

## > {epistack}

An R/Bioconductor package to visualise stack profiles of epigenomic signals

Safia Saci & Guillaume Devailly

R-Toulouse



# Bioconductor [www.bioconductor.org](http://www.bioconductor.org)

- An alternative R package repository
- Dedicated to bioinformatics, the data science of biology.
- More opinionated than CRAN:
  - package reviewing
  - BiocCheck()

The screenshot shows the Bioconductor website homepage. At the top left is the Bioconductor logo with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right is a search bar and a navigation menu with links for Home, Install, Help, Developers, and About. The main content area is divided into several sections: "About Bioconductor" with a mission statement, "News" with a list of recent updates, "Install" with links to get started, "Learn" with links to resources, "Use" with links to create solutions, and "Develop" with links to contribute. The "News" section includes items like "Bioconductor Bioc 3.14 Released" and "Bioconductor browsable code base now available".



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# Bioconductor

- Release cycle every 6 months
- Everything should work in *release*
- Some things may be broken in *devel*
- Package are tested, vignettes are built, nightly
- Install packages using:

```
BiocManager::install("packagename")
```

- A bioconductor release will work only (mostly) with the latest version of R

[Home](#) » [Bioconductor 3.14](#) » [Software Packages](#) » [epistack](#)

## epistack



DOI: [10.18129/B9.bioc.epistack](https://doi.org/10.18129/B9.bioc.epistack)



[Home](#) » [Bioconductor 3.15](#) » [Software Packages](#) » [epistack \(development version\)](#)

## epistack



DOI: [10.18129/B9.bioc.epistack](https://doi.org/10.18129/B9.bioc.epistack)



This is the **development** version of epistack; for the stable release version, see [epistack](#).



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# Bioconductor in numbers

Release 3.14 (for R 4.1.0)

- 2.083 software packages
- 408 data packages
- 904 annotation packages

{DESeq2} in 2020: 370.000 downloads from 124.000 distinct IPs.

[Home](#) » [Bioconductor 3.14](#) » [Software Packages](#) » DESeq2

## DESeq2



DOI: [10.18129/B9.bioc.DESeq2](https://doi.org/10.18129/B9.bioc.DESeq2)  

Differential gene expression analysis based on the negative binomial distribution



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# Bioconductor strength 1

A core set of methods / parser / classed to work with biological data

## Common Bioconductor Methods and Classes

---

We strongly recommend reusing existing methods for importing data, and reusing established classes for representing data. Here are some suggestions for importing different file types and commonly used *Bioconductor* classes. For more classes and functionality also try searching in [BiocViews](#) for your data type.

### Importing

---

- GTF, GFF, BED, BigWig, etc., – [rtracklayer::import\(\)](#)
- VCF – [VariantAnnotation::readVcf\(\)](#)
- SAM / BAM – [Rsamtools::scanBam\(\)](#), [GenomicAlignments::readGAlignment\\*\(\)](#)
- FASTA – [Biostrings::readDNASTringSet\(\)](#)
- FASTQ – [ShortRead::readFastq\(\)](#)
- MS data (XML-based and mgf formats) – [Spectra::Spectra\(\)](#), [MSnbase::readMSData\(\)](#),  
[Spectra::Spectra\(source = MsBackendMgf::MsBackendMgf\(\)\)](#), [MSnbase::readMgfData\(\)](#)

### Common Classes

---

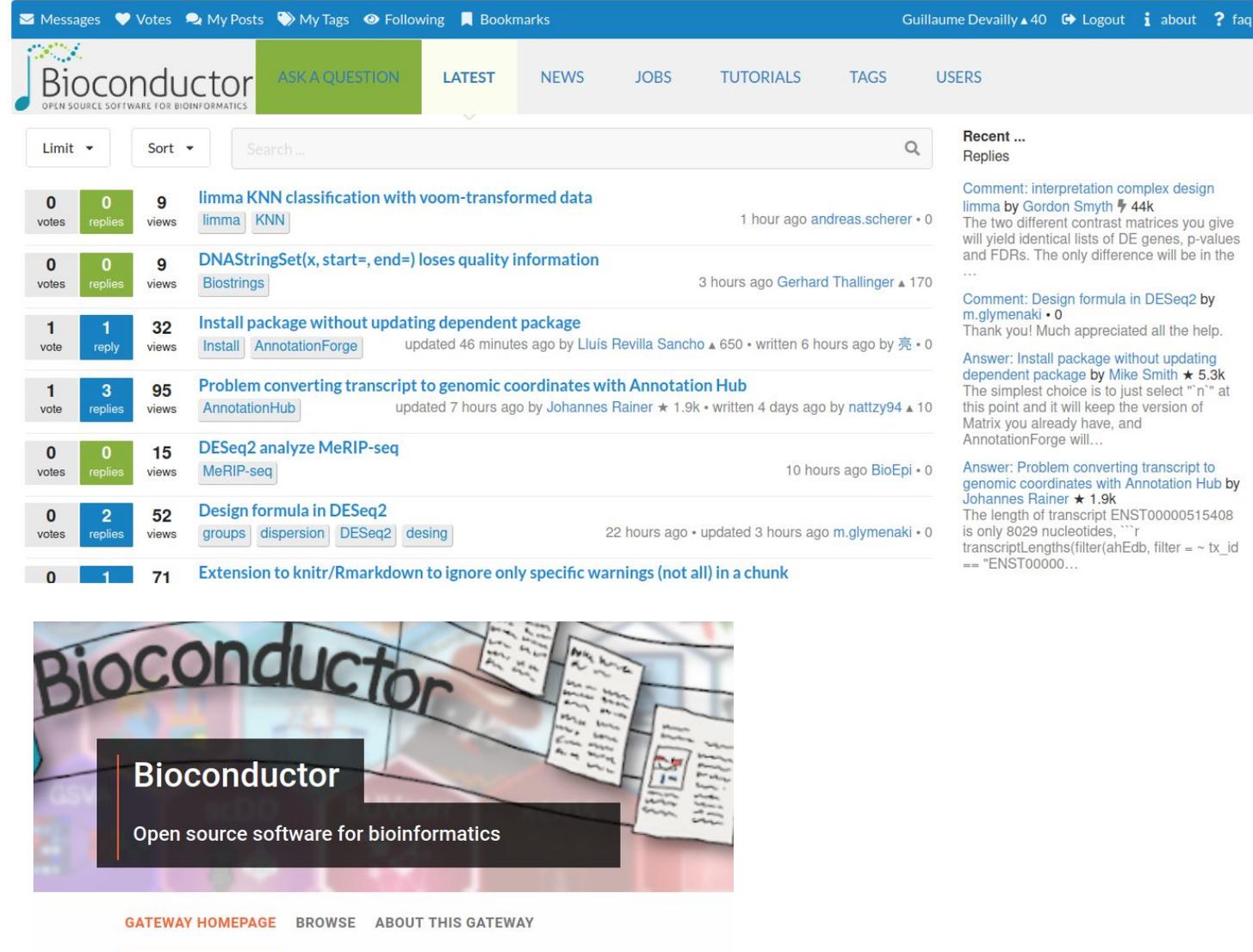
- Rectangular feature x sample data – [SummarizedExperiment::SummarizedExperiment\(\)](#) (RNAseq count matrix, microarray, ...)
- Genomic coordinates – [GenomicRanges::GRanges\(\)](#) (1-based, closed interval)
- Genomic coordinates from multiple samples – [GenomicRanges::GRangesList\(\)](#)
- Ragged genomic coordinates – [RaggedExperiment::RaggedExperiment\(\)](#)
- DNA / RNA / AA sequences – [Biostrings::\\*StringSet\(\)](#)
- Gene sets – [BiocSet::BiocSet\(\)](#), [GSEABase::GeneSet\(\)](#), [GSEABase::GeneSetCollection\(\)](#)
- Multi-omics data – [MultiAssayExperiment::MultiAssayExperiment\(\)](#)
- Single cell data – [SingleCellExperiment::SingleCellExperiment\(\)](#)
- Mass spec data – [Spectra::Spectra\(\)](#), [MSnbase::MSnExp\(\)](#)



# Bioconductor strength 2

## Community:

- forum *à la* stackoverflow
- conferences & events
- F1000 Research gateway



The screenshot shows the Bioconductor forum interface. At the top, there are navigation links for Messages, Votes, My Posts, My Tags, Following, and Bookmarks. The main header includes the Bioconductor logo and navigation options: ASK A QUESTION, LATEST, NEWS, JOBS, TUTORIALS, TAGS, and USERS. Below the header, there are filters for Limit and Sort, and a search bar. The main content area displays a list of recent forum posts, each with a title, a brief description, and statistics (votes, replies, views). The posts include:

- limma KNN classification with voom-transformed data** (1 hour ago, 0 votes, 0 replies, 9 views)
- DNASTringSet(x, start=, end=) loses quality information** (3 hours ago, 0 votes, 0 replies, 9 views)
- Install package without updating dependent package** (updated 46 minutes ago, 1 vote, 1 reply, 32 views)
- Problem converting transcript to genomic coordinates with Annotation Hub** (updated 7 hours ago, 1 vote, 3 replies, 95 views)
- DESeq2 analyze MeRIP-seq** (10 hours ago, 0 votes, 0 replies, 15 views)
- Design formula in DESeq2** (22 hours ago, 0 votes, 2 replies, 52 views)
- Extension to knitr/Rmarkdown to ignore only specific warnings (not all) in a chunk** (0 votes, 1 reply, 71 views)

On the right side, there is a 'Recent ... Replies' section with comments and answers related to the forum posts.



[GATEWAY HOMEPAGE](#) [BROWSE](#) [ABOUT THIS GATEWAY](#)

This gateway highlights Bioconductor package-based vignettes and cross-package workflows.



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# How to submit a Bioconductor package?

