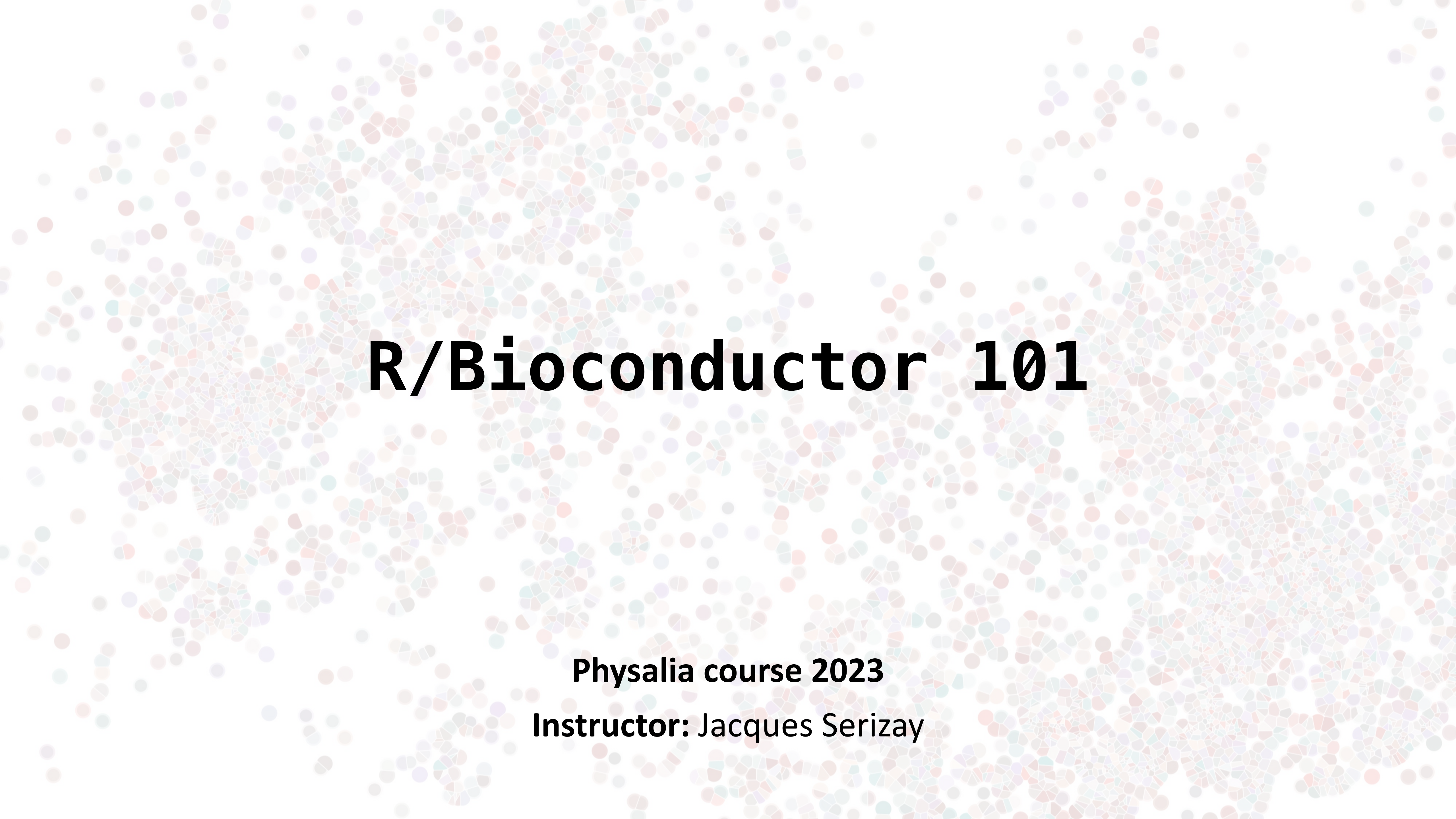


R/Bioconductor 101

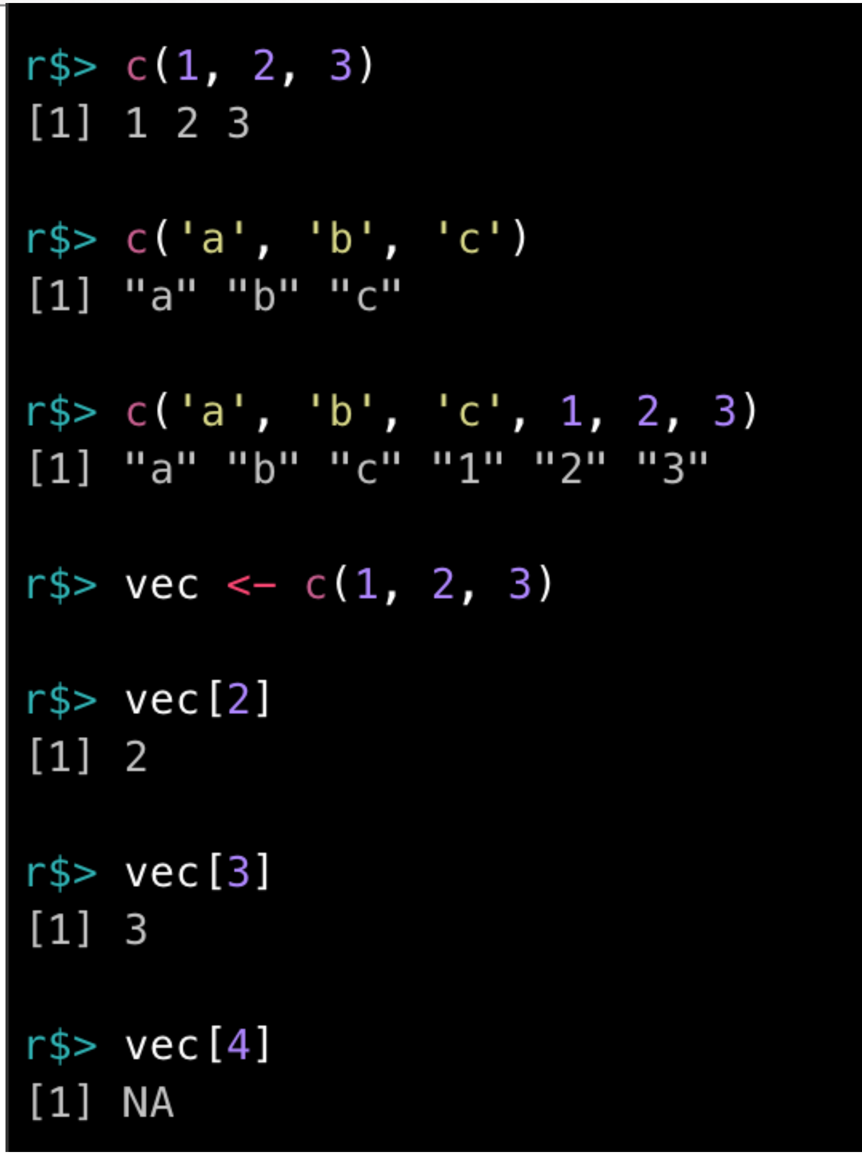
Physalia course 2023

Instructor: Jacques Serizay

Vectors

Vectors

- Defined with c() function
- All the elements must be from the same class
- Can be subset with […]



```
r$> c(1, 2, 3)
[1] 1 2 3

r$> c('a', 'b', 'c')
[1] "a" "b" "c"

r$> c('a', 'b', 'c', 1, 2, 3)
[1] "a" "b" "c" "1" "2" "3"

