UMI4Cats: An R package for analyzing UMI-4C chromatin contact data

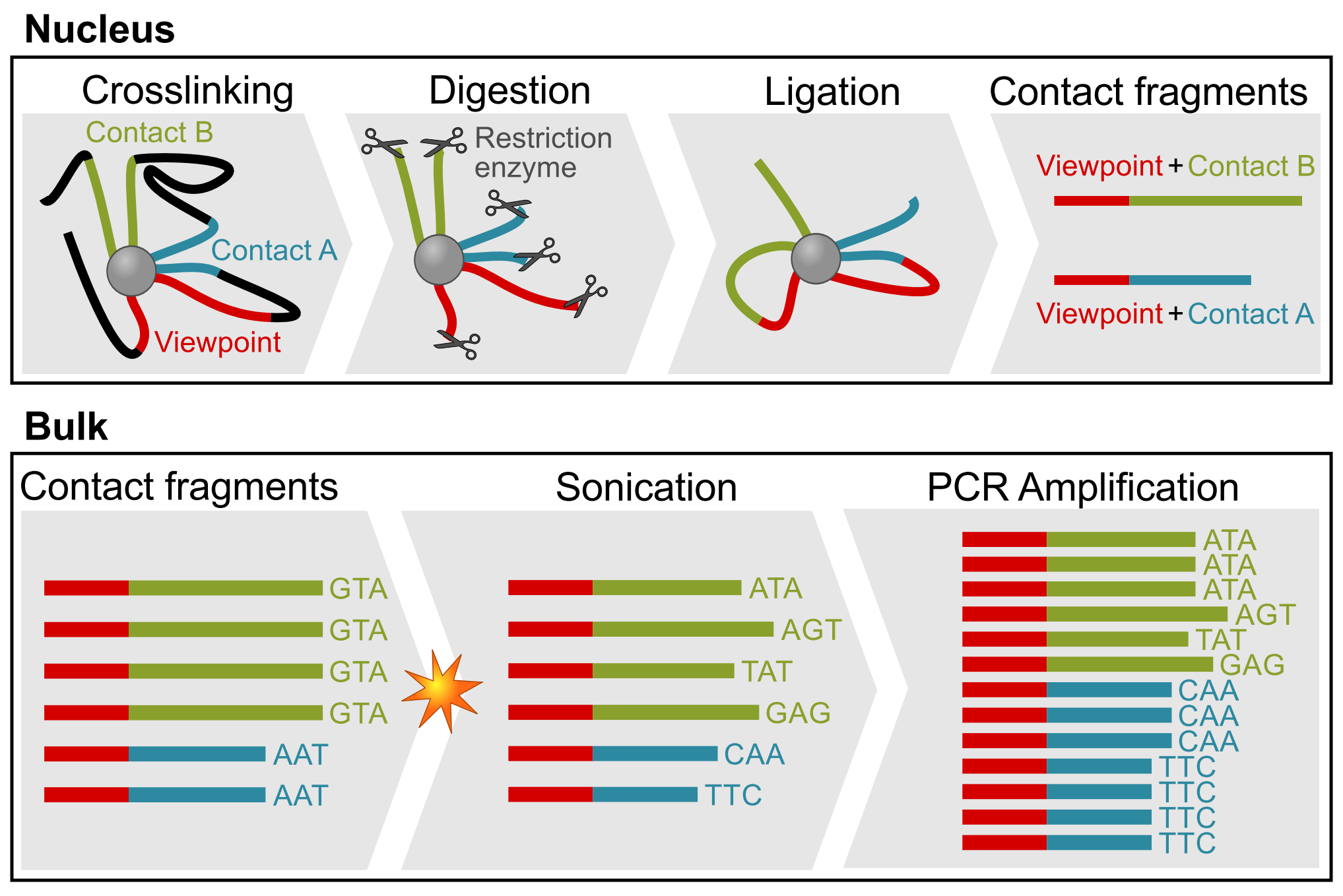
Mireia Ramos-Rodríguez[1], Marc Subirana-Granés[1], Lorenzo Pasquali[1,2,3]

[1] Endocrine Regulatory Genomics Lab, Germans Trias i Pujol University Hospital and Research Institute, Badalona, Spain
[2] Josep Carreras Leukaemia Research Institute, Barcelona, Spain
[3] CIBER de Diabetes y Enfermedades Metabólicas Asociadas, Barcelona, Spain

Background

Chromatin physical interactions guide the association of enhancers to target genes. 3C-derived techniques allow us to detect such chromatin contacts, but share a PCR amplification step that limits the quantitative comparison of the contact intensities detected in different cell types or conditions.