Raise an issue at [github.com/Bioconductor/Contributions](https://github.com/Bioconductor/Contributions)  
with a link to a GitHub repo of your package



# How to submit a Bioconductor package?

Bioconductor has expectations for your package:

- Mandatory vignette
- Use Bioconductor recommended classes and methods
- 0 error, 0 warning, 0 note in R CMD check 🤖
- 0 error, 0 warning, minimal number of notes in BiocCheck::BiocCheck() 🤖
  - including mandatory subscription to a daily mailing list storing your password in clear 😞

```
$error
character(0)

$warning
[1] "y of x.y.z version should be even in release"

$note
[1] "Consider adding these automatically suggested biocViews: ChipOnChip"

[2] "The Description field in the DESCRIPTION is made up by less than 3 sentences. Please consider expanding this field, and\nstructure it as a full paragraph"
[3] "Recommended function length ≤ 50 lines."

[4] "Consider shorter lines; 31 lines (1%) are > 80 characters long."

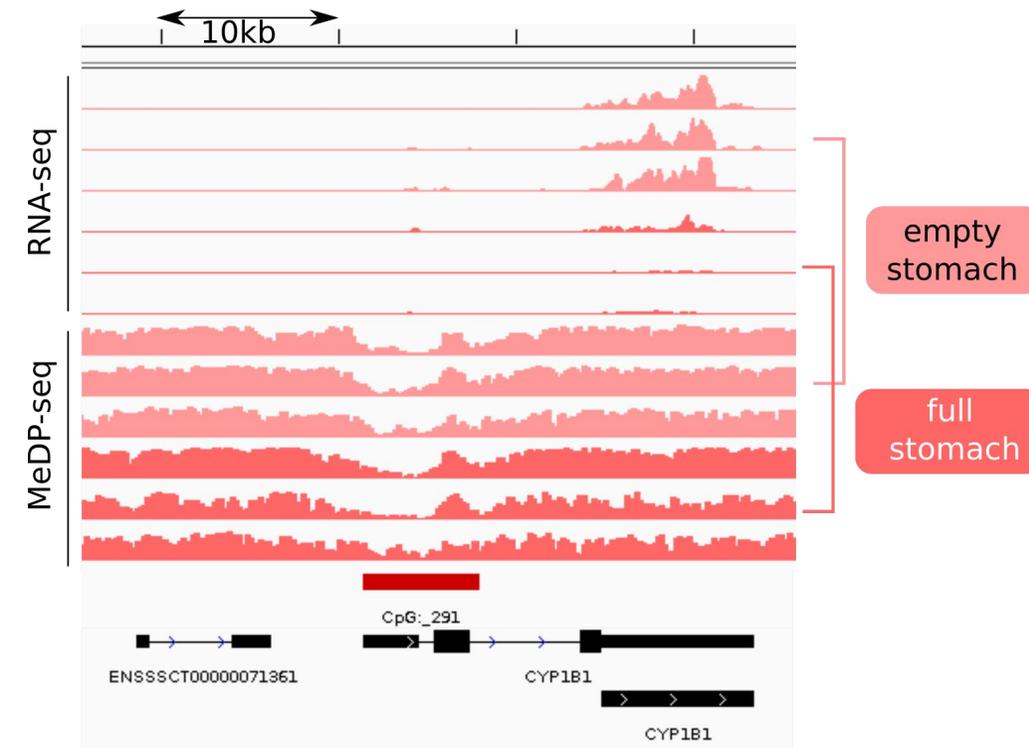
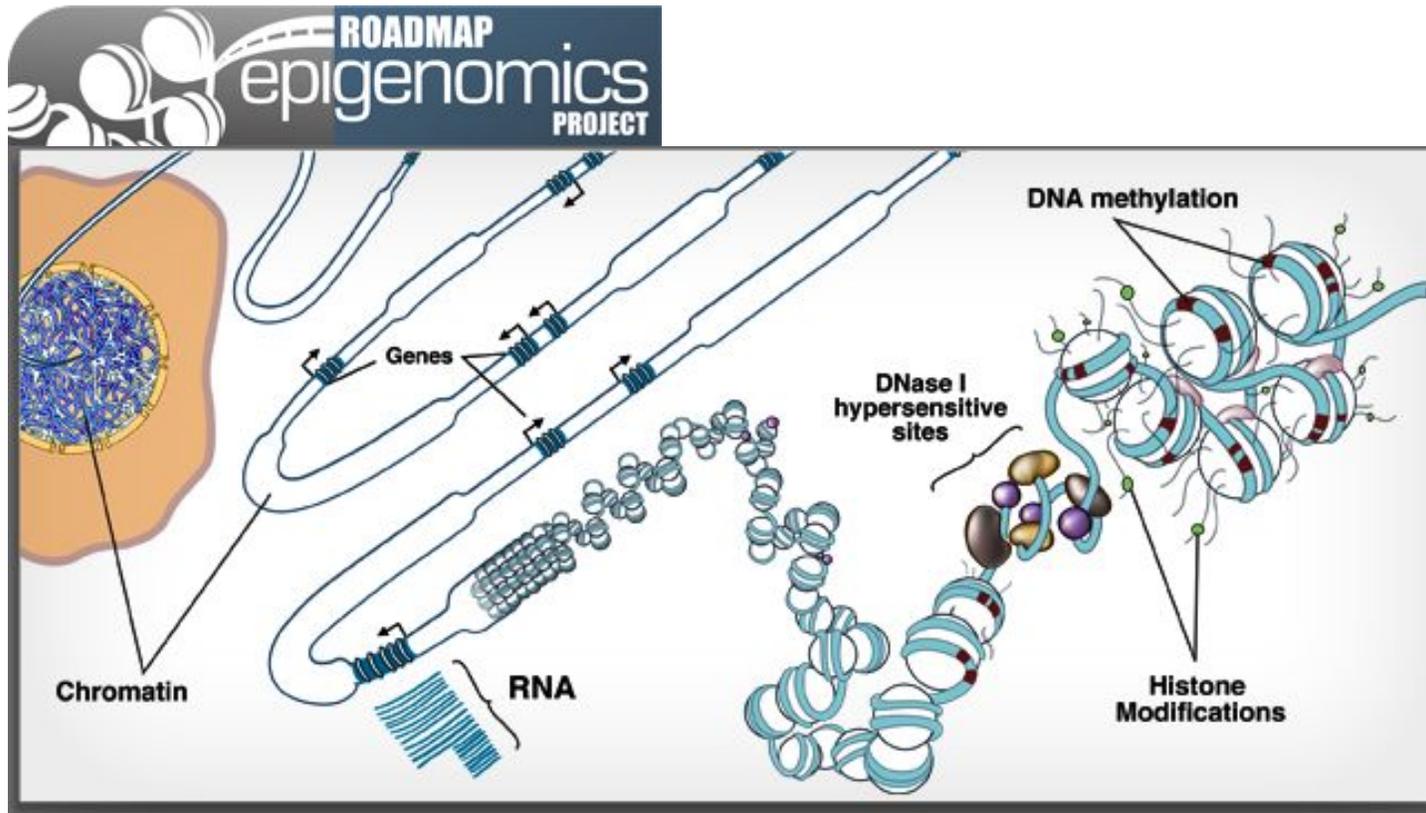
[5] "Consider multiples of 4 spaces for line indents, 184 lines(7%) are not."

[6] "Cannot determine whether maintainer is subscribed to the bioc-devel mailing list (requires admin credentials).\nSubscribe here: http://stat.ethz.ch/mailman/listinfo/bioc-devel"
```



# Epigenomic data

Tracks of genomic scores



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

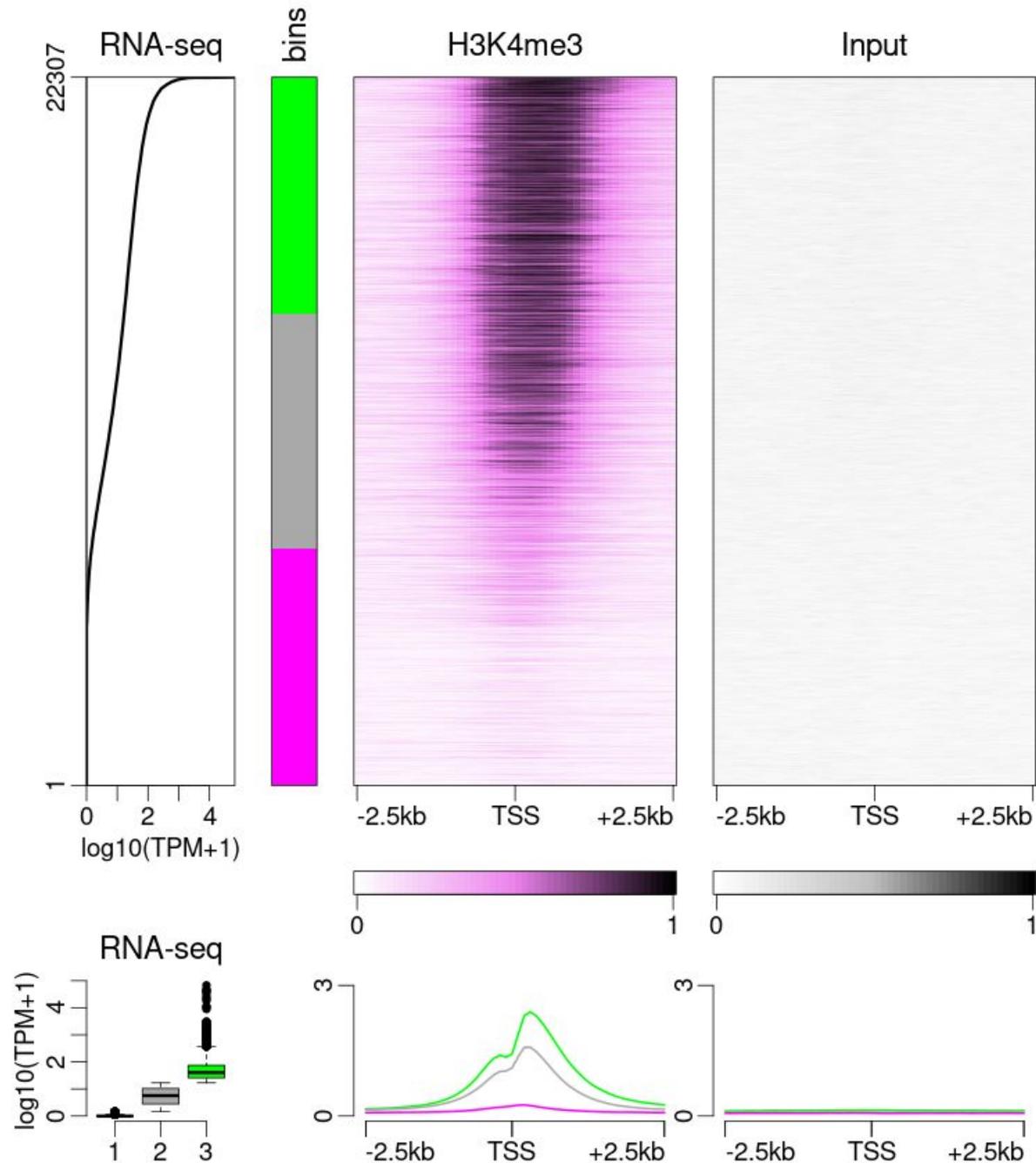
# {epistack} epigenetic stacks

Visualise stacks of epigenetic signals on anchor regions:

- ◇ Gene starts
- ◇ Peak center
- ◇ CpG islands
- ◇ Differentially methylated regions
- ◇ ...

Sort regions:

- ◇ Gene expression levels
- ◇ P-values
- ◇ Fold changes
- ◇ Clustering
- ◇ Region widths
- ◇ Distance to closest TSS
- ◇ Gene types
- ◇ ...