r$> vec <- c(1, 2, 3)

r$> vec[2]
[1] 2

r$> vec[3]
[1] 3

r$> vec[4]
[1] NA
```

Tibbles

Tibbles

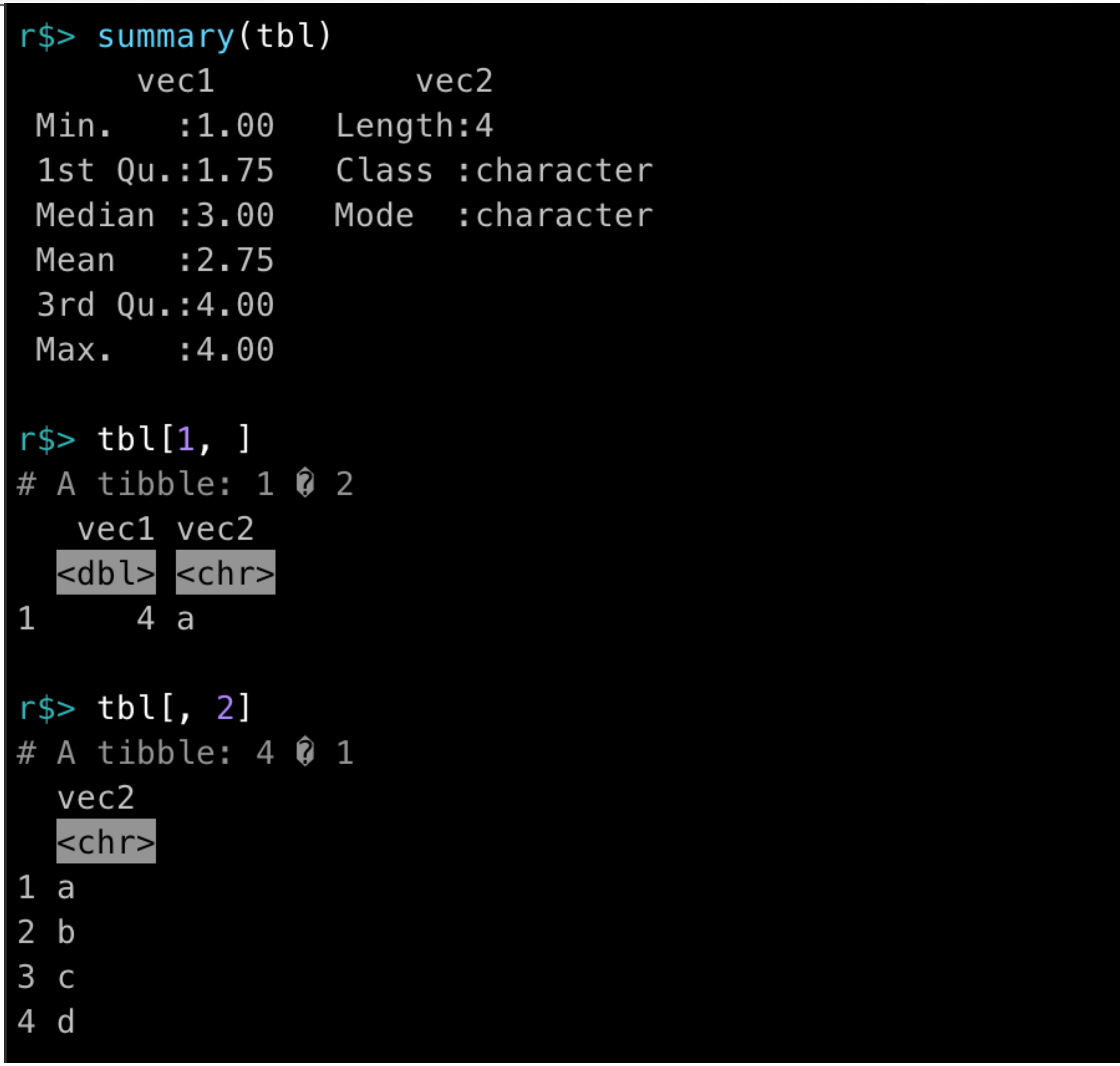
- Modern data.frame, tabular shape
- Created with tibble()

```
r$> library(tibble)

r$> tbl <- tibble(
    "vec1" = c(4, 1, 2, 4),
    "vec2" = c('a', 'b', 'c', 'd')
    )

r$> tbl
# A tibble: 4 � 2
   vec1 vec2
  <dbl> <chr>
1     4 a
2     1 b
3     2 c
4     4 d
```