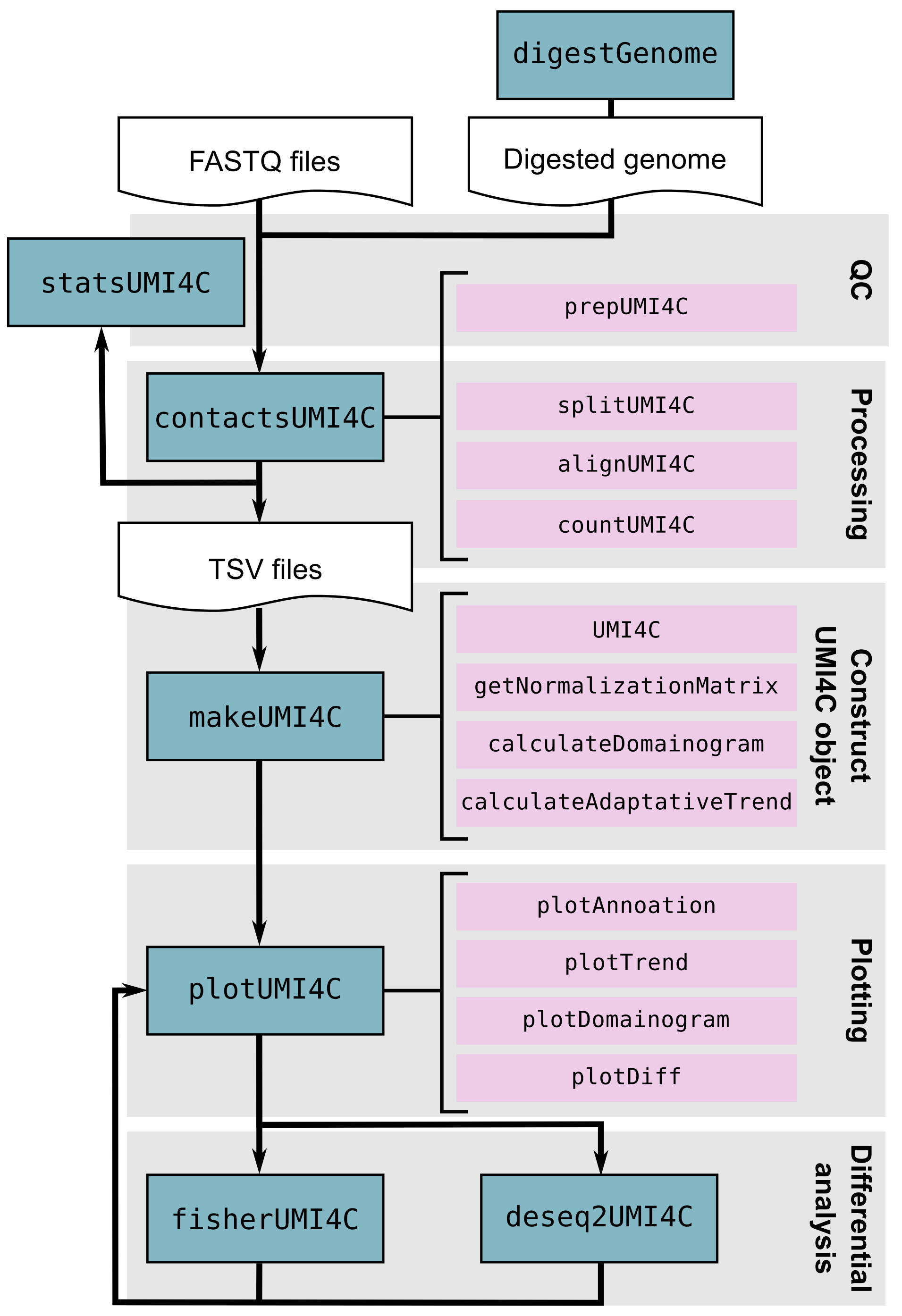
UMI-4C (*Schwartzmann et al., 2016*) is a technique that allows deriving high resolution and high complexity quantitative contact profiles of a selected viewpoint of interest. A key advance of this technique is the inclusion of a **sonication** step, in which molecules are randomly cut. This produces a sort of **Unique Moleuclar Identifier (UMI),** making it possible to recognize PCR duplicates and derive contact intensities accurately and quantitatively.