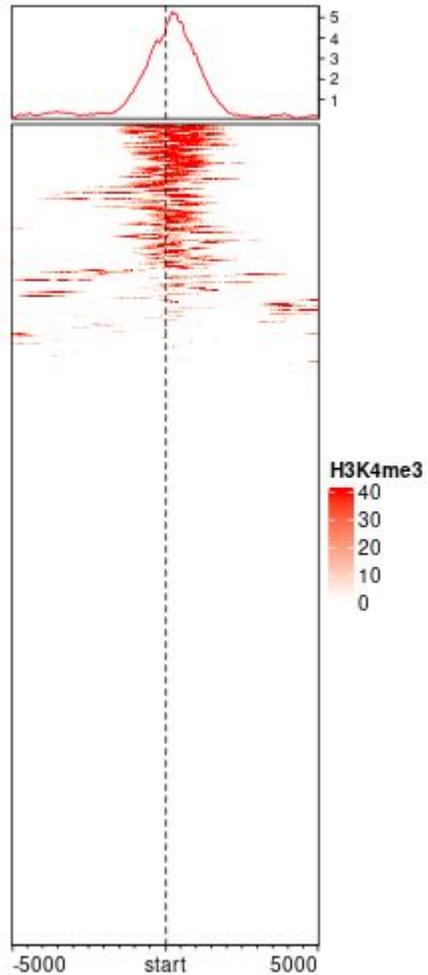
INRAE

{epistack}

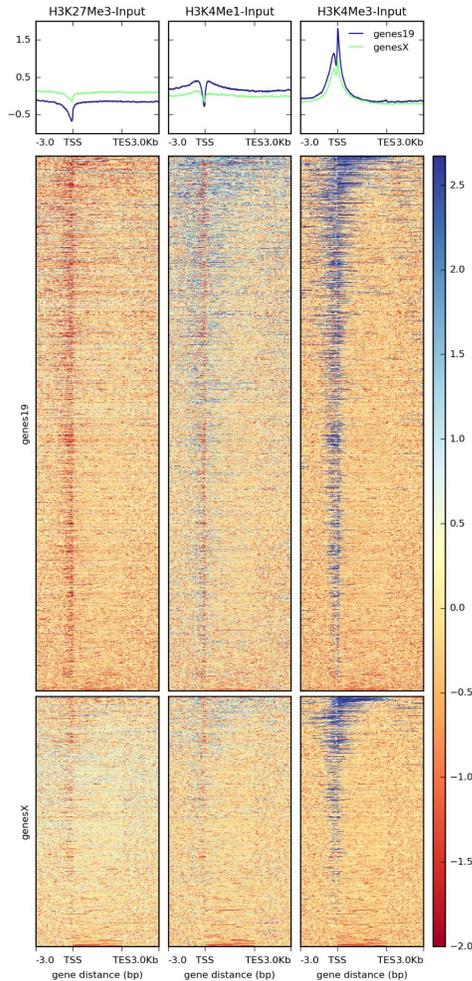
2021-12-10 / R-Toulouse / Guillaume Devailly

# {epistack} alternatives

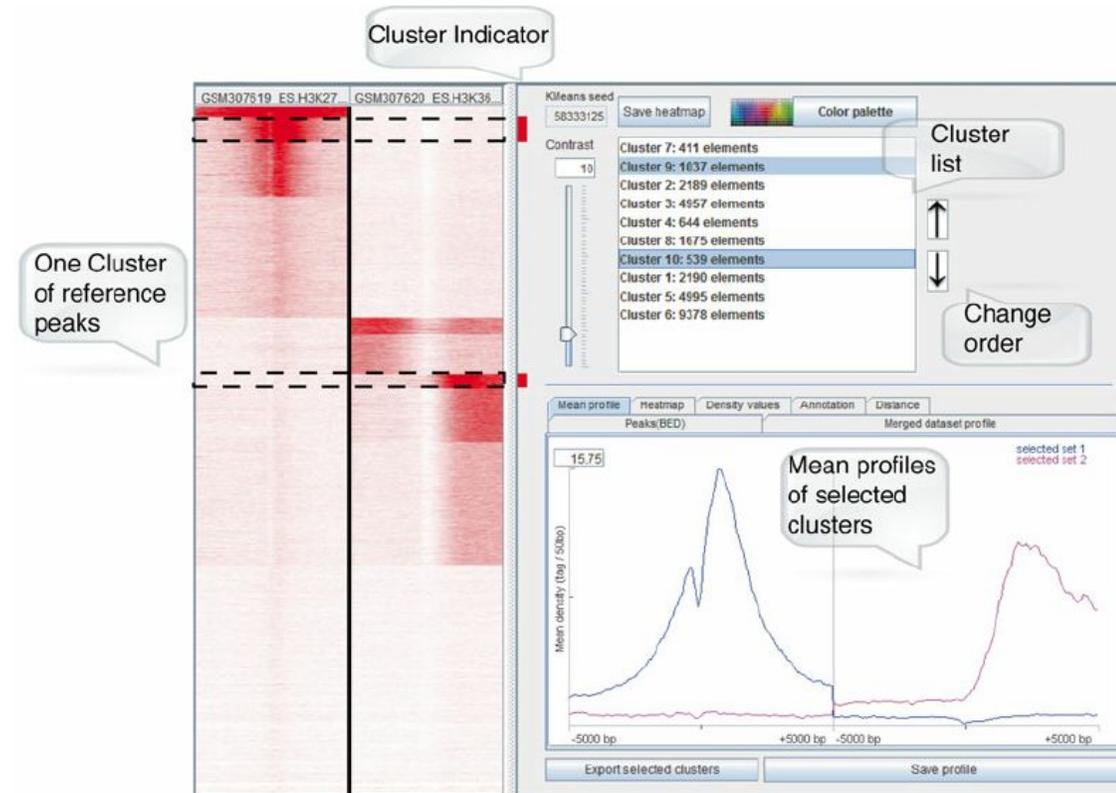
EnrichedHeatmap  
(bioconductor)



deepTools plotHeatmap  
(CLI - python)



seqMINER  
(GUI - java)



And also: seqplots, Repitools, ChIPseeker, ...  
@Bioconductor



# {epistack} is a visualisation package

**Input:** A SummarizedExperiment object, with signal matrices embedded as assays

```
library(SummarizedExperiment)
library(epistack)

data("stackepi")
dim(stackepi)
#> [1] 693 51
stackepi
#> class: RangedSummarizedExperiment
#> dim: 693 51
#> metadata(0):
#> assays(1): DNAME
#> rownames(693): ENSSSCG00000016737 ENSSSCG00000036350 ... ENSSSCG00000024209
#> ENSSSCG00000048227
#> rowData names(3): gene_id exp score
#> colnames(51): window_1 window_2 ... window_50 window_51
#> colData names(0):
```



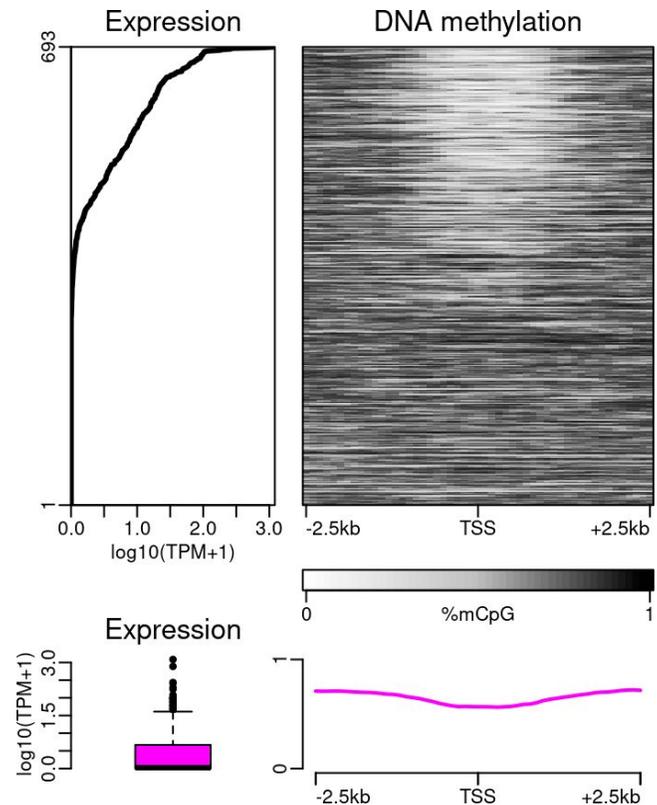
# {epistack} is a visualisation package

**Input:** A SummarizedExperiment object, with signal matrices embedded as assays

```
library(SummarizedExperiment)
library(epistack)

data("stackepi")
dim(stackepi)
#> [1] 693 51
stackepi
#> class: RangedSummarizedExperiment
#> dim: 693 51
#> metadata(0):
#> assays(1): DNAm
#> rownames(693): ENSSSCG00000016737 ENSSSCG00000036350 ... ENSSSCG00000024209
#> ENSSSCG00000048227
#> rowData names(3): gene_id exp score
#> colnames(51): window_1 window_2 ... window_50 window_51
#> colData names(0):
```

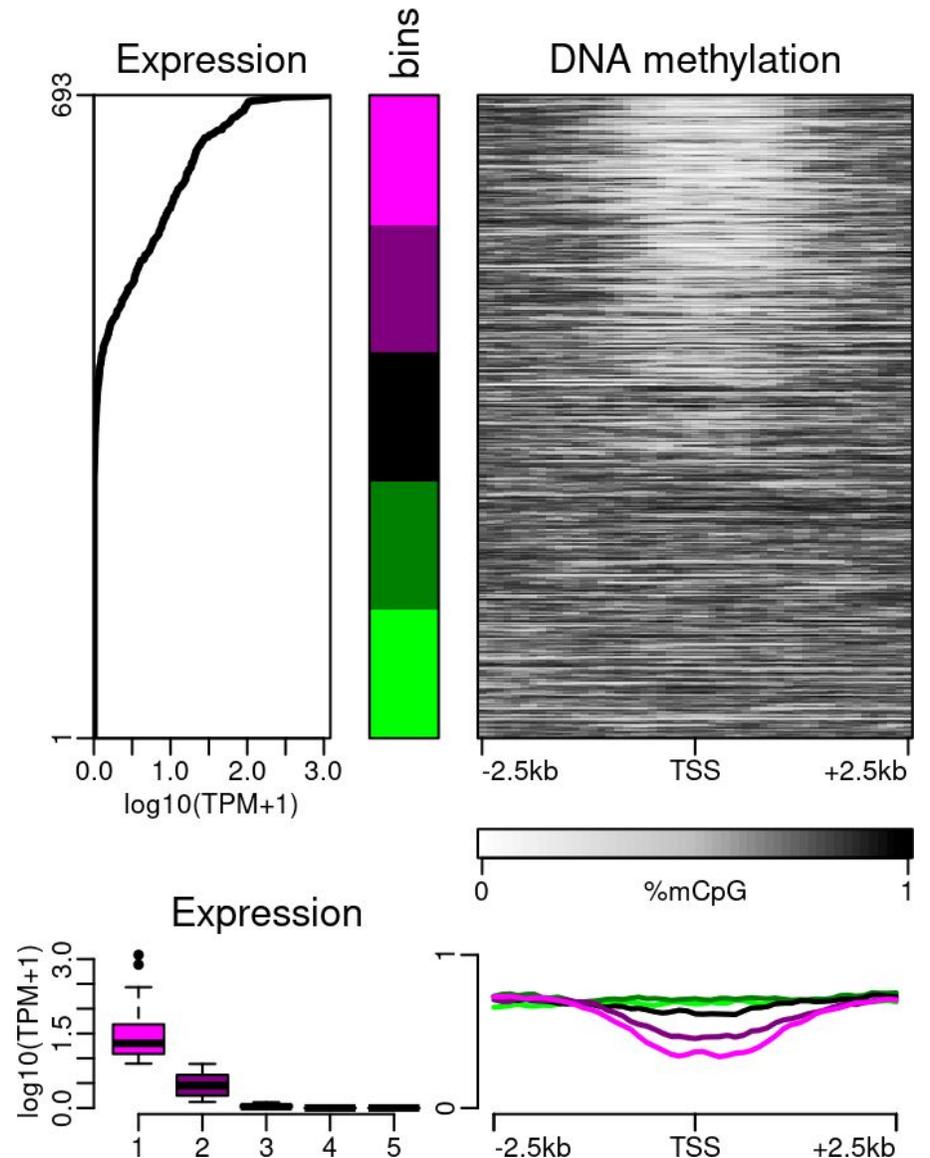
```
plotEpistack(
  stackepi,
  assays = "DNAm", metric_col = "exp",
  ylim = c(0, 1), zlim = c(0, 1),
  x_labels = c("-2.5kb", "TSS", "+2.5kb"),
  titles = "DNA methylation", legends = "%mCpG",
  metric_title = "Expression", metric_label = "log10(TPM+1)",
  metric_transfunc = function(x) log10(x+1)
)
```