Tibbles

Tibbles

- Modern data.frame, tabular shape
- Created with tibble()
- Subset with […, …]



```
r$> summary(tbl)
      vec1           vec2          
 Min.   :1.00   Length:4          
 1st Qu.:1.75   Class :character  
 Median :3.00   Mode  :character  
 Mean   :2.75                     
 3rd Qu.:4.00                     
 Max.   :4.00                     

r$> tbl[1, ]
# A tibble: 1 � 2
   vec1 vec2 
  <dbl> <chr>
1     4 a    

r$> tbl[, 2]
# A tibble: 4 � 1
  vec2 
  <chr>
1 a    
2 b    
3 c    
4 d    
```

Tibbles

Tibbles

- Modern data.frame, tabular shape
- Created with tibble()
- Subset with […, …]
- Columns can also be accessed with [[…]] or $

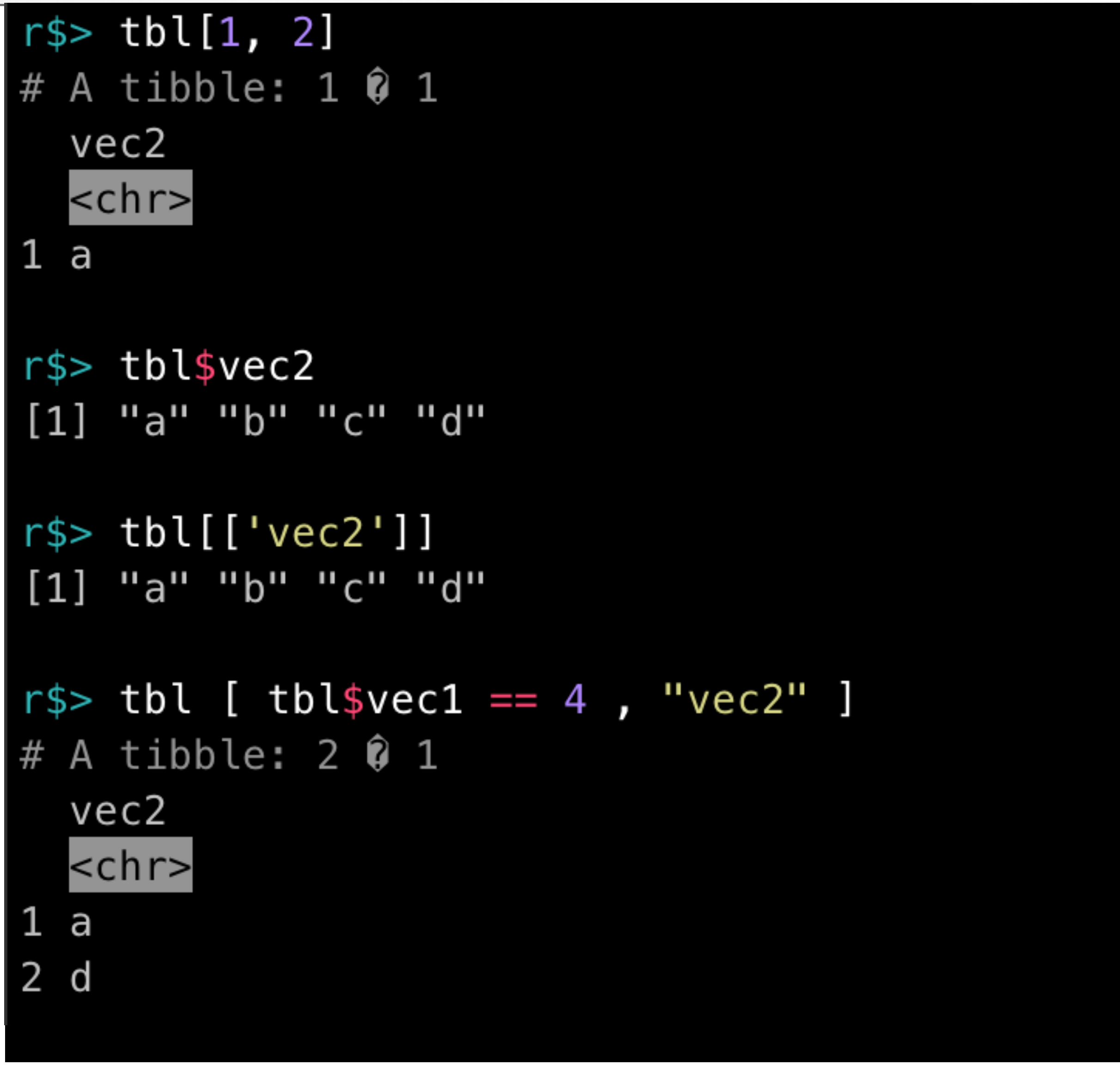
```
r$> tbl[1, 2]
# A tibble: 1 � 1
  vec2
  <chr>
1 a

r$> tbl$vec2
[1] "a" "b" "c" "d"

r$> tbl[['vec2']]
[1] "a" "b" "c" "d"
```

Tibbles

Tibbles

- Modern data.frame, tabular shape
- Created with tibble()
- Subset with […, …]
- Columns can also be accessed with [[…]] or $

```
tbl [ tbl$vec1 == 4 , "vec2" ]
```

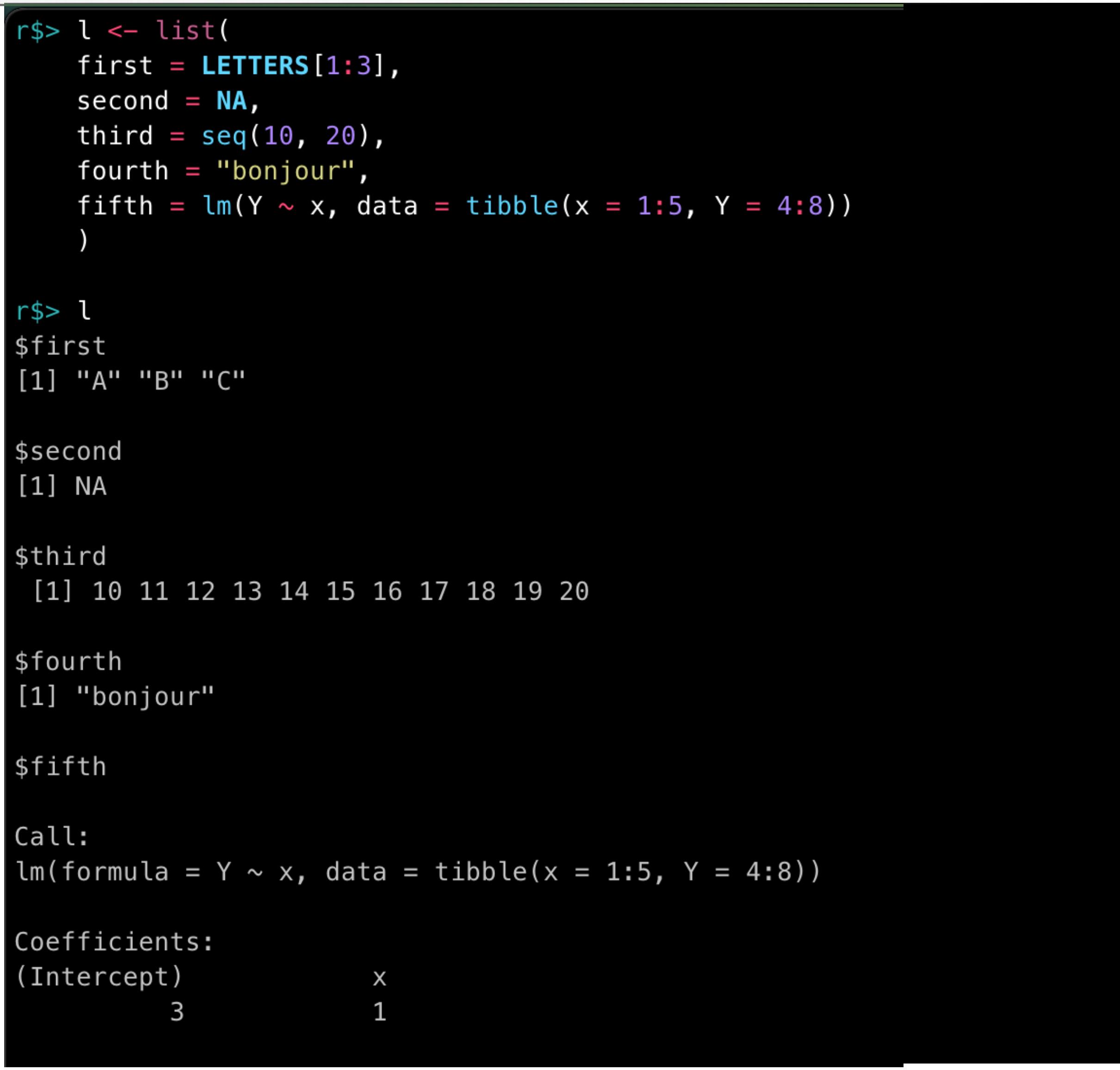


```
r$> tbl[1, 2]
# A tibble: 1 � 1
  vec2 
  <chr>
1 a    

r$> tbl$vec2
[1] "a" "b" "c" "d"

r$> tbl[['vec2']]
[1] "a" "b" "c" "d"

r$> tbl [ tbl$vec1 == 4 , "vec2" ]
# A tibble: 2 � 1
  vec2 
  <chr>
1 a    
2 d    
```