UMI4Cats (**UMI-4C** **A**nalysis **T**urned **S**imple) is the first R package that allows easy and fast processing and analysis of UMI-4C experiments.

Methods

Before analyzing a UMI-4C experiment, a digested genome needs to be generated. This can be done using `digestGenome` with any restriction sequence.



1) Quality control

The experiment quality control (QC) metrics are performed at different steps: 1) on the raw FastQ files and 2) after the read alignment.

A summary of the sample statistics is returned and can be plotted using the `statsUMI4C` function.

2) Processing

The processing step converts UMI-4C FastQ files into a list of fragments an the number of UMIs supporting the contacts.

This is done in three steps: (1) FastQ reads are **digested** (split) at the restriction sequences, (2) digested reads are **aligned** to the reference genome using {RBowtie2} and (3) the **UMI filtration algorithm** collapses contacts with the exact same position or with <2 mismatches.

3) Construct `UMI4C-class` object

An object of the `UMI4C` class, representative of the locus contact intensities, is constructed by computing a normalization matrix, domainograms and the adaptative smoothen trend (see ***Results***).

4) Differential analysis

UMI-4C experiments allow for accurate differential testing. UMI4Cats implements two different methods: (1) **Fisher's Exact test**, performed at specific regions of inerest (for example, genomic annotations) or at binned windows and (2) **Wald's Test** (from {DESeq2}) at fragment ends.

5) Plotting

The `UMI4C` object can be plotted before or after performing differential analysis (see ***Results***).

Results

UMI-4C experiment statistics

During the processing of UMI-4C reads, statistics from several key filtering processes are outputed:

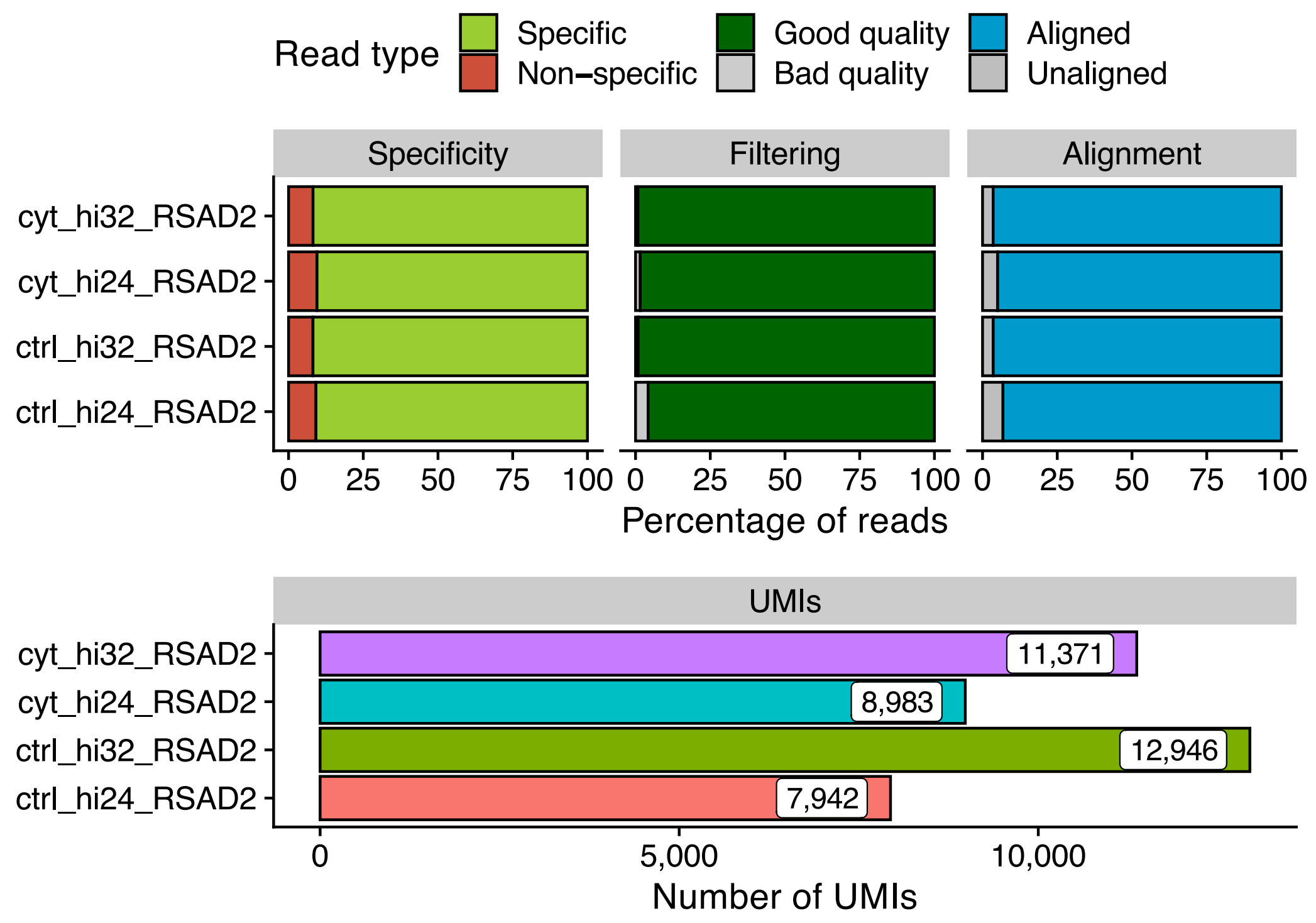
(1) **Specificity** of the sequenced fragments, dtermined by the presence of the {bait + padding + restriction} sequences.

