue, kinase] as columns. Kinase field may be blank and is shown for informative purposes.

`flank` numeric for selecting region size around active sites considered important for site activity. Default value is 7. Ignored in case of simplified analysis.

`mid_flank` numeric for splitting flanking region size into proximal (<=X) and distal (>X). Default value is 2. Ignored in case of simplified analysis.

`mc.cores` numeric for indicating number of computing cores dedicated to computation. Default value is 1.

`simplified` true/false for selecting simplified analysis. Default value is FALSE. If TRUE, no flanking regions are considered and only indicated sites are tested for mutations.

`return_records` true/false for returning a collection of gene records with more data regarding sites and mutations. Default value is FALSE.

`skip_mismatch` true/false for skipping mutations whose reference protein residue does not match expected residue from FASTA sequence file.

`regression_type`

’nb’ for negative binomial, ’poisson’ for poisson GLM. The latter is default.

`enriched_only` true/false to indicate whether only sites with enriched active site mutations will be included in the final p-value estimation (TRUE is default). If FALSE, sites with less than expected mutations will be also included.

Value

list with the following components: @return all_active_mutations - table with mutations that hit or flank an active site. Additional columns of interest include Status (DI - direct active mutation; N1 - proximal flanking mutation; N2 - distal flanking mutation) and Active_region (region ID of active sites in that protein).

all_active_sites -

all_region_based_pval - p-values for regions of sites, statistics on observed mutations (obs) and expected mutations (exp, low, high based on mean and s.d. from Poisson sampling). The field Region identifies region in all_active_sites.

Author(s)

Juri Reimand <juri.reimand@utoronto.ca>

References

Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers (2013, Molecular Systems Biology) by Juri Reimand and Gary Bader.

Examples

```
data(ActiveDriver_data)

phos_results = ActiveDriver(sequences, sequence_disorder, mutations, phosphosites)
ovarian_mutations = mutations[grep("ovarian", mutations$sample_id),]
phos_results_ovarian = ActiveDriver(sequences, sequence_disorder, ovarian_mutations, phosphosites)
GBM_muts = mutations[grep("glioblastoma", mutations$sample_id),]
kin_rslt_GBM = ActiveDriver(sequences, sequence_disorder, GBM_muts, kinase_domains, simplified=TRUE)

kin_results = ActiveDriver(sequences, sequence_disorder, mutations, kinase_domains, simplified=TRUE)
```

`kinase_domains` *Example kinase domains for ActiveDriver*

Description

A dataset describing kinase domains. The variables are as follows:

Usage

```
data(ActiveDriver_data)
```

Format

A data frame with 1 observation of 4 variables

Details

- gene. the gene symbol of the gene where the kinase domain occurs
- position. the position in the protein sequence where the kinase domain begins
- phos. TRUE
- residue. the kinase domain residues

mutations *Example mutations for ActiveDriver*

Description

A dataset describing mis-sense mutations (i.e., substitutions in proteins). The variables are as follows:

Usage

```
data(ActiveDriver_data)
```

Format

A data frame with 408 observations of 5 variables

Details

- gene. the mutated gene
- sample_id. the sample where the mutation originates
- position. the position in the protein sequence where the mutation occurs
- wt_residue. the wild-type residue
- mut_residue. the mutant residue

phosphosites *Example phosphosites for ActiveDriver*

Description

A dataset describing protein phosphorylation sites. The variables are as follows:

Usage

```
data(ActiveDriver_data)
```

Format

A data frame with 131 observations of 4 variables

Details

- gene. the gene symbol the phosphosite occurs in
- position. the position in the protein sequence where the phosphosite occurs
- residue. the phosphosite residue
- kinase. the kinase that phosphorylates this site

`read_fasta` *Read FASTA file as character vector.*

Description

Read FASTA file as character vector.

Usage

```
read_fasta(fname)
```

Arguments

`fname`	name of file to be read.

Value

character vector with names corresponding to annotations from FASTA.

`sequences` *Example protein sequences for ActiveDriver*

Description

A dataset containing the sequences of four proteins.

Usage

```
data(ActiveDriver_data)
```

Format

A named character vector with 4 elements

`sequence_disorder` *Example protein disorder for ActiveDriver*

Description

A dataset containing the disorder of four proteins.

Usage

```
data(ActiveDriver_data)
```

Format

A named character vector with 4 elements

Index