Lists

Lists

- Created with list() function
- Each element can be whatever object you want
- Each element can be named

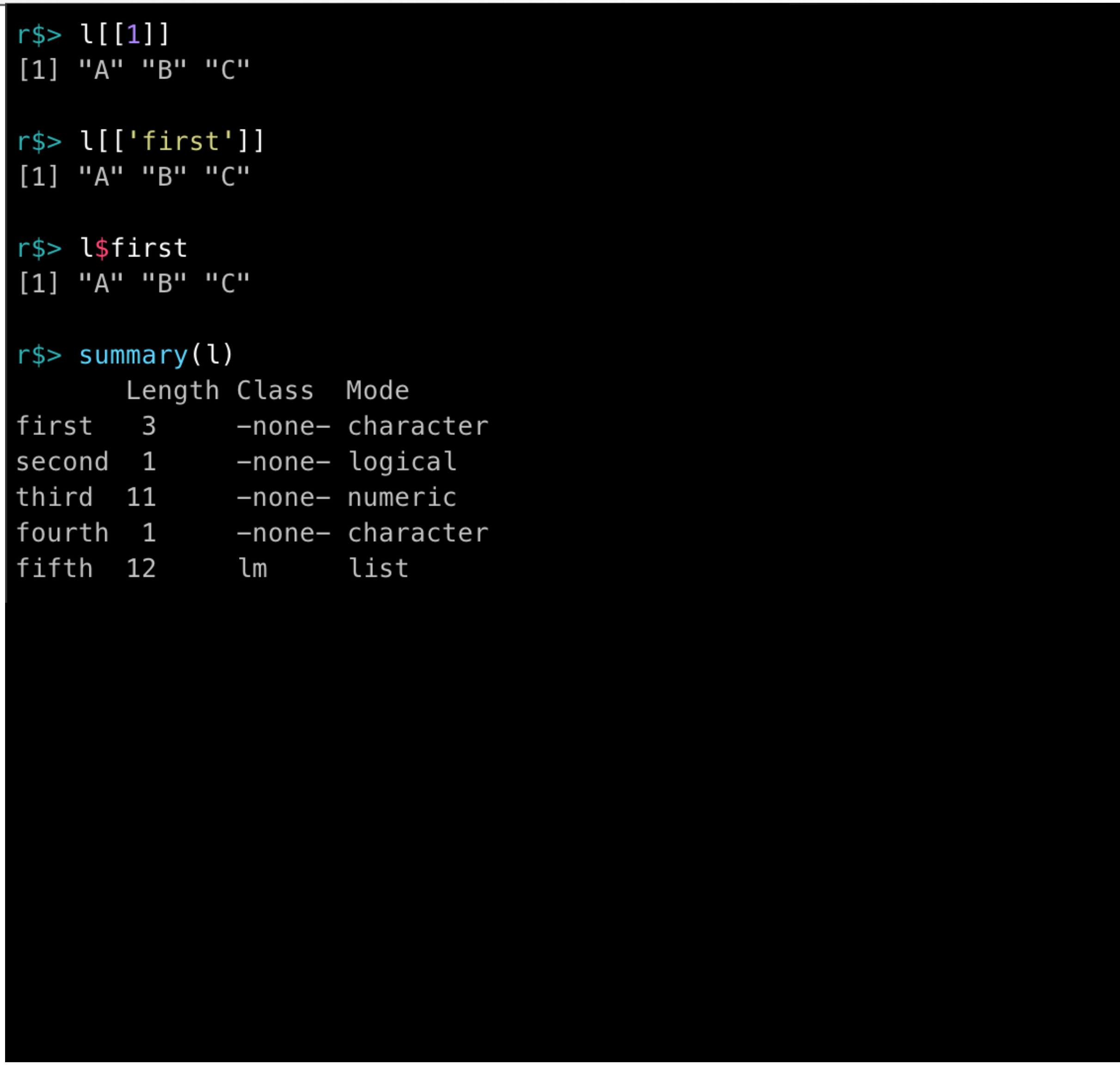


```
r$> l <- list(
    first = LETTERS[1:3],
    second = NA,
    third = seq(10, 20),
    fourth = "bonjour",
    fifth = lm(Y ~ x, data = tibble(x = 1:5, Y = 4:8))
    )

r$> l
$first
[1] "A" "B" "C"

$second
[1] NA

$third
 [1] 10 11 12 13 14 15 16 17 18 19 20

$fourth
[1] "bonjour"

$fifth

Call:
lm(formula = Y ~ x, data = tibble(x = 1:5, Y = 4:8))

Coefficients:
(Intercept)            x
          3            1
```