(2) Read **quality**, reads with mean Phred scores < 20 are filtered out.

(3) **Alignment** statistics, only reads aligned with MAPQ>30 are kept.

(4) Detected **UMIs** after collapsing PCR duplicates using the UMI filtering algorithm.

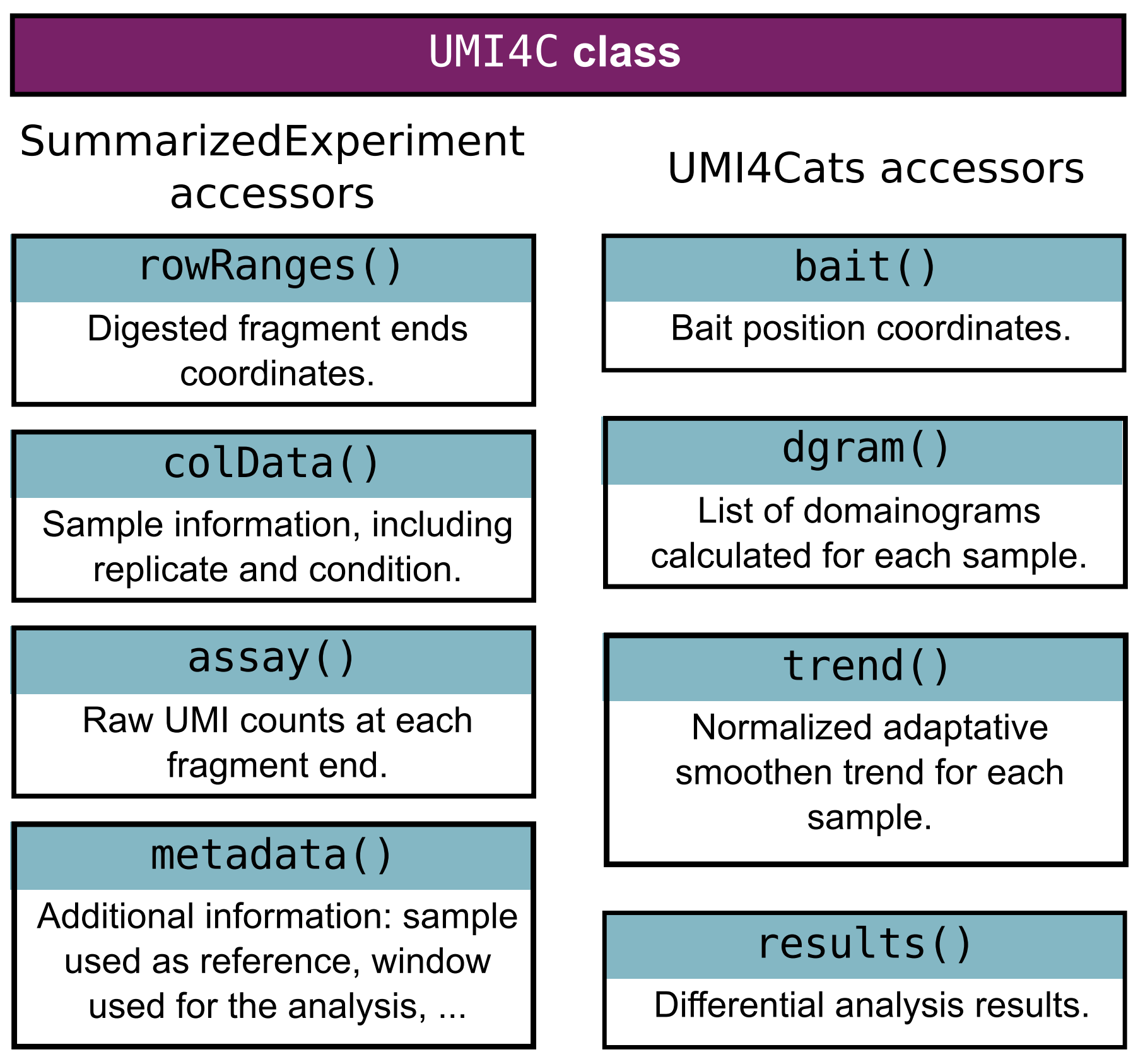
The resulting values can then be plotted using the `statsUMI4C` function.



