Package: iscream (via r-universe)

May 30, 2026

Title Make fast and memory efficient BED file queries, summaries and matrices

Version 1.3.0

Description BED files store ranged genomic data that can be queried even when the files are compressed. iscream can query data from BED files and return them in muliple formats: parsed records or their summary statistics as data frames or GenomicRanges objects, and matrices as matrix, GenomicRanges, or SummarizedExperiment objects. iscream also provides specialized support for importing methylation data.

URL https://huishenlab.github.io/iscream/, https://github.com/huishenlab/iscream/

BugReports https://github.com/huishenlab/iscream/issues/

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

Depends R (>= 4.4)

LinkingTo Rcpp, RcppArmadillo, RcppProgress, RcppSpdlog, Rhtslib, stringfish

Imports Rcpp, Matrix, data.table, methods, pbapply, parallelly, stringfish,

Suggests BiocFileCache, BiocStyle, bsseq, ggplot2, ggridges, knitr, microbenchmark, rmarkdown, GenomicRanges, IRanges, Rsamtools, SummarizedExperiment, S4Vectors, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation yes

SystemRequirements htslib: htslib-devel (rpm) or libhts-dev (deb) & tabix: htslib-tools (rpm) or tabix (deb) & GNU make

biocViews DataImport, Software, Sequencing, SingleCell, DNAMethylation

Config/pak/sysreqs make libbz2-dev liblzma-dev xz-utils zlib1g-dev

Repository https://bioc.r-universe.dev

Date/Publication 2026-04-28 13:05:51 UTC

RemoteUrl https://github.com/bioc/iscream

RemoteRef HEAD

RemoteSha a142d5418b2cf78de1793786a23be25523c6715b

Contents

`get_df_string` *DataFrame to region strings*

Description

Convert DataFrame to a vector of strings. Set feature names in a "name" column

Usage

```
get_df_string(regions_df, feature_col = NULL)
```

Arguments

`regions_df`	A data frame with "chr", "start" and "end" columns
`feature_col`	The data frame column to use as the names of the output string vector

Value

A character vector

Examples

```
(df <- data.frame(chr = c("chr1", "chr2"), start = c(1, 5), end = c(4, 10)))
get_df_string(df)
```

---

get_granges_string *GRanges to region strings*

Description

Coerces GenomicRanges to `chr:start-end` strings with `as.character`. If any regions have the same start and end, `as.character` returns `chr:start` strings which are invalid for the htslib API. These are corrected to `chr:start-start`.

Usage

```
get_granges_string(gr, feature_col = NULL)
```

Arguments

`gr`	A GRanges object
`feature_col`	The `mcols` column to use as the names of the output string vector

Value

A character vector

Examples

```
if (requireNamespace("GenomicRanges", quietly = TRUE)) {
  get_granges_string(GenomicRanges::GRanges(c("chr1:1-10", "chr2:15-20")))
}
```

---

get_threads *Get the number of available threads*

Description

Gets the number of threads iscream is currently set to use, whether the `"iscream.threads"` option is set and how many threads are available for use. To set the number of threads use `set_threads()` or set the `iscream.threads` option in your `~/.Rprofile`. See `?set_threads` for more information.

Usage

```
get_threads()
```

Value

A named vector:

- use_threads = the number of threads iscream will use
- opt_set = whether the option was set by the user
- avail_threads = The number of available threads as reported by parallelly::availableCores

Examples

```
get_threads()
```

---

htslib_version *Get htslib version and available features*

Description

Returns the version of htslib being used by iscream and whether features such as libdeflate support are available. This information may not always correspond to the htslib version used during iscream's installation if a different htslib version is available for linking at runtime.

Usage

```
htslib_version()
```

Value

named vector with "version" containing the version number and "features" containing the available features

Examples

```
htslib_version()
```

---

make_mat *Make a matrix from a numeric column of BED files*

Description

Queries the provided regions and produces a matrix along with genomic positions as a named list (make_mat()), a RangedSummarizedExperiment (make_mat_se()), GRanges (make_mat_gr()). Parallelized across files using threads from the "iscream.threads" option.