Lists

Lists

- Created with list() function
- Each element can be whatever object you want
- Each element can be named
- Elements can be accessed using $ or [[…]]

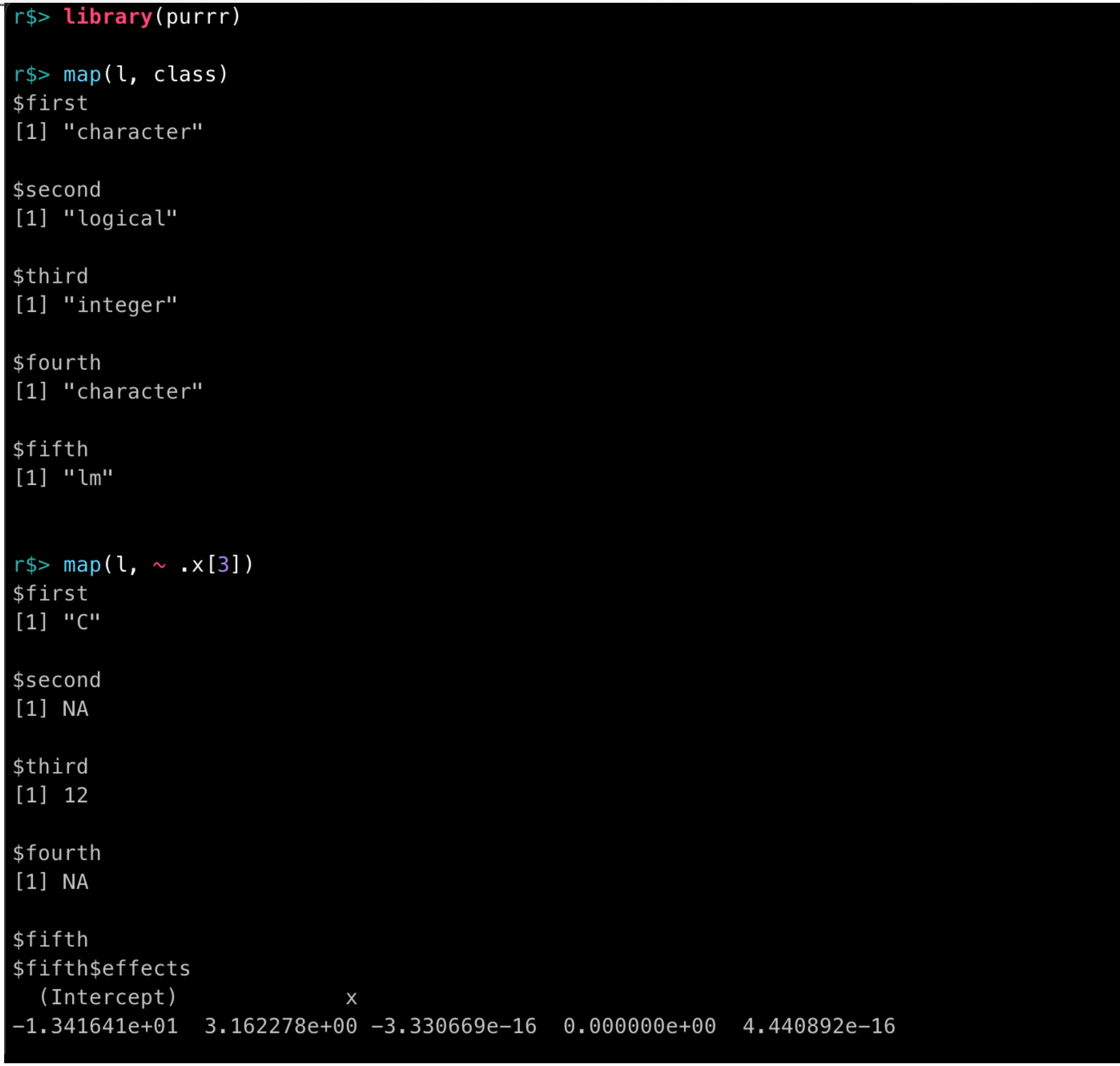


```
r$> l[[1]]
[1] "A" "B" "C"

r$> l[['first']]
[1] "A" "B" "C"

r$> l$first
[1] "A" "B" "C"

r$> summary(l)
       Length Class  Mode     
first   3     -none- character
second  1     -none- logical  
third  11     -none- numeric  
fourth  1     -none- character
fifth  12     lm     list     
```