# plotEpistack() parameters highlights: bins

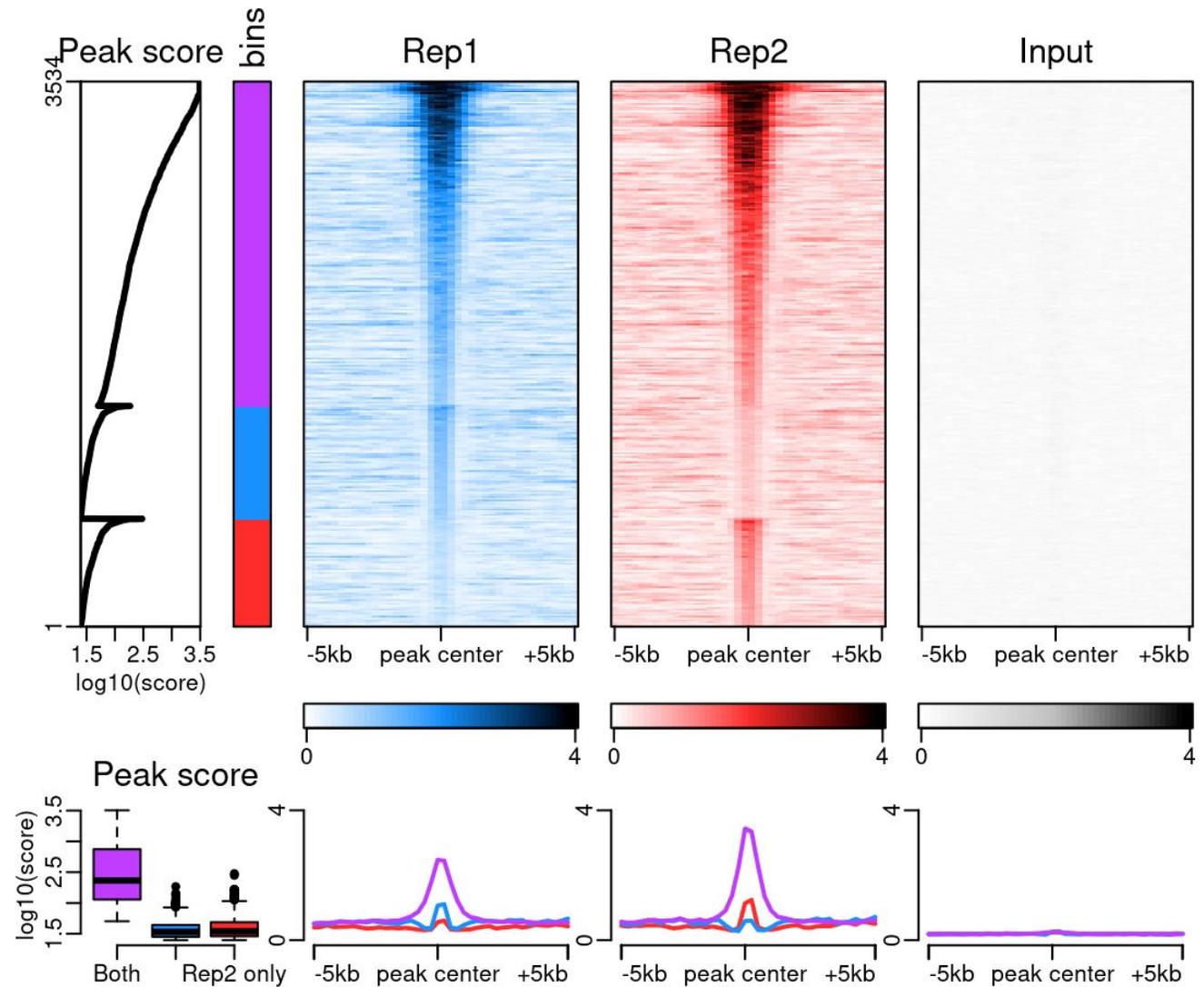
```
stackepi <- addBins(stackepi, nbins = 5)

plotEpistack(
  stackepi,
  assays = "DNAm", metric_col = "exp",
  ylim = c(0, 1), zlim = c(0, 1),
  x_labels = c("-2.5kb", "TSS", "+2.5kb"),
  titles = "DNA methylation", legends = "%mCpG",
  metric_title = "Expression", metric_label = "log10(TPM+1)",
  metric_transfunc = function(x) log10(x+1)
)
```



# plotEpistack() parameters highlights: several tracks

```
plotEpistack(  
  meDP,  
  assays = c("Rep1", "Rep2", "input"),  
  ...  
)
```



# Building epistack's input RangedSummarizedExperiment

- epigenetic tracks:
  - .bam, .bigwig
  - load into R with `GenomicAlignment::readGAlignment()` or `rtracklayer::import()`
- anchors:
  - .bed, .gtf/.gff
  - load into R with `rtracklayer::import()`
- get epigenetic stacks matrices
  - `EnrichedHeatmap::normalizeToMatrix()`
  - `Repitools::annotationCounts()`
  - `ChIPseeker::getTagMatrix()`
- an additional ordering vector
  - gene expression, p-value, fold change, etc.
  - `epitack::addMetricAndArrange*`

# Building epistack's input RangedSummarizedExperiment

```
anchors <- rtracklayer::import( "my_peaks.bed" )
signal <- rtracklayer::import( "my_coverage.bw" )

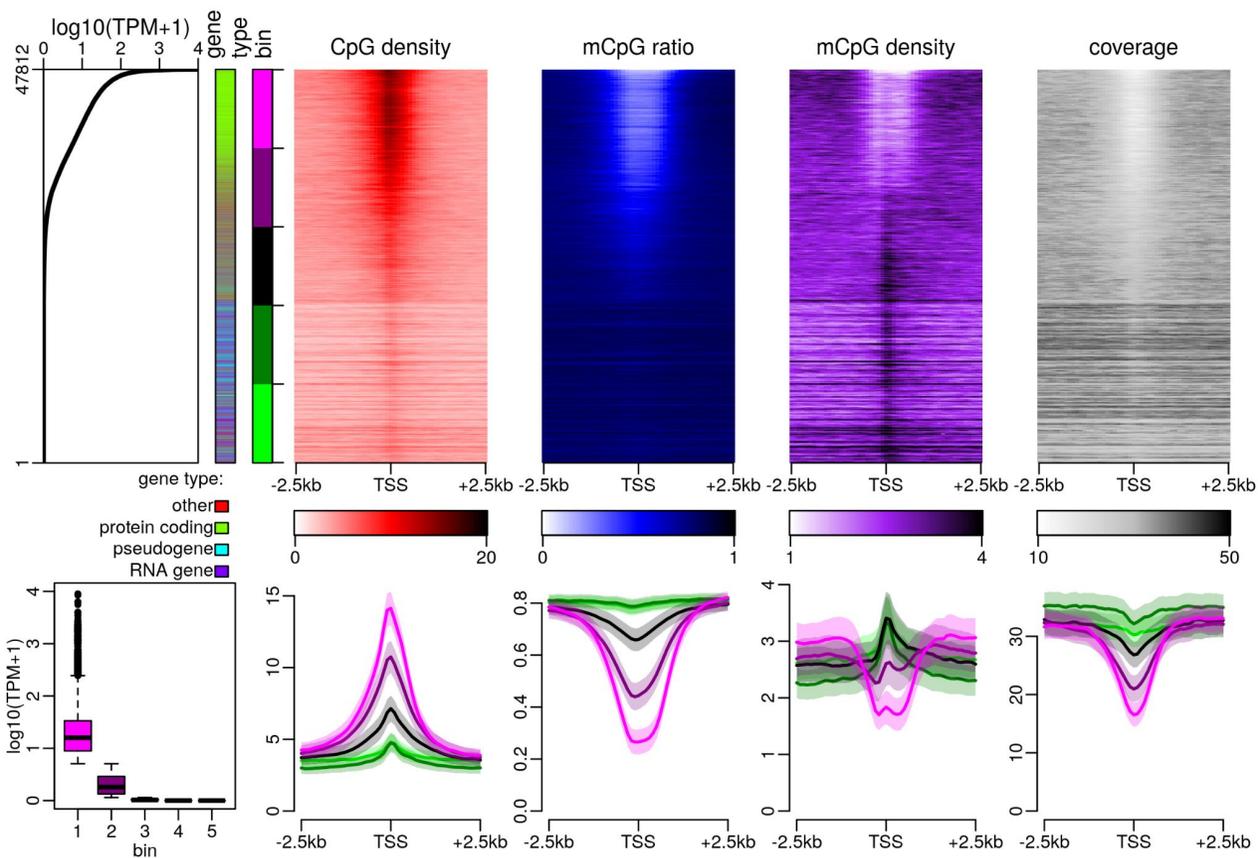
stack <- EnrichedHeatmap::normalizeToMatrix(
  signal,
  anchors,
  extend = 2500, w = 50
)

pack <- SummarizedExperiment(
  rowRanges = anchors,
  assays = list(stack = stack)
)

plotEpistack(
  pack,
  assays = "stack",
  ...
)
```