Usage

```
make_mat(
  bedfiles,
  regions,
  column,
  mat_name = "value",
  sparse = FALSE,
  prealloc = 10000,
  nthreads = NULL
)

make_mat_se(
  bedfiles,
  regions,
  column,
  mat_name = "value",
  sparse = FALSE,
  prealloc = 10000,
  nthreads = NULL
)

make_mat_gr(
  bedfiles,
  regions,
  column,
  mat_name = "value",
  prealloc = 10000,
  nthreads = NULL
)
```

Arguments

bedfiles	A vector of BED file paths
regions	A vector, data frame or GenomicRanges of genomic regions. See details.
column	The index of the data column needed for the matrix
mat_name	What to name the matrix in the returned object
sparse	Whether to return a sparse matrix
prealloc	The number of rows to initialize the matrices with. If the number of loci are approximately known, this can reduce runtime as fewer resizes need to be made.
nthreads	Set the number of threads to use. Overrides the "iscream.threads" option. See ?set_threads for more information.

Details

The input regions may be string vector in the form "chr:start-end" or a GRanges object. If a data frame is provided, they must have "chr", "start", and "end" columns.

Value

- `make_mat()`: A named list of
  - the matrix with the value of interest
  - a character vector of chromosomes and numeric vector of base positions
  - a character vector of the input sample BED file names
- `make_mat_gr()`: if `GenomicRanges` is available, a `GRanges`
- `make_mat_se()`: if `SummarizedExperiment` is available, a `RangedSummarizedExperiment`

Examples

```
bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)
# examine the bedfiles
colnames <- c("chr", "start", "end", "beta", "coverage")
lapply(bedfiles, function(i) knitr::kable(read.table(i, col.names = colnames)))

# make a vector of regions
regions <- c("chr1:1-6", "chr1:7-10", "chr1:11-14")
# make matrix of beta values
make_mat(bedfiles, regions, column = 4)
```

`make_mat_bsseq` *Make M/beta and coverage matrices from WGBS BED files*

Description

Queries the CpG/CpH loci from provided regions and produces M/beta and coverage matrices with their genomic positions. Parallelized across files using threads from the `"iscream.threads"` option. The output of `make_mat_bsseq` may be used to create a BSseq object: `do.call(BSseq, make_mat_bsseq(...))`.

Usage

```
make_mat_bsseq(
  bedfiles,
  regions,
  aligner = "biscuit",
  mval = TRUE,
  merged = TRUE,
  sparse = FALSE,
  prealloc = 10000,
  nthreads = NULL
)
```

Arguments

bedfiles	A vector of BED file paths
regions	A vector, data frame or GenomicRanges of genomic regions. See details.
aligner	The aligner used to produce the BED files - one of "biscuit", "bismark", "bsbolt".
mval	Whether to return M-values or beta-values with the coverage matrix. Defaults to M-value. Set mval=FALSE to get beta value matrix.
merged	Whether the input strands have been merged/collapsed
sparse	Whether to return a sparse matrix
prealloc	The number of rows to initialize the matrices with. If the number of loci are approximately known, this can reduce runtime as fewer resizes need to be made.
nthreads	Set the number of threads to use. Overrides the "iscream.threads" option. See ?set_threads for more information.

Details

The input regions may be string vector in the form "chr:start-end" or a GRanges object. If a data frame is provided, they must have "chr", "start", and "end" columns.

Value

A named list of

- coverage and either a beta- or M-value matrix
- a character vector of chromosomes and numeric vector of corresponding CpG base positions
- a character vector of the input sample names

Bitpacking limits

make_mat_bsseq() makes two matrices: M-value (or beta-value) and coverage. For speed and memory efficiency these two values are bitpacked during matrix creation so that only one matrix needs to be populated and resized. This matrix is unpacked into the two required matrices only after the matrix dimensions are known after querying all input files. The two values are packed using the INT16 type, which has an upper limit of 32,767, into one INT32. If the coverage values exceed 32,767, the upper limit of a 16-bit signed integer, it will be capped at the limit. Beta values will also be capped similarly, but any such beta values would indicate a bug in the aligner that produced the data.