Lists

Lists

- Created with list() function
- Each element can be whatever object you want
- Each element can be named
- Elements can be accessed using $ or [[…]]
- One can “map” or “apply” (`lapply`) a function over each element of a list



```
r$> library(purrr)

r$> map(l, class)
$first
[1] "character"

$second
[1] "logical"

$third
[1] "integer"

$fourth
[1] "character"

$fifth
[1] "lm"


r$> map(l, ~ .x[3])
$first
[1] "C"

$second
[1] NA

$third
[1] 12

$fourth
[1] NA

$fifth
$fifth$effects
  (Intercept)             x
-1.341641e+01  3.162278e+00 -3.330669e-16  0.000000e+00  4.440892e-16
```

R essentials

Tidyverse (https://rstudio-education.github.io/tidyverse-cookbook/program.html)

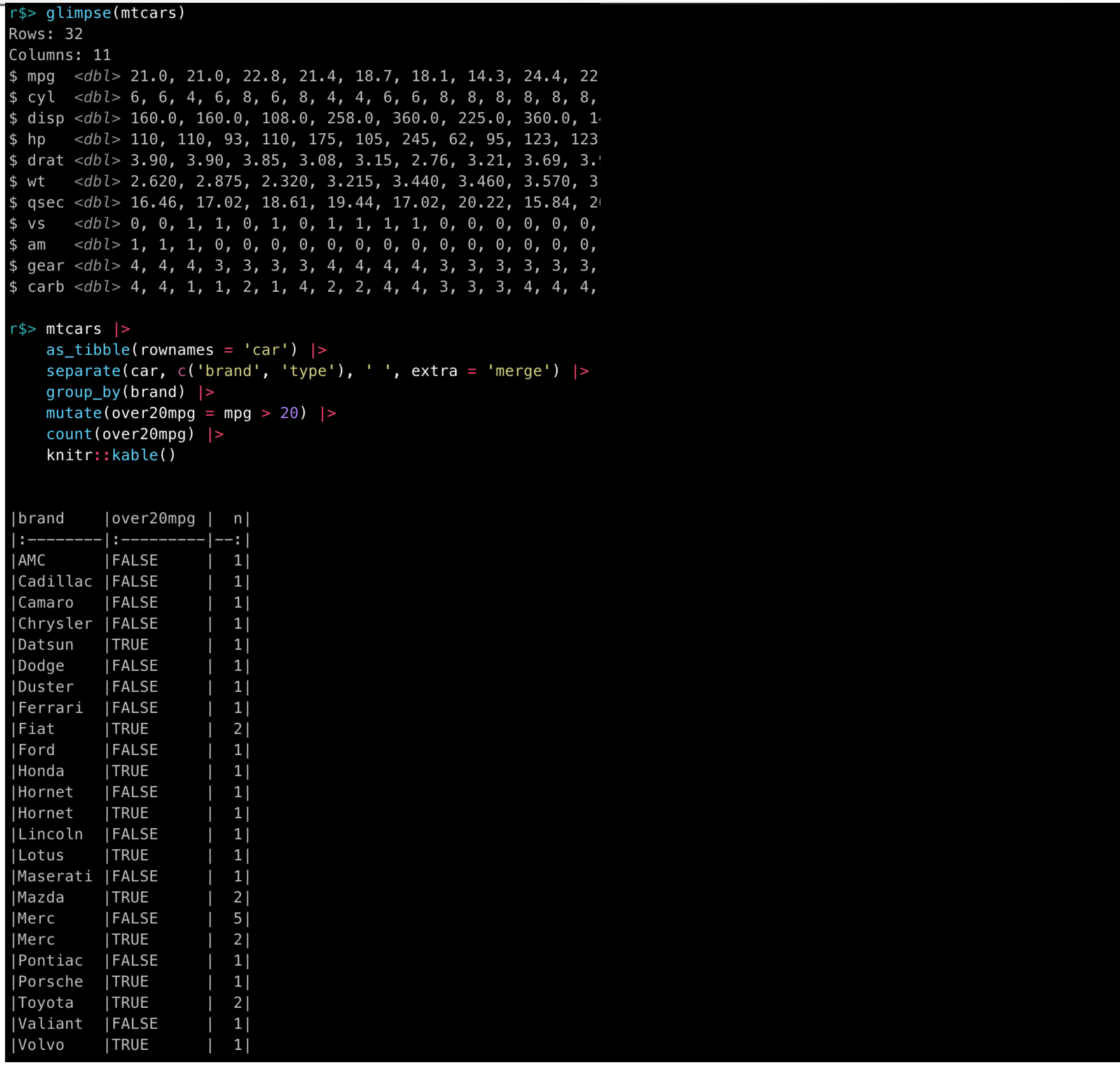
- Verb-based ecosystem
- dplyr::filter
- dplyr::arrange
- dplyr::mutate
- purrr::map
- tidyr::pivot_*
- ggplot2 plotting functions