# Use case: WGBS coverage at TSS

ROADMAP, H1 cell line



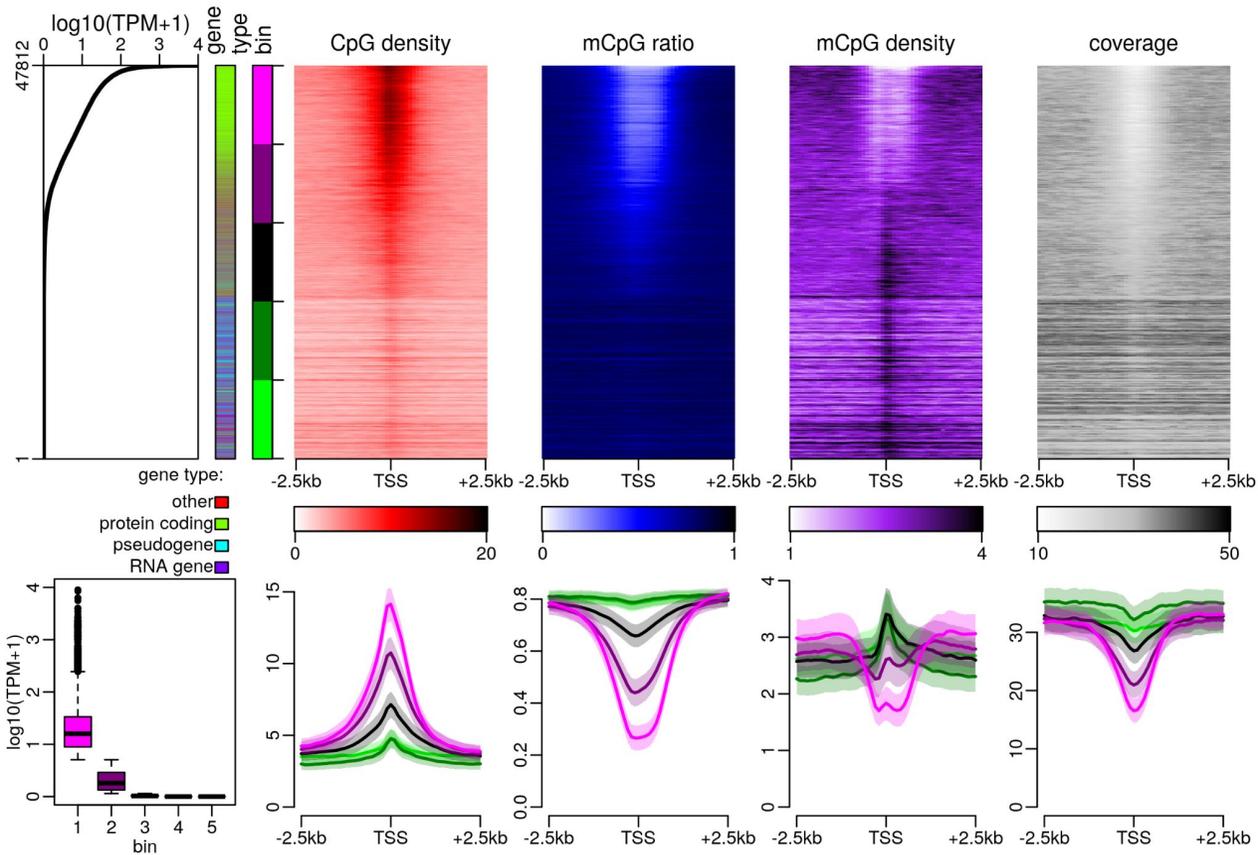
INRAE

{epistack}

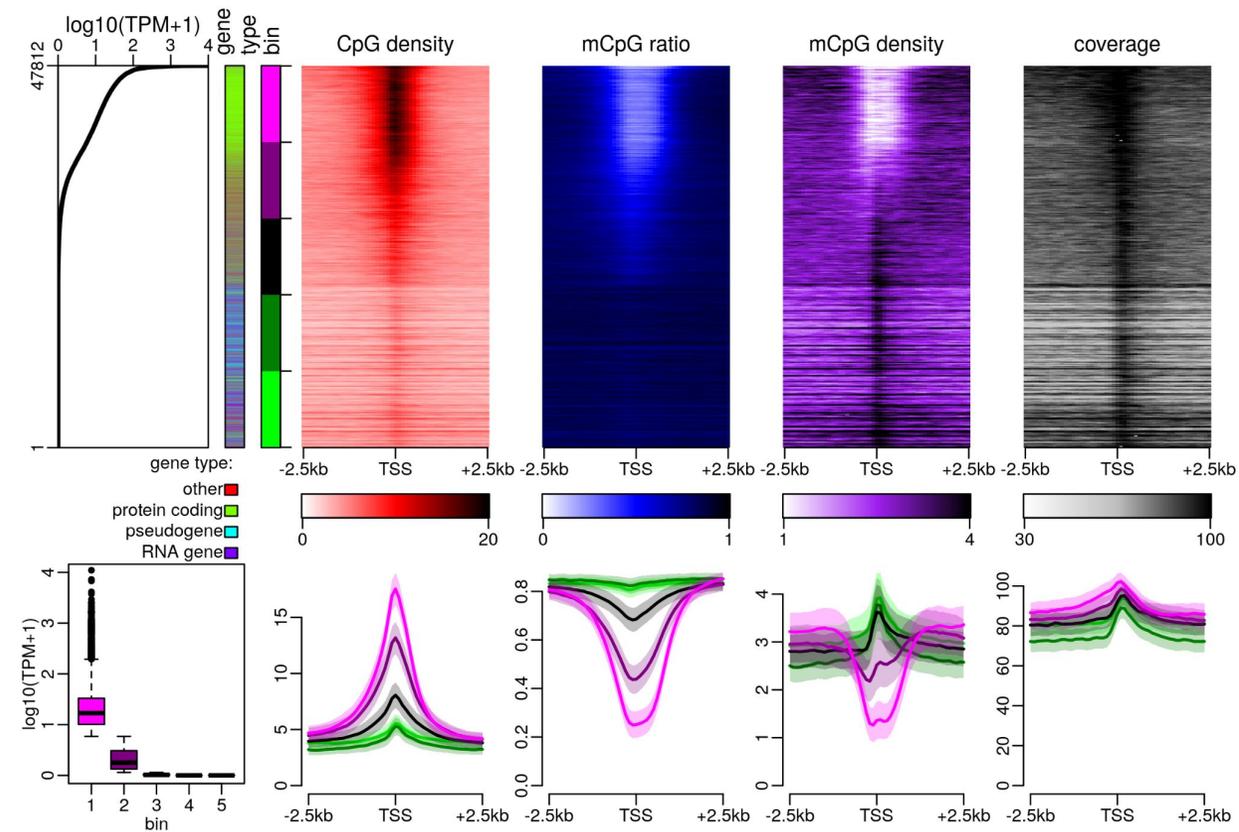
2021-12-10 / R-Toulouse / Guillaume Devailly

# Use case: WGBS coverage at TSS

ROADMAP, H1 cell line



ROADMAP, ESC derived CD56+ Ectoderm



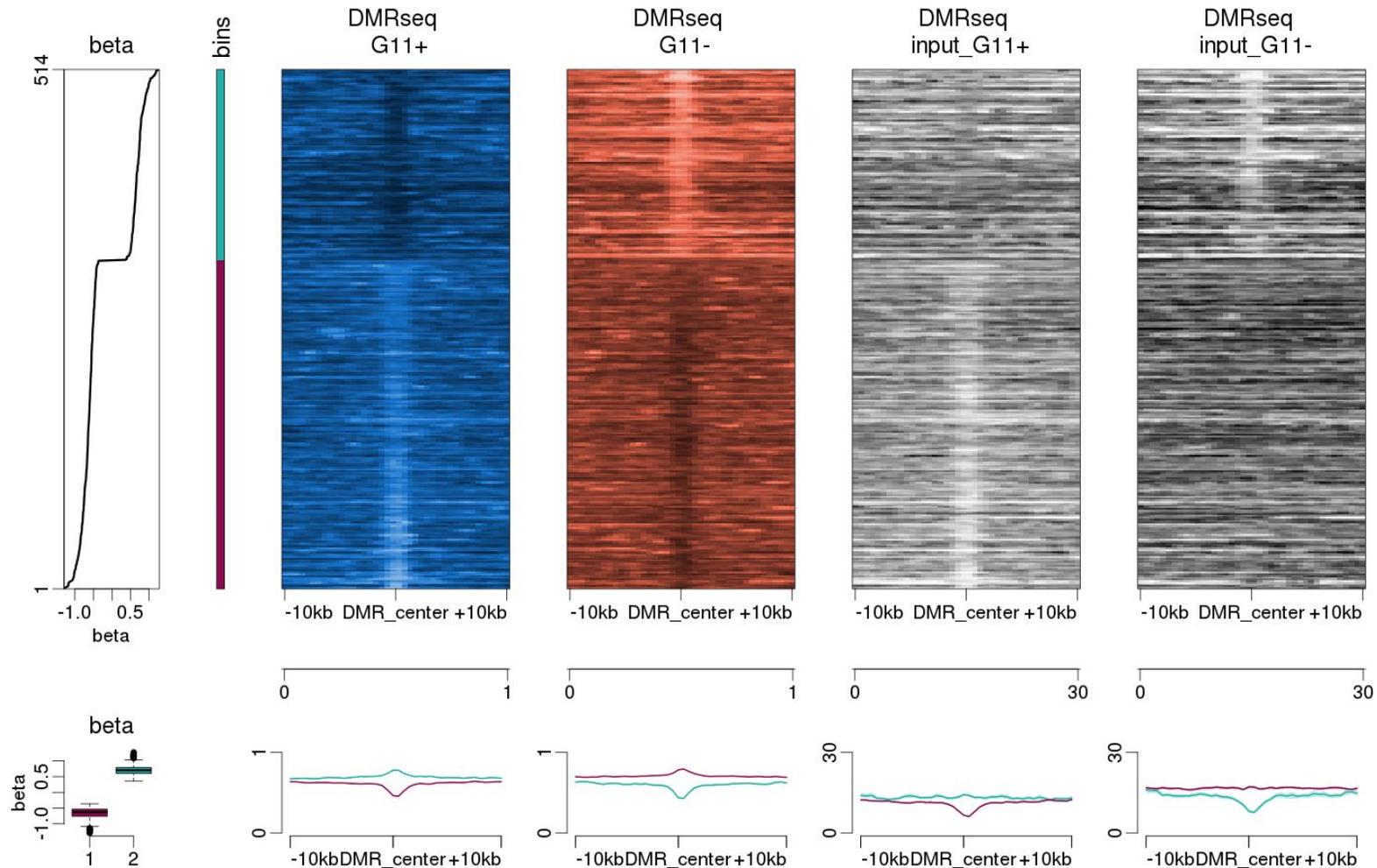
INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# Use case: visualisations of DMR