`UMI4C` class

The `UMI4C` class is defined in the UMI4Cats package and inherits from the SummarizedExperiment S4 class.

This class has some specific accessors to retrieve relevant UMI-4C experiment informaton.



UMI4C class

SummarizedExperiment accessors

`rowRanges()`: Digested fragment ends coordinates.

`colData()`: Sample information, including replicate and condition.

`assay()`: Raw UMI counts at each fragment end.

`metadata()`: Additional information: sample used as reference, window used for the analysis, ...

UMI4Cats accessors

`bait()`: Bait position coordinates.

`dgram()`: List of domainograms calculated for each sample.

`trend()`: Normalized adaptative smoothen trend for each sample.

`results()`: Differential analysis results.

```
> umi_obj

class: UMI4C
dim: 2303 4
metadata(5): bait scales min_win_factor region ref_umi4c
assays(6): umis norm_mat ... scale sd
rownames(2303): frag_14737 frag_14738 ... frag_17049
frag_17050
rowData names(2): id_contact position
colnames(4): ctrl_hi24_RSAD2 ctrl_hi32_RSAD2 cyt_hi24_RSAD2
cyt_hi32_RSAD2
colData names(5): sampleID condition replicate viewpoint
file
```

UMI4Cats output plot

The plot outputed by `plotUMI4C` contains the following information:

