Package 'seq.hotSPOT'

September 16, 2024

Type Package

Title Targeted sequencing panel design based on mutation hotspots

Version 1.4.0

Description seq.hotSPOT provides a resource for designing effective sequencing panels to help improve mutation capture efficacy for ultradeep sequencing projects. Using SNV datasets, this package designs custom panels for any tissue of interest and identify the genomic regions likely to contain the most mutations. Establishing efficient targeted sequencing panels can allow researchers to study mutation burden in tissues at high depth without the economic burden of whole-exome or whole-genome sequencing. This tool was developed to make high-depth sequencing panels to study low-frequency clonal mutations in clinically normal and cancerous tissues.

License Artistic-2.0

Encoding UTF-8

LazyData FALSE

RoxygenNote 7.2.3

biocViews Software, Technology, Sequencing, DNASeq, WholeGenome

Imports R.utils, hash, stats, base, utils

Suggests BiocStyle, knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

URL https://github.com/sydney-grant/seq.hotSPOT

BugReports https://github.com/sydney-grant/seq.hotSPOT/issues

Config/testthat/edition 3

git_url https://git.bioconductor.org/packages/seq.hotSPOT

git_branch RELEASE_3_19

git_last_commit 8b7e497

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-09-15

Author Sydney Grant [aut, cre],
Lei Wei [aut],
Gyorgy Paragh [aut]

Maintainer Sydney Grant <sydney.grant@roswellpark.org>

Contents

`amp_pool` *create amplicon pool*

Description

create a dataframe containing the coordinates of all potential amplicons for hotspot testing

Usage

```
amp_pool(data, amp)
```

Arguments

`data`	A dataframe containing the location of each mutation.
`amp`	The length of amplicons in number of base pairs

Details

This algorithm searches the mutational dataset (input) for mutational hotspot regions on each chromosome:

1. Starting at the mutation with the lowest chromosomal position (primary mutation), using a modified rank and recovery system, the algorithm searches for the closest neighboring mutation.

2. If the neighboring mutation is less than one amplicon, in distance, away from the primary mutation, the neighboring mutation is included within the hotspot region. a. This rank and recovery system is repeated, integrating mutations into the hotspot region until the neighboring mutation is greater than or equal to the length of one amplicon in distance, from the primary mutation. b. Once neighboring mutations equal or exceed one amplicon in distance from the primary mutation, incorporation into the hotspot region, halts incorporation.

3. For hotspots within the one amplicon range, from the lowest to highest mutation location, this area is covered by a single amplicon and added to an amplicon pool, with a unique ID. a. The center of these single amplicons is then defined by the weighted distribution of mutations.

4. For all hotspots larger than one amplicon, the algorithm examines 5 potential amplicons at each covered mutation in the hotspot: a. one amplicon directly upstream of the primary mutation b. one amplicon directly downstream of the primary mutation c. one amplicon including the mutation at the end of the read and base pairs (amplicon length - 1) upstream d. one amplicon including the mutation at the beginning of the read and base pairs (amplicon length - 1) downstream e. one amplicon with the mutation directly in the center.

5. All amplicons generated for each hotspot region of interest, are assigned a unique ID and added to the amplicon pool.

The mutation dataset should include two columns containing the chromosome and genomic position, the columns should be names "chr" and "pos" respectively. Optionally the gene names for each mutation may be included under a column names "gene".

Value

A dataframe containing the genomic coordinates of all potential amplicons

Examples

```
data("mutation_data")
amp_pool(mutation_data, 100)
```

`com_hotspot` *comprehensive amplicon ranking*

Description

create a targeted sequencing panel by finding which amplicons will likely capture the most mutations using a pseudo-exhaustive selection method

Usage

```
com_hotspot(fw_panel, bins, data, amp, len, size, include_genes)
```

Arguments

`fw_panel`	a dataframe containing the sequencing panel designed by fw_hotspot
`bins`	A dataframe containing all potential amplicons
`data`	A dataframe containing the location of each mutation.
`amp`	The length of amplicons in number of base pairs
`len`	The total length of sequencing panel in number of base pairs
`size`	the threshold of hotspot size to split up in number of amplicons
`include_genes`	True or False based on whether dataset includes gene names

Details

Comprehensive Selection Sequencing Panel Identifier (Optimal mutation capture)

1. To conserve computational power, the forward selection sequencing panel identifier is run to determine the lowest number of mutations per amplicon (mutation frequency) that need to be included in the predetermined length sequencing panel. a. any amplicon generated by the algorithm, which is less than this threshold value, will be removed.

2. For the feasible exhaustive selection of amplicon combinations covering hotspot areas larger than the predefined number of amplicons in length, the algorithm breaks these large regions into multiple smaller regions. a. The amplicons covering these regions are pulled from the amplicon pool, based on their unique IDs.

3. The algorithm finds both the minimum number of amplicons overlap and all positions with this value and identifies the region with the longest continuous spot of minimum value. a. The region is split at the center of this longest continuous minimum post values and continues the splitting process until all smaller regions are less than the "n" number amplicon length set by the user. i. As this set number of amplicons decreases, the computation time required also often decreases.