## Differentially Methylated Regions

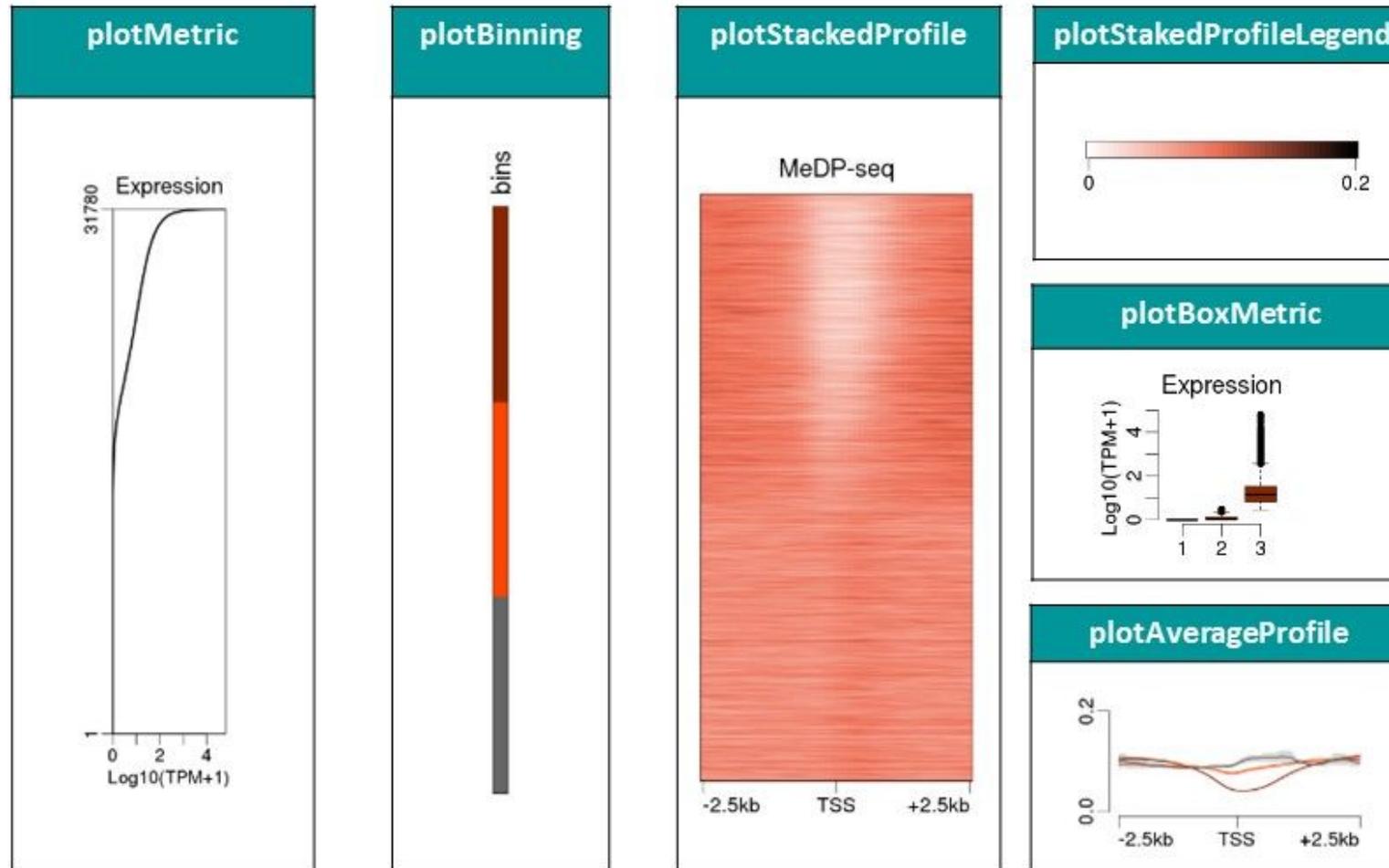


INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# Individual plotting functions



Assemble panels in the order you wish

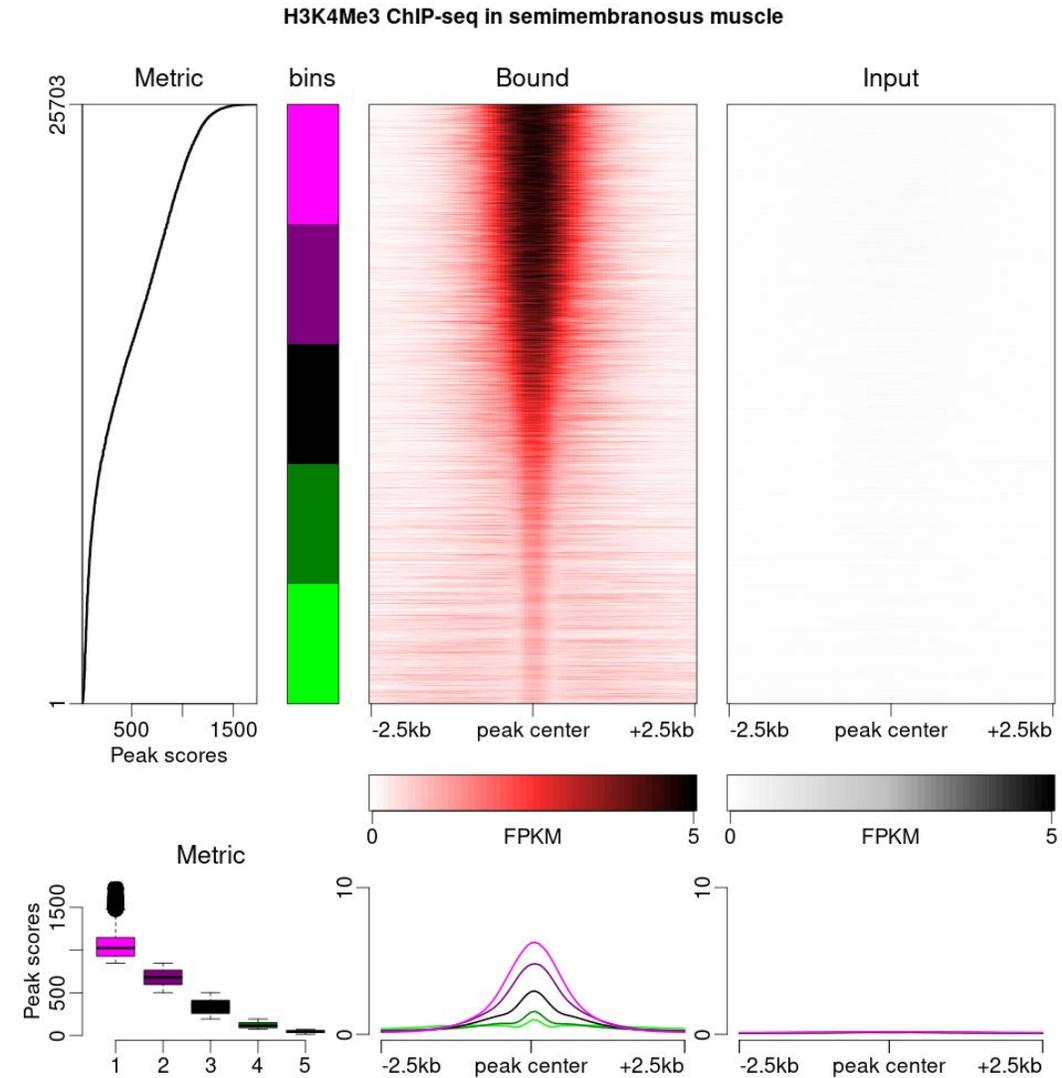
```
layout()  
{gridExtra}  
{patchwork}
```

# Experimental: CLI interface

Not feature complete, but may be useful (at least to me!)

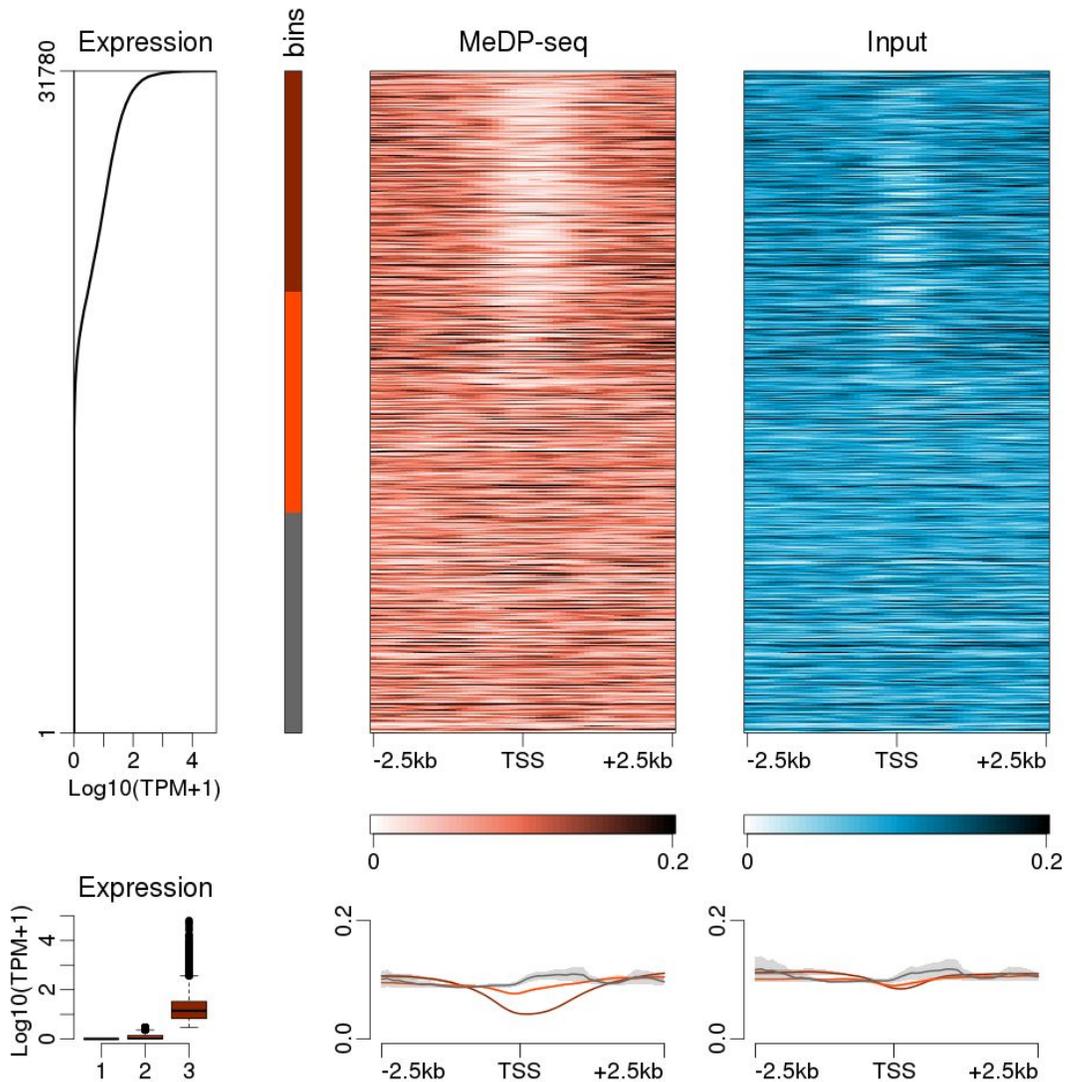
```
epistack.R \  
  -a ERX5798633_R1_peaks.narrowPeak \  
  -b ERX5798633_R1.bigWig \  
  -i ERX5798633_R1.bigWig \  
  -p ERX5798633.png \  
  -t 'H3K4Me3 ChIP-seq in semimembranosus muscle' \  
  -r center -y 10 -z 5 -c 2 -v -g 5 -m 99999 -f ci95
```

```
> Parsing files... done!  
> Processing... done!  
> Plotting... done!  
> Job completed for file: ERX5798633_R1.bigWig
```