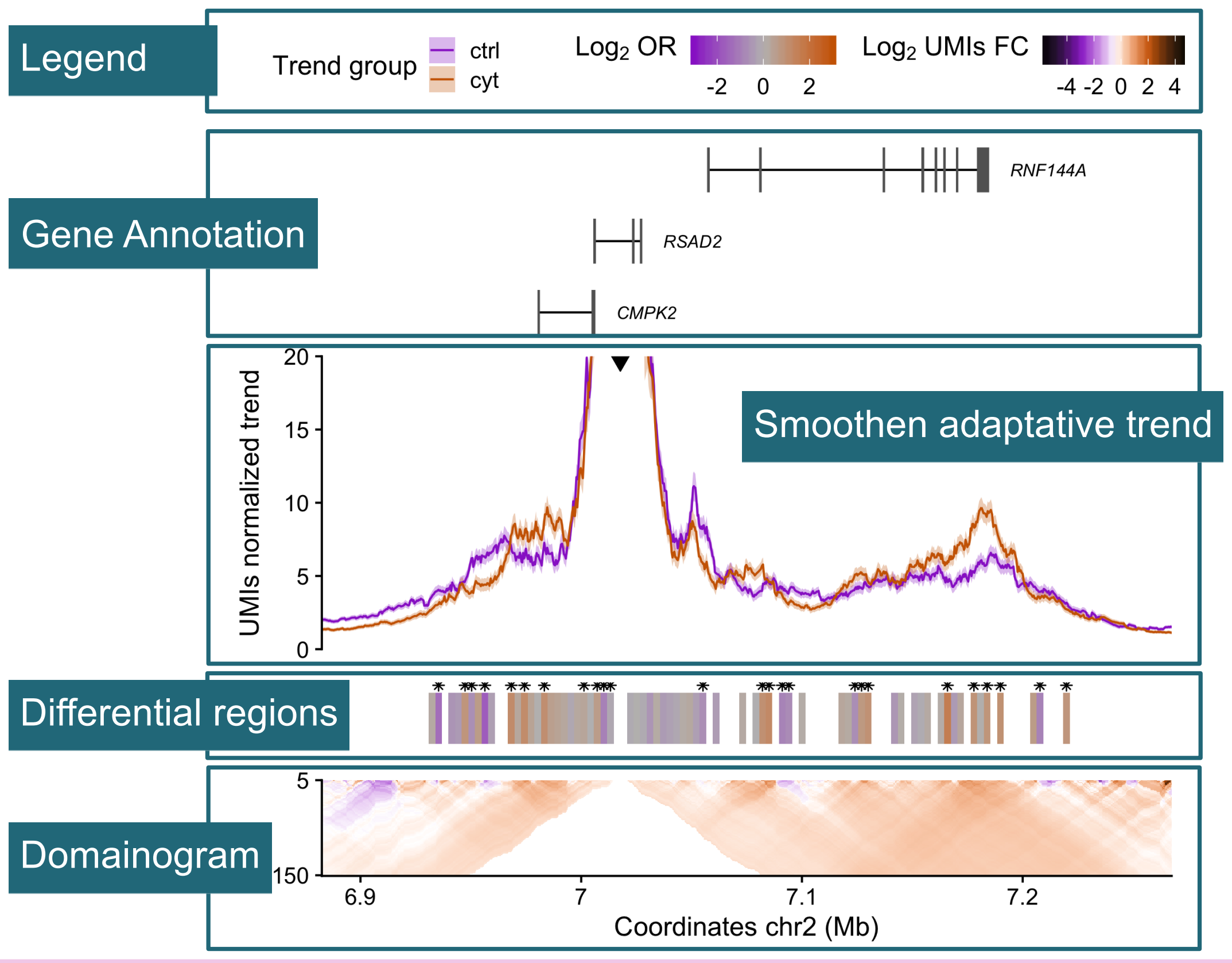
(1) **Legend**, representing values and scales of the different plot elements.

(2) **Gene Annotation**, showing the annotation of coding genes in the region of interest.

(3) **Smoothen adaptative trend**, representing the normalized profile of the UMI-4C contacts at each position.

(4) **Differential test results**, showing the position of the tested differential regions. An asterisk indacates statistical significande (adjusted *P*-value < 0.05).

(5) **Domainogram**, representing intensities of contacts merging the number of digested contacts defined in the y axis.



Conclusions

UMI4Cats is a user friendly package that allows processing and analysis of UMI-4C experiments. Moreover, it takes advantage of well-known Bioconductor packages such as {GenomicRanges}, {Biostrings}, {Rsamtools}, {SummarizedExperiment} and {DESeq2} to process and analyze these complex data.

This package provides several accessor functions that allow easy retrieval of processed UMI-4C data, as well as highly customizable plots that summarize all the information contained in the `UMI4C` object. It also includes several statistical methods to determine differential contact intensities.

UMI4Cats is currently under development and available upon request. We plan to submit it to Bioconductor soon.

References

R packages used by UMI4Cats:

`Biostrings, BSgenome, cowplot, DESeq2, dplyr, GenomicAlignments, GenomicRanges, ggplot2, magrittr, Rbowtie2, regioneR, reshape2, Rsamtools, S4Vectors, scales, ShortRead, SummarizedExperiment`

Acknowledgements:

IGTP Germans Trias i Pujol Research Institute · Josep Carreras LEUKAEMIA Research Institute · ciberdem · AGAUR Agència de Gestió d'Ajuts Universitaris i de Recerca · UNIÓN EUROPEA Fondo Social Europeo · GOBIERNO DE ESPAÑA MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL · La Marató 3