4. All amplicons contained in these bins are added back to the amplicon pool, based on a new unique ID.

5. Amplicons covering hotspots less than or equal to one amplicon length are added to the final sequencing panel dataset.

6. To determine the optimal combination of amplicons for each region, the number of amplicons necessary for full coverage of the bin is calculated.

7. A list is generated of every possible combination of n, number of amplicons, needed. For each combination of amplicons: a. amplicons that would not meet the threshold of unique mutations are filtered out, and the number of all mutations captured by these amplicons is calculated. b. the combination of amplicons that yields the highest number of mutations is added to the final sequencing panel.

8. All amplicons in the final sequencing panel are ranked from highest to lowest based on the number of mutations they cover.

9. All amplicons capturing the number of mutations equal to the cutoff are further ranked to favor amplicons that have mutations closer in location to the center of the amplicon.

10. Cumulative base-pair length and cumulative mutations covered by each amplicon are calculated. a. Depending on the desired length of the targeted panel, a cutoff may be applied to remove all amplicons which fall below a set cumulative length.

Value

A dataframe containing the genomic coordinates for targeted sequencing panel

Examples

```
data("mutation_data")
my_bins <- amp_pool(mutation_data, 100)

my_fw_panel <- fw_hotspot(my_bins, mutation_data, 100, 1000, TRUE)

com_hotspot(my_fw_panel, my_bins, mutation_data, 100, 1000, 3, TRUE)
```

`fw_hotspot` *forward amplicon ranking*

Description

create a targeted sequencing panel by finding which amplicons will likely capture the most mutations

Usage

```
fw_hotspot(bins, data, amp, len, include_genes)
```

Arguments

`bins`	A dataframe containing all potential amplicons
`data`	A dataframe containing the location of each mutation.
`amp`	The length of amplicons in number of base pairs
`len`	The total length of sequencing panel in number of base pairs
`include_genes`	True or False based on whether dataset includes gene names

Details

Forward Selection Sequencing Panel Identifier

1. Amplicons covering hotspots less than or equal to one amplicon in length, are added to the final sequencing panel dataset.

2. For amplicons covering larger hotspot regions, the algorithm uses a forward selection method to determine the optimal combination of amplicons to use in the sequencing panel: a. the algorithm first identifies the amplicon containing the highest number of mutations b. the algorithm then identifies the next amplicon, which contains the highest number of new mutations. c. this process continues until all mutations are covered by at least one amplicon

3. Each of these amplicons are then added to the final sequencing panel, with their own unique IDs.

4. All amplicons in the final sequencing panel are ranked from highest to lowest based on the number of mutations they cover.

5. The algorithm then calculates the cumulative base-pair length and the cumulative mutations covered by each amplicon.

6. Dependent on the desired length of the targeted panel, a cutoff may be applied to remove all amplicons which fall below a set cumulative length.

Value

A dataframe containing the genomic coordinates for targeted sequencing panel

Examples

```
data("mutation_data")
my_bins <- amp_pool(mutation_data, 100)

fw_hotspot(my_bins, mutation_data, 100, 1000, TRUE)
```

`mutation_data` *Single Nucleotide Variants in Clinically-Normal Epidermis*

Description

dataframe containing the chromosome and base pair position from single nucleotide variants of ultradeep sequencing epidermis.

Usage

```
mutation_data
```

Format

'mutation_data' a dataframe with 3 columns and 201 rows:

chr Chromosome which the mutation is located on

pos Base pair position of mutation

gene Name of gene affected by mutation

Details

mutation_data

Value

A dataframe containing the chromosome number, base pair location and optional gene name of mutations

References

Wei L, Christensen SR, Fitzgerald ME, Graham J, Hutson ND, Zhang C, Huang Z, Hu Q, Zhan F, Xie J, Zhang J, Liu S, Remenyik E, Gellen E, Colegio OR, Bax M, Xu J, Lin H, Huss WJ, Foster BA, Paragh G. Ultradeep sequencing differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. Sci Adv. 2021 Jan 1;7(1):eabd7703. doi: 10.1126/sciadv.abd7703. PMID: 33523857; PMCID: PMC7775785.

seq.hotSPOT *Targeted sequencing panel design based on mutation hotspots*

Description

seq.hotSPOT provides a resource for designing effective sequencing panels to help improve mutation capture efficacy for ultradeep sequencing projects. Using SNV datasets, this package designs custom panels for any tissue of interest and identify the genomic regions likely to contain the most mutations. Establishing efficienttargeted sequencing panels can allow researchers to study mutation burden in tissues at high depth without the economic burden of whole-exome or whole-genome sequencing. This tool was developed to make high-depth sequencing panels to study low-frequency clonal mutations in clinically normal and cancerous tissues.

Value

A package containing functions which generate optimal targeted sequencing panels based on mutation hotspots.

Index