

Package ‘scShardSplitRef’

July 23, 2026

Title Build Custom Reference Genome for Single-Cell Multiome ATAC and Gene Expression Data of Large-Genome Species

Version 0.1.0

Description Provides utilities that enable researchers to build custom references for the Multiome ATAC Gene Expression sequencing data for large-genome species without manual intervention.

License GPL (>= 3)

Depends R (>= 4.3.0)

Imports cli, rlang, stringi, utils, withr

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

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Config/roxyglobals/filename globals.R

Config/roxyglobals/unique FALSE

Config/roxygen2/version 8.0.0

Config/Needs/build moodymudskipper/devtag

NeedsCompilation no

Author Irina Kuznetsova [aut, cre] (ORCID: <https://orcid.org/0009-0001-5411-6063>),
Luke Pembleton [aut] (ORCID: <https://orcid.org/0000-0001-8131-8210>),
Luca Curci [aut] (ORCID: <https://orcid.org/0000-0002-5780-3029>),
Adam H. Sparks [aut] (ORCID: <https://orcid.org/0000-0002-0061-8359>),
Curtin University [fnd, cph] (ROR: <https://ror.org/02n415q13>),
Grains Research and Development Corporation [fnd, cph] (GRDC Project CUR2210-005OPX (AAGI-CU), ROR: <https://ror.org/02xwr1996>)

Maintainer Irina Kuznetsova <irina.kuznetsova@curtin.edu.au>

Repository CRAN

Date/Publication 2026-07-23 12:30:02 UTC

Contents

determine_split_regions	2
process_gtf	3
scShardSplitRef_options	5
Index	6

determine_split_regions
<i>Determine split regions for reference sharding</i>

Description

Builds BED-like split intervals from genomic regions and GTF gene annotations. If the start position in a genomic region is not 0, it is assumed to mark the start of a centromere and the region is split around that point. Proposed boundaries are shifted rightward when they fall too close to gene intervals, and oversized chunks are repeatedly re-split until all segments satisfy the requested size limit.

Usage

```
determine_split_regions(  
  bed,  
  gtf,  
  output_bed,  
  limit = 2L^29L,  
  shift_by = 1L,  
  clearance = 1L  
)
```

Arguments

bed	Path to a 3-column BED-like file (chr, start, end). Columns should be TAB-delimited with chromosome name, region start (0-based), and region end coordinates.
gtf	Path to a GTF file containing gene annotations. Must contain standard 9 GTF columns. Comments starting with '#' are ignored.
output_bed	Path to the BED file to be written. Directory must exist.
limit	Maximum allowed width (bp) for output intervals. Defaults to 2^29 (536,870,912 bp). Must be larger than any single gene plus the shift distance.
shift_by	Step size (bp) used to iteratively walk boundaries to the right when they violate gene clearance rules. Defaults to 1 bp.
clearance	Minimum required clearance (bp) between split boundaries and gene intervals on both sides. Defaults to 1 bp.

Details

The function performs the following steps:

1. Reads genomic regions and gene annotations from input files.
2. Checks that the provided limit is feasible given gene positions and requested two-sided boundary clearance.
3. For each region, splits it at the centromere (if start > 0) and applies the size limit.
4. Adjusts boundaries so each is at least clearance bp away from genes on both sides, moving by shift_by bp per iteration when needed.
5. Iteratively subdivides intervals exceeding the limit until all intervals satisfy the constraint.
6. Writes final split regions to BED file in the active R session's tempdir().

Value

Invisible character: path to the written BED file.

Examples

```
# load example files from this package and write the output bed file to
# R's tempdir()

chromosomes <- system.file("extdata",
                           "AlgorithmToy.bed",
                           package = "scShardSplitRef",
                           mustWork = TRUE)
genes <- system.file("extdata",
                    "AlgorithmToy.gtf",
                    package = "scShardSplitRef",
                    mustWork = TRUE)

determine_split_regions(
  bed = chromosomes,
  gtf = genes,
  output_bed = file.path(tempdir(), "split_regions.bed"),
  limit = 2L^29L,
  shift_by = 1L,
  clearance = 1L
)
```

process_gtf

Process a GTF file by assigning split-region sequence names

Description

Reads a BED-like "split regions" file and a GTF, assigns each GTF feature to exactly one split-region interval, and replaces the GTF Chr field with a split-region name of the form Chr-RegionStart-RegionEnd.

Usage

```
process_gtf(
  split_regions_bed,
  gtf,
  genome_name,
  genome_version = NULL,
  keep_attributes = NULL,
  out_path = "."
)
```

Arguments

split_regions_bed	Character: File path to a BED-like file containing split intervals as produced by determine_split_regions() . The first three columns must be: Chr, RegionStart, RegionEnd.
gtf	Path to a GTF file containing gene annotations. Must contain standard 9 GTF columns. Comments starting with '#' are ignored.
genome_name	Genome identifier used in the output filename.
genome_version	Genome version string used in the output filename.
keep_attributes	Character vector of attribute keys to keep (e.g., c("gene_id", "transcript_id", "gene_name")). If provided, the GTF attribute column will be filtered to keep only these fields. Defaults to NULL (no filtering).
out_path	Directory where the output GTF will be written. Defaults to the current working directory (" . ").

Details

A feature is considered inside a region when:

- start >= RegionStart
- start < RegionEnd
- end <= RegionEnd

The function aborts if:

- any GTF feature matches more than one region (overlapping regions), or
- any GTF feature matches no region (regions don't fully cover the GTF).

Value

The function writes the filtered GTF file to disk.

Output file

The function will automatically name the file based on the input file and prepends the following string onto the filename: "B1_FINAL_MODIFIED_GTF_genome_name_genome_version.gtf", placing the file in the directory specified in the out_path.

Examples

```
reg_file <- system.file(
  "extdata/AlgorithmToy_DetermineSplitReg.bed",
  package = "scShardSplitRef",
  mustWork = TRUE
)
gtf_file <- system.file(
  "extdata/AlgorithmToy.gtf",
  package = "scShardSplitRef",
  mustWork = TRUE
)

process_gtf(
  split_regions_bed = reg_file,
  gtf = gtf_file,
  genome_name = "Example",
  genome_version = "v1.0",
  out_path = tempdir()
)
```

`scShardSplitRef_options`*Get or Set scShardSplitRef Options*

Description

A convenience function to get or set options used by **scShardSplitRef**.

Usage

```
scShardSplitRef_options(...)
```

Arguments

... Named options to set, or no arguments to retrieve current values.

Value

A list of current option values.

Examples

```
# See currently set options for scShardSplitRef
scShardSplitRef_options()

# Set verbose to FALSE to suppress info/success messages
scShardSplitRef_options(scShardSplitRef.verbose = FALSE)
scShardSplitRef_options()
```

Index

* **scShardSplitRef-options**

scShardSplitRef_options, [5](#)

determine_split_regions, [2](#)

determine_split_regions(), [4](#)

process_gtf, [3](#)

scShardSplitRef_options, [5](#)