

# Package ‘copyseparator’

July 6, 2022

**Type** Package

**Title** Assembling Long Gene Copies from Short Read Data

**Version** 1.1.0

**Author** Lei Yang

**Maintainer** Lei Yang <leiyangslu@gmail.com>

**Description** Assembles two or more gene copies from short-read Next-Generation Sequencing data. Works best when there are only two gene copies and read length  $\geq 250$  base pairs. High and relatively even coverage are important.

**License** GPL-2

**URL** <https://github.com/LeiYang-Fish/copyseparator>

**BugReports** <https://github.com/LeiYang-Fish/copyseparator/issues>

**Depends** R ( $\geq 3.5.0$ )

**Encoding** UTF-8

**Imports** ape, seqinr, stringr, kmer, DECIPHER, beepr, Biostrings, grDevices

**RoxxygenNote** 7.1.2

**Suggests** knitr, rmarkdown, testthat ( $\geq 3.0.0$ )

**VignetteBuilder** knitr

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2022-07-06 13:10:02 UTC

## R topics documented:

copy_assemble . . . . .	2
copy_detect . . . . .	2
copy_separate . . . . .	3
copy_validate . . . . .	4
sep_assem . . . . .	5
subset_downsize . . . . .	6

**Index****7**


---

copy_assemble	<i>copy_assemble</i>
---------------	----------------------

---

**Description**

Assembles a small number of overlapping DNA sequences into their respective gene copies.

**Usage**

```
copy_assemble(filename, copy_number, verbose = 1)
```

**Arguments**

filename	A fasta alignment of a small number of overlapping DNA sequences (results from "copy_separate") covering the entire length of the target gene. Check the alignment carefully before proceeding.
copy_number	An integer (e.g. 2,3, or 4) giving the anticipated number of gene copies. Must be the same value as used for "copy_separate".
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

**Value**

A fasta alignment of the anticipated number of full-length gene copies.

**Examples**

```
## Not run:
copy_assemble("inst/extdata/combined_con.fasta",2,1)

## End(Not run)
```

---

copy_detect	<i>copy_detect</i>
-------------	--------------------

---

**Description**

Separates two or more gene copies from a single subset of short reads.

**Usage**

```
copy_detect(filename, copy_number, verbose = 1)
```

**Arguments**

filename	A fasta file contains short reads from a single subset generated by "subset_downsize".
copy_number	An integer (e.g. 2,3, or 4) giving the anticipated number of gene copies in the input file.
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

**Value**

A fasta alignment of the anticipated number of gene copies.

**Examples**

```
## Not run:
copy_detect("inst/extdata/toysubset.fasta",2,1)

## End(Not run)
```

---

copy_separate	<i>copy_separate</i>
---------------	----------------------

---

**Description**

Separates two or more gene copies from short-read Next-Generation Sequencing data into a small number of overlapping DNA sequences.

**Usage**

```
copy_separate(
  filename,
  copy_number,
  read_length,
  overlap = 225,
  rare_read = 10,
  verbose = 1
)
```

**Arguments**

filename	A fasta file contains thousands of short reads that have been mapped to a reference. The reference and reads that are not directly mapped to the reference need to be removed after mapping.
copy_number	An integer (e.g. 2,3, or 4) giving the anticipated number of gene copies in the input file.
read_length	An integer (e.g. 250, or 300) giving the read length of your Next-generation Sequencing data. This method is designed for read length $\geq 250$ bp.

overlap	An integer describing number of base pairs of overlap between adjacent subsets. More overlap means more subsets. Default 225.
rare_read	A positive integer. During clustering analyses, clusters with less than this number of reads will be ignored. Default 10.
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

### Value

A fasta alignment of a small number of overlapping DNA sequences covering the entire length of the target gene. Gene copies can be assembled by reordering the alignment manually or use the function "copy\_assemble".

### Examples

```
## Not run:
copy_separate("inst/extdata/toydata.fasta", 2, 300, 225, 10, 1)

## End(Not run)
```

---

copy_validate	<i>copy_validate</i>
---------------	----------------------

---

### Description

A tool to help identify incorrectly assembled chimeric sequences.

### Usage

```
copy_validate(filename, copy_number, read_length, verbose = 1)
```

### Arguments

filename	A DNA alignment in fasta format that contains sequences of two or more gene copies (e.g. results from "copy_assemble").
copy_number	An integer (e.g. 2,3, or 4) giving the number of gene copies in the input file.
read_length	An integer (e.g. 250, or 300) giving the read length of your Next-generation Sequencing data.
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

### Value

A histogram in pdf format showing the relationships between the physical distance between neighboring variable sites and read length.

**Examples**

```
## Not run:
copy_validate("inst/extdata/Final_two_copies.fasta",2,300,1)

## End(Not run)
```

---

```
sep_assem          sep_assem
```

---

**Description**

Separates two or more gene copies from short-read Next-Generation Sequencing data into a small number of overlapping DNA sequences and assemble them into their respective gene copies.

**Usage**

```
sep_assem(
  filename,
  copy_number,
  read_length,
  overlap = 225,
  rare_read = 10,
  verbose = 1
)
```

**Arguments**

filename	A fasta file contains thousands of short reads that have been mapped to a reference. The reference and reads that are not directly mapped to the reference need to be removed after mapping.
copy_number	An integer (e.g. 2,3, or 4) giving the anticipated number of gene copies in the input file.
read_length	An integer (e.g. 250, or 300) giving the read length of your Next-generation Sequencing data. This method is designed for read length $\geq 250$ bp.
overlap	An integer describing number of base pairs of overlap between adjacent subsets. More overlap means more subsets. Default 225.
rare_read	A positive integer. During clustering analyses, clusters with less than this number of reads will be ignored. Default 10.
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

**Value**

A fasta alignment of the anticipated number of full-length gene copies.

**Examples**

```
## Not run:
sep_assem("inst/extdata/toydata.fasta", 2, 300, 225, 10, 1)

## End(Not run)
```

---

subset_downsize	<i>subset_downsize</i>
-----------------	------------------------

---

**Description**

Subdivides the imported read alignment into subsets and then downsizes each subset by deleting those sequences that have too many gaps or missing data.

**Usage**

```
subset_downsize(filename, read_length, overlap, verbose = 1)
```

**Arguments**

filename	A fasta file contains thousands of short reads that have been mapped to a reference. The reference and reads that are not directly mapped to the reference need to be removed after mapping.
read_length	An integer (e.g. 250, or 300) giving the read length of your Next-generation Sequencing data. This method is designed for read length $\geq 250$ bp.
overlap	An integer describing number of base pairs of overlap between adjacent subsets. More overlap means more subsets.
verbose	Turn on (verbose=1; default) or turn off (verbose=0) the output.

**Value**

A number of overlapping subsets (before and after downsizing) of the input alignment.

**Examples**

```
## Not run:
subset_downsize("inst/extdata/toydata.fasta", 300, 225, 1)

## End(Not run)
```

# Index

`copy_assemble`, 2

`copy_detect`, 2

`copy_separate`, 3

`copy_validate`, 4

`sep_assem`, 5

`subset_downsize`, 6