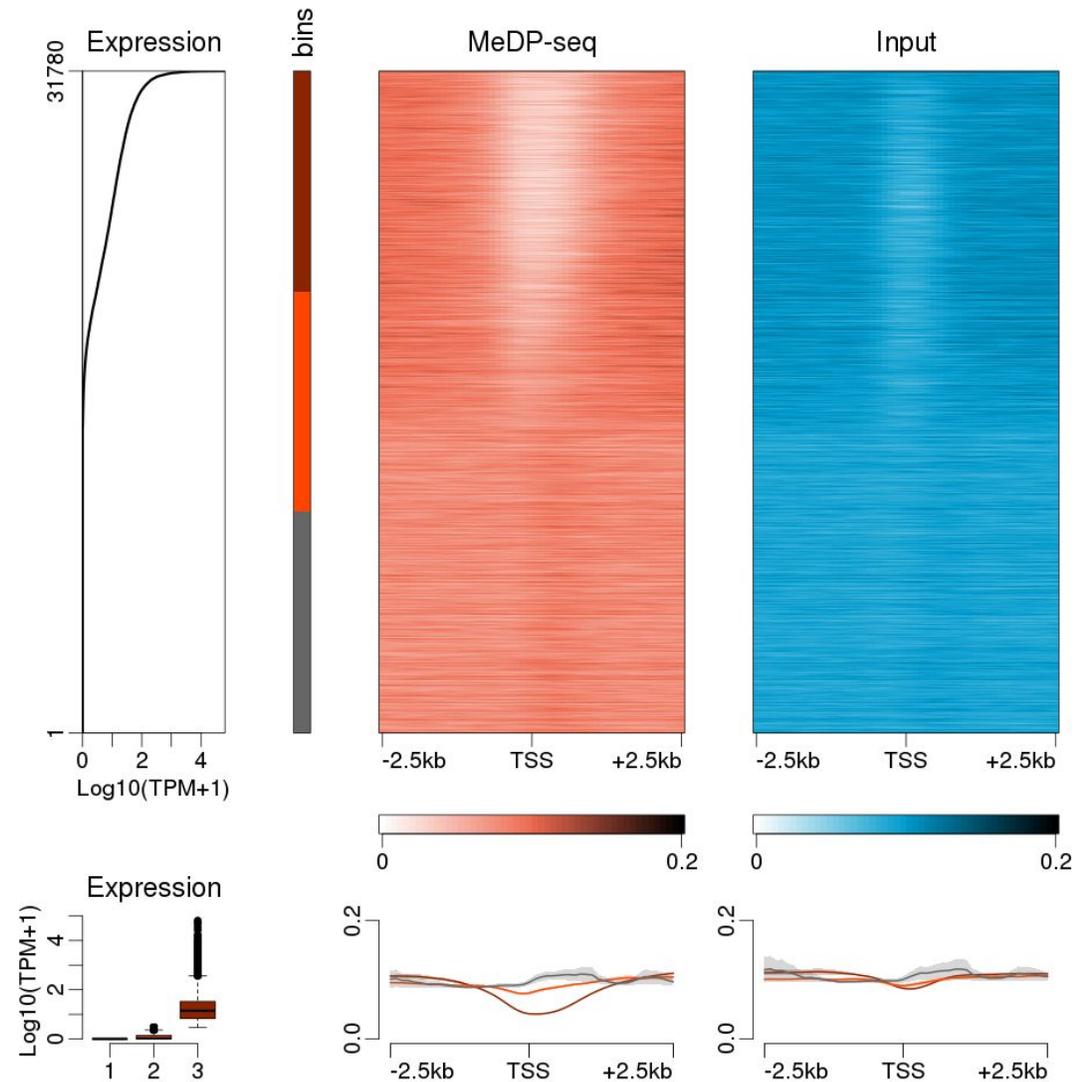


# The overplotting issue: more regions than pixels

## Default R behaviour

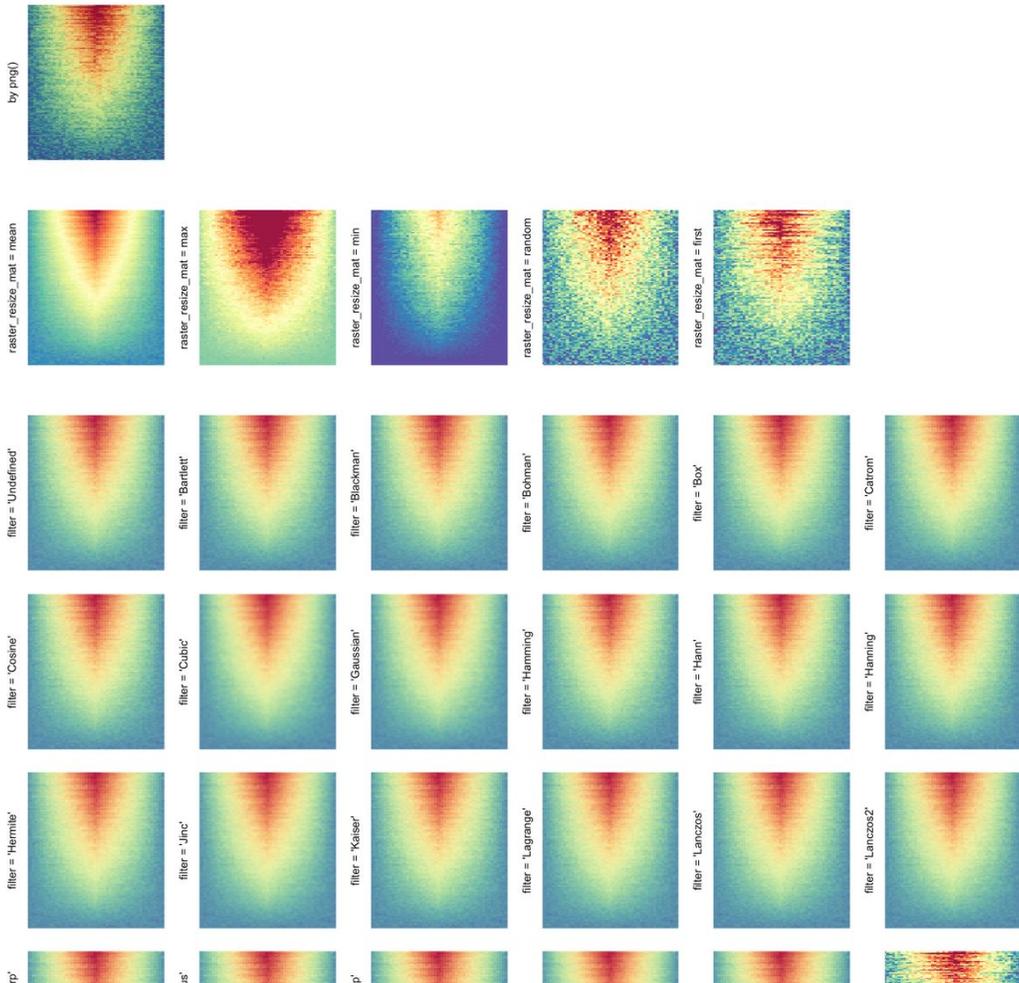


## Default {epistack} behaviour

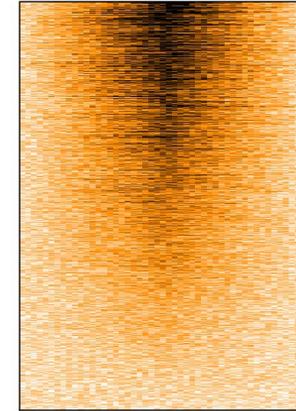


# The overplotting issue: more regions than pixels

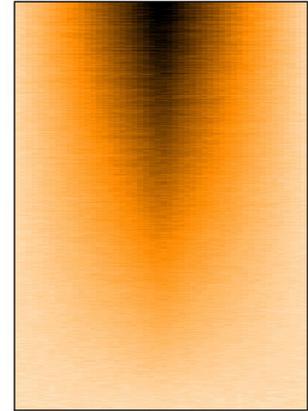
- `epistack::redimMatrix(mat, target_height, target_width, summary_func)`
- [bioinfo-fr.net/creer-des-heatmaps-a-partir-de-grosses-matrices-en-r](http://bioinfo-fr.net/creer-des-heatmaps-a-partir-de-grosses-matrices-en-r)
- [jokergoo.github.io/2020/06/30/rasterization-in-complexheatmap/](http://jokergoo.github.io/2020/06/30/rasterization-in-complexheatmap/)