Everything documented here: https://www.r-bloggers.com/2020/12/the-tidyverse-in-a-table/

R essentials

Native `|>` pipe

- Just like a pipe in bash, for R
- Very useful in combination with tidyverse's dplyr for data wrangling



```
r$> glimpse(mtcars)
Rows: 32
Columns: 11
$ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22
$ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8,
$ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 1
$ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123
$ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.
$ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3
$ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 2
$ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
$ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3,
$ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4,

r$> mtcars |>
    as_tibble(rownames = 'car') |>
    separate(car, c('brand', 'type'), ' ', extra = 'merge') |>
    group_by(brand) |>
    mutate(over20mpg = mpg > 20) |>
    count(over20mpg) |>
    knitr::kable()


|brand    |over20mpg |  n|
|:--------|:---------|--:|
|AMC      |FALSE     |  1|
|Cadillac |FALSE     |  1|
|Camaro   |FALSE     |  1|
|Chrysler |FALSE     |  1|
|Datsun   |TRUE      |  1|
|Dodge    |FALSE     |  1|
|Duster   |FALSE     |  1|
|Ferrari  |FALSE     |  1|
|Fiat     |TRUE      |  2|
|Ford     |FALSE     |  1|
|Honda    |TRUE      |  1|
|Hornet   |FALSE     |  1|
|Hornet   |TRUE      |  1|
|Lincoln  |FALSE     |  1|
|Lotus    |TRUE      |  1|
|Maserati |FALSE     |  1|
|Mazda    |TRUE      |  2|
|Merc     |FALSE     |  5|
|Merc     |TRUE      |  2|
|Pontiac  |FALSE     |  1|
|Porsche  |TRUE      |  1|
|Toyota   |TRUE      |  2|
|Valiant  |FALSE     |  1|
|Volvo    |TRUE      |  1|
```

R essential packages

R *per se* is useful for statistical analyses.

Why do bioinformaticians keep talking about R then?

In other words, **how do we unlock the power of R-stats in genomics**?

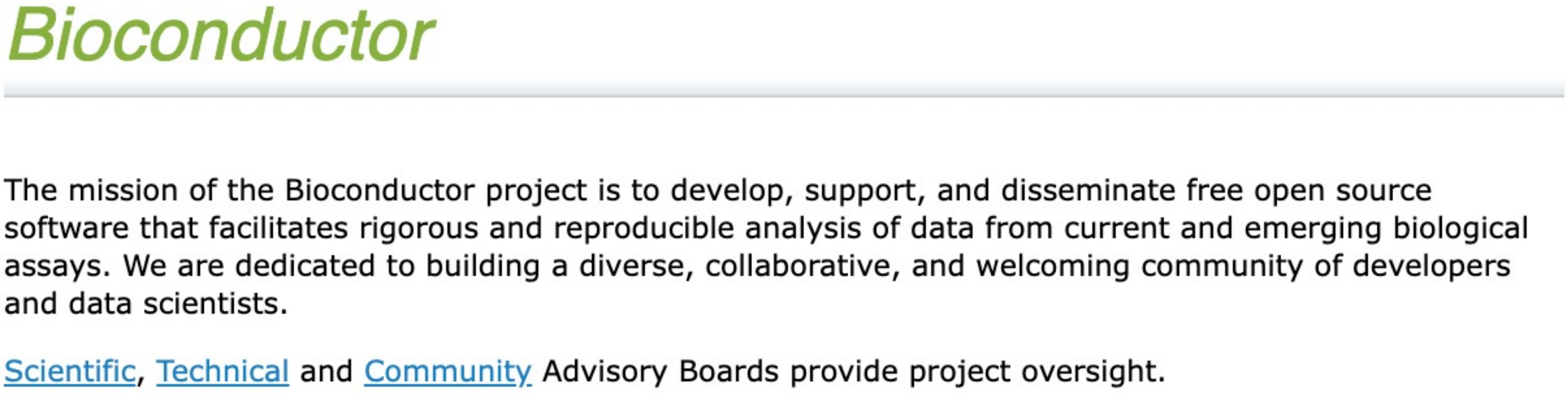
What do you need in bioinformatics to study genomics?

Most common genomic files:

- **BED** format: essentially a set of chromosomal ranges
- **BigWig** format: essentially veeeeeeeee…eeeeery long numerical vectors
- **Fasta** format: letters, letters, letters
- **Others** (bam, GFF, …): can usually be described/built on as one of the two options above

Bioconductor

https://bioconductor.org/