Examples

```
bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)
# examine the BED files
colnames <- c("chr", "start", "end", "beta", "coverage")
lapply(bedfiles, function(i) knitr::kable(read.table(i, col.names = colnames)))

# make a vector of regions
regions <- c("chr1:1-6", "chr1:7-10", "chr1:11-14")
```

```
mat <- make_mat_bsseq(bedfiles, regions)
# for BSseq object run
if (requireNamespace("bsseq", quietly = TRUE)) {
  do.call(bsseq::BSseq, mat)
}
```

---

query_chroms *Query the chromosomes or seqnames from a vector of BED files*

Description

Query the chromosomes or seqnames from a vector of BED files

Usage

```
query_chroms(bedfiles, nthreads = NULL)
```

Arguments

bedfiles	The vector of BED file paths
nthreads	Set number of threads to use overriding the "iscream.threads" option. See ?set_threads for more information.

Value

A vector of seqnames

Examples

```
bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)
query_chroms(bedfiles)
```

---

set_log_level *Set and get logging level*

Description

Set and get logging level

Usage

```
set_log_level(level = "info")

get_log_level()
```

Arguments

`level`	The logging verbosity level to use • `"info"`: the default that gives provides basic information about the number of files and regions used in a function • `"debug"`: more verbose about row allocations, how many CpGs were found in a region, filename parsing etc. This mode cannot be used on more than one thread as R cannot output messages from multiple threads without crashing. • `"off"`: no logging

Value

- `set_log_level()`: None; sets the log level to the provided level
- `get_log_level()`: The current logging level as a string

Examples

```
set_log_level("info")
get_log_level()
```

`set_threads` *Set the number of available threads*

Description

Sets the `"iscream.threads"` option to `n_threads`. To see how many threads you have available see `?get_threads()`.

Usage

```
set_threads(n_threads)
```

Arguments

`n_threads`	The number of threads to use

Details

iscream uses OpenMP to parallelize certain functions. You can use as many threads as are available to you on your system to varying degrees of performance improvements. The `get_threads()` function uses `parallelly::availableCores()` to report the number of available threads. Although OpenMP can detect the number of available cores, on high performance computers (HPCs) with resource allocating job schedulers like SLURM, OpenMP may detect all available threads across the HPC and not limit itself to the cores that were allocated to you by the scheduler. If your system administrator has not set up any limits, this may result in your job taking resources from other jobs. If there are limits, trying to use more threads that those available will reduce iscream's performance. Job schedulers will typically have an environment variable (e.g. `SLURM_CPUS_ON_NODE` with

SLURM) that gives you the actual number of available cores. Further, on hyperthreaded systems, this count may be double that of the available processors. Using hyperthreading does not guarantee any performance improvement - it may be better to set the number of threads to half the reported number. `parallelly::availableCores()` takes HPC scheduler/CRAN/Bioconductor limits into account when reporting the number of available threads but it may not reliably report hyperthreading ('system' or 'nproc'). To set the number of threads without having to call `set_threads()` in every session, put

```
options(iscream.threads = [n_threads])
```

in your `.Rprofile` See `help('Rprofile')` for information on startup options.

Functions currently using multithreading:

- `tabix()`, `tabix_gr()`, `tabix_raw()`
- `query_chroms()`
- `make_mat()`, `make_mat_se()`, `make_mat_gr()`, `make_mat_bsseq()`
- `summarize_regions()`, `summarize_meth_regions()`

Value

None. Sets the `iscream.threads` option to the requested number of threads if available

Examples

```
(ncores <- parallelly::availableCores())

set_threads(ncores)
```

---

summarize_meth_regions

Summarize methylation information over genomic regions

Description

Run summarizing functions on the CpG/CpH loci in BED files across genomic regions. Parallelized across files using threads from the `"iscream.threads"` option.

Usage

```
summarize_meth_regions(
  bedfiles,
  regions,
  fun = "all",
  aligner = "biscuit",
  feature_col = NULL,
  mval = TRUE,
  nthreads = NULL
)
```

Arguments

`bedfiles`	A vector of BED file paths
`regions`	A vector, data frame or GenomicRanges of genomic regions. See details.
`fun`	Function(s) to apply over the region. See details.
`aligner`	The aligner used to produce the BED files - one of "biscuit", "bismark", "bsbolt".
`feature_col`	Column name of the input `regions` data frame or GRanges `mcols` containing a name for each genomic region. Set only if the using a data frame-like object as the input regions format, not for string vectors. See details.
`mval`	Whether to calculate the M value (coverage $\times\beta$) or use the beta value when applying the function.
`nthreads`	Set number of threads to use overriding the `"iscream.threads"` option. See `?set_threads` for more information.