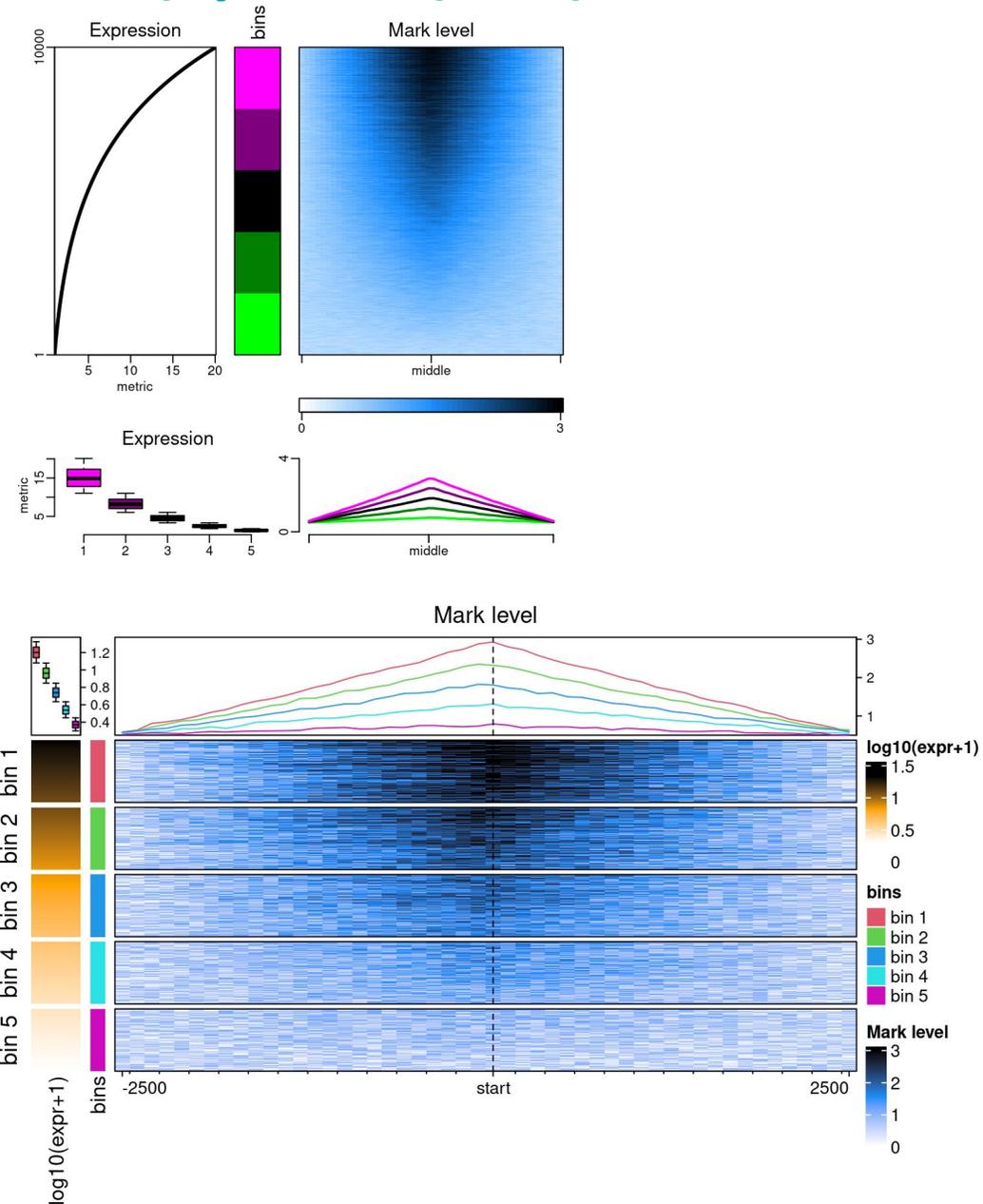
Original matrix



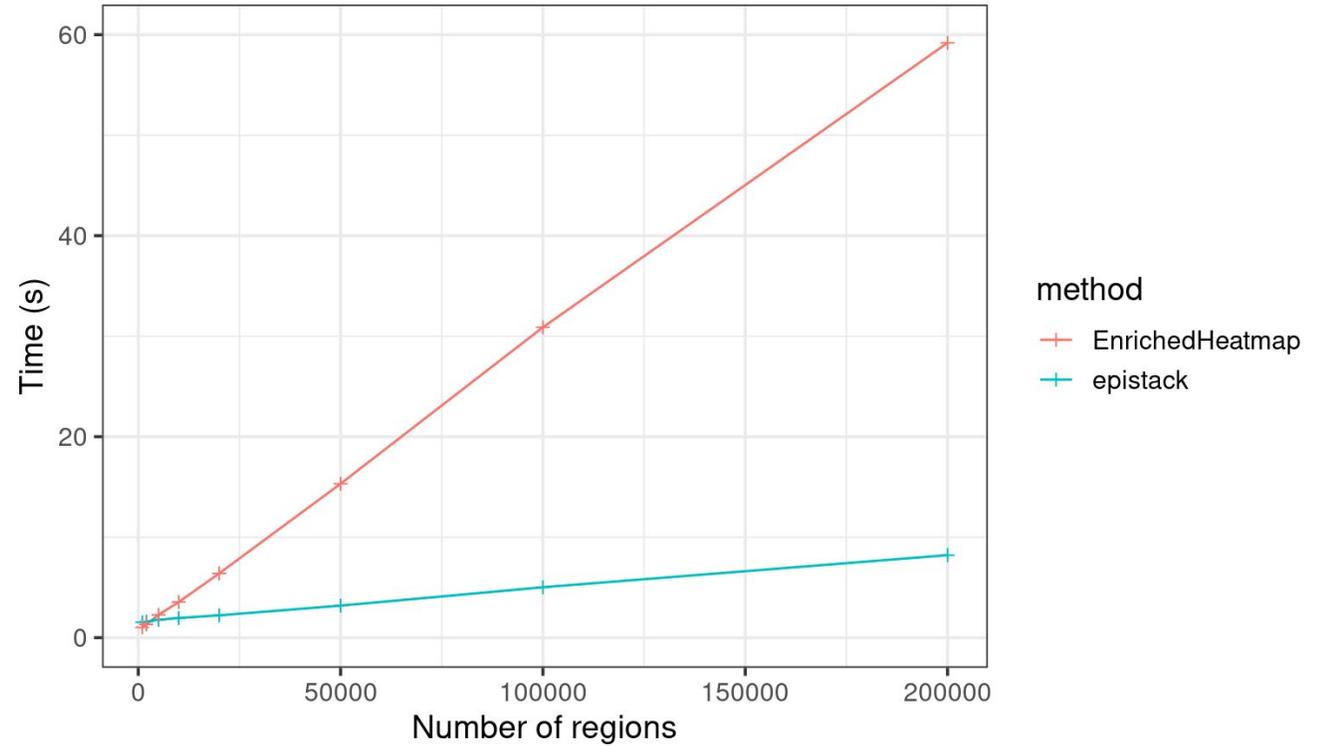
Reduced matrix



# {epistack} vs {EnrichedHeatmap} benchmark



plotEpistack() benchmark



# {epistack} in the real world

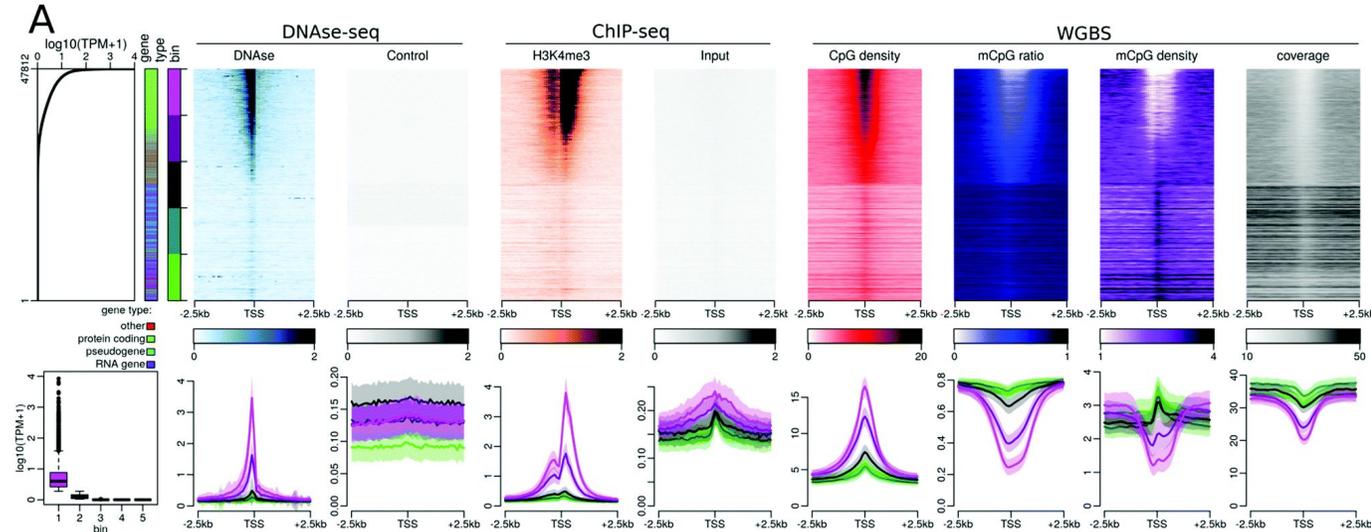
- inspired by our work on ROADMAP data

[doi.org/10.1039/d0mo00130a](https://doi.org/10.1039/d0mo00130a)

[joshiapps.cbu.uib.no/perepigenomics\\_app/](http://joshiapps.cbu.uib.no/perepigenomics_app/)

- Used internally to visualise DMR in various team projects

- To be applied on ALL FAANG CHIP-seq / ATAC-seq / WGBS data in the ANR VizFaDa



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly

# {epistack}

Available on github:

[github.com/GenEpi-GenPhySE/epistack](https://github.com/GenEpi-GenPhySE/epistack)

```
remotes::install_github("GenEpi-GenPhySE/epistack")
```

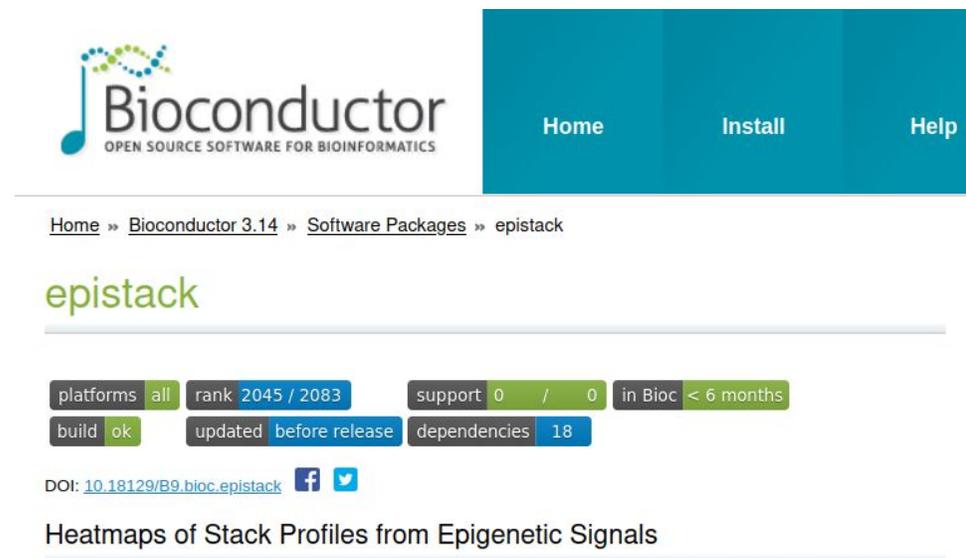
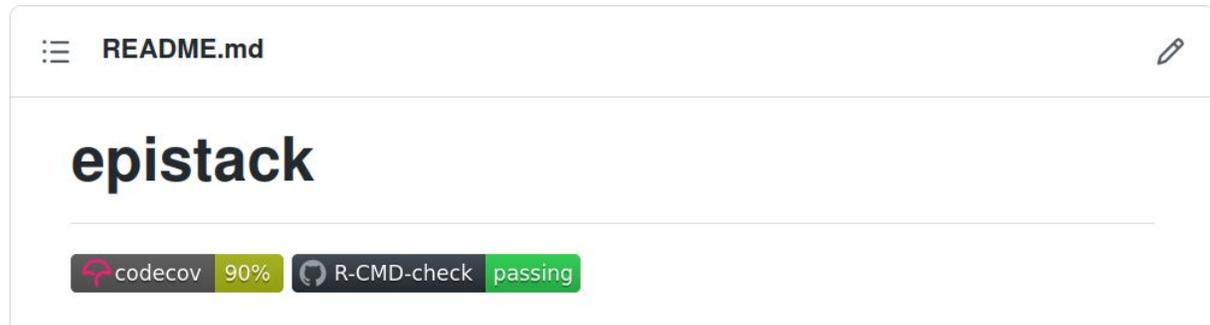
Available on Bioconductor:

[bioconductor.org/packages/epistack/](https://bioconductor.org/packages/epistack/)

Vignette:

[gdevailly.github.io/using\\_epistack.html](https://gdevailly.github.io/using_epistack.html)

Built using {devtools}, {testthat}, {roxygen2}, Github Actions



INRAE

{epistack}

2021-12-10 / R-Toulouse / Guillaume Devailly