Value

A data.table

Supported functions

- Sum: `"sum"`
- Mean: `"mean"`
- Median: `"median"`
- Mode: `"mode"`
- Anti-mode: `"antimode"`
- Sample standard deviation: `"stddev"` (`sstdev` in `bedtools map`)
- Population standard deviation: `"pstddev"` (`stdev` in `bedtools map`)
- Variance: `"variance"`
- Minimum: `"min"`
- Maximum: `"max"`
- Minimum of absolute values: `"absmin"`
- Maximum of absolute values: `"absmax"`
- Range: `"range"`
- First element: `"first"`
- Last element: `"last"`
- No. of records in the region: `"count"`
- No. of records in the region with unique data values: `"count_distinct"`

Most summarizing computations are backed by the Armadillo library. See https://arma.sourceforge.net/docs.html#stats_fns for futher details on the supported functions

Using feature identifiers

regions may be a string vector in the form "chr:start-end", a GRanges object or a data frame with "chr", "start", and "end" columns. If the input data.frame or GRanges has a column with feature identifiers, like gene names for a set of gene regions, pass that column's name to feature_col. If regions is a vector, set its names() to those identifiers. These will be used to populate a 'feature' column in the summary. See examples.

Examples

```
# also see examples from ?summarize_regions

bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)

# make a vector of regions
regions <- c("chr1:1-6", "chr1:7-10", "chr1:11-14")
summarize_meth_regions(bedfiles, regions)

# add names to the regions to populate the 'feature' column
names(regions) <- c("gene1", "gene2", "gene3")
summarize_meth_regions(bedfiles, regions, fun = c("mean", "stddev"), mval = FALSE)
summarize_meth_regions(bedfiles, regions, fun = "sum")
```

---

summarize_regions *Summarize information over genomic regions from any BED file*

Description

Run summarizing functions on BED file records across genomic regions. Parallelized across files using threads from the "iscream.threads" option.

Usage

```
summarize_regions(
  bedfiles,
  regions,
  columns,
  col_names = NULL,
  fun = "all",
  feature_col = NULL,
  nthreads = NULL
)
```

Arguments

bedfiles	A vector of BED file paths
regions	A vector, data frame or GenomicRanges of genomic regions. See details.

columns	A vector of indices of the numeric columns to be summarized
col_names	A vector of names to use for columns in the output
fun	Function(s) to apply over the region. See details.
feature_col	Column name of the input regions data frame or GRanges mcols containing a name for each genomic region. Set only if the using a data frame-like object as the input regions format, not for string vectors. See details.
nthreads	Set number of threads to use overriding the "iscream.threads" option. See ?set_threads for more information.

Value

A data.table

Supported functions

- Sum: "sum"
- Mean: "mean"
- Median: "median"
- Mode: "mode"
- Anti-mode: "antimode"
- Sample standard deviation: "stddev" (sstdev in bedtools map)
- Population standard deviation: "pstddev" (stdev in bedtools map)
- Variance: "variance"
- Minimum: "min"
- Maximum: "max"
- Minimum of absolute values: "absmin"
- Maximum of absolute values: "absmax"
- Range: "range"
- First element: "first"
- Last element: "last"
- No. of records in the region: "count"
- No. of records in the region with unique data values: "count_distinct"

Most summarizing computations are backed by the Armadillo library. See https://arma.sourceforge.net/docs.html#stats_fns for futher details on the supported functions

Using feature identifiers

regions may be a string vector in the form "chr:start-end", a GRanges object or a data frame with "chr", "start", and "end" columns. If the input data.frame or GRanges has a column with feature identifiers, like gene names for a set of gene regions, pass that column's name to feature_col. If regions is a vector, set its names() to those identifiers. These will be used to populate a 'feature' column in the summary. See examples.

Examples

```
bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)
# examine the bedfiles
colnames <- c("chr", "start", "end", "beta", "coverage")
lapply(bedfiles, function(i) knitr::kable(read.table(i, col.names = colnames)))

# make a vector of regions
regions <- c("chr1:1-6", "chr1:7-10", "chr1:11-14")
summarize_regions(bedfiles, regions, columns = c(4, 5), col_names = c("beta", "cov"))

# select functions
summarize_regions(
  bedfiles,
  regions,
  fun = c("mean", "stddev"),
  columns = c(4, 5),
  col_names = c("beta", "cov")
)

# add names to the regions
names(regions) <- c("gene1", "gene2", "gene3")
summarize_regions(
  bedfiles,
  regions,
  fun = "sum",
  columns = 5,
  col_names = "coverage"
)

# using `feature_col`
library(data.table)

# convert string vector to a data.table
regions_df <- data.table::as.data.table(regions) |>
_[, tstrsplit(regions, ":|-", fixed = FALSE, names = c("chr", "start", "end"))] |>
_[, feature := names(regions)][]
regions_df

summarize_regions(
  bedfiles,
  regions_df,
  fun = "sum",
  columns = 5,
  col_names = "coverage",
  feature_col = "feature"
)
```

---

tabix *Query records from tabixed BED files*

Description

Query records from tabixed BED files

Usage

```
tabix(bedfiles, regions, aligner = NULL, col.names = NULL, nthreads = NULL)

tabix_gr(
  bedfiles,
  regions,
  aligner = NULL,
  col.names = NULL,
  zero_based = TRUE,
  nthreads = NULL
)

tabix_raw(bedfiles, regions, nthreads = NULL)
```

Arguments

bedfiles	The BED files to be queried
regions	A vector, data frame or GenomicRanges of genomic regions. See details.
aligner	The bisulfite aligner used to produce the BED files - one of "biscuit", "bismark", "bsbolt". Will set the result data.table's column names based on this argument.
col.names	A vector of column names for the data columns of the result, not including "chr", "start", and "end". Set for non-WGBS BED files or WGBS BED files not from the supported aligners. If the BED files have a strand column, use "strand" and `tabix_gr` will use it in the output GRanges object.
nthreads	Set number of threads to use overriding the `"iscream.threads"` option. See `?set_threads` for more information.
zero_based	Whether the input BED file has a zero-based start column - used when coverting the result data frame to GenomicRanges.

Details

Query method:

'*iscream* has two methods to query records from BED files:

- the *tabix* shell executable: fast since its output can be redirected to a file (which `data.table::fread()` can then read) instead of having to allocate memory and store it during the query
- *iscream's* tabix implementation, based on the *tabix* executable using *htslib*, but slower on large queries since it stores the records as they are found instead of writing to a file. However it's able to store each region's records independently instead of in a single file and is used in `make_mat()`, `make_mat_bsseq()`, and `summarize_regions()`.

When *iscream* is attached, it checks that the *tabix* executable is available with `Sys.which()` and, if available, sets `options("tabix.method" = "shell")`. `tabix()` then uses the *tabix* executable to make queries, except for `tabix_raw()`. If *tabix* is not found, *iscream* uses its tabix implementation. To use only *iscream's* tabix implementation, set `options("tabix.method" = "htslib")`.

Input region formats:

The input regions format may be string vector in the form "chr:start-end", a dataframe with "chr", "start" and "end" columns or a `GRanges` object. Input regions must be 1-based. When using `"htslib"` as the query method, if the input `GRanges` object of regions contains any single locus regions where the start and end positions are the same, iscream will notify that such regions were found and fixed as `chr:start` format strings are invalid for the htslib API (see `?get_granges_string`).

Value

- `tabix()`: A data frame
- `tabix_gr()`: A `GRanges` object for single files and `GRangesList` for multiple files. When making `GRanges`, the 0-based records from BED-files will be converted to 1-based with `GenomicRanges::makeGRangesFromDataFrame()`. Bismark's coverage files will not be converted as they are already 1-based and the `ranges` slot will be only one position.
- `tabix_raw()`: A named list of raw strings from the regions in the style of `Rsamtools::scanTabix`

Examples

```
bedfiles <- system.file("extdata", package = "iscream") |>
  list.files(pattern = "[a|b|c|d].bed.gz$", full.names = TRUE)
regions <- c("chr1:1-6", "chr1:7-10", "chr1:11-14")
tabix(bedfiles, regions, col.names = c("beta", "coverage"))
if (require("GenomicRanges", quietly = TRUE)) {
  tabix_gr(bedfiles, regions, col.names = c("beta", "coverage"))
}
tabix_raw(bedfiles, regions)
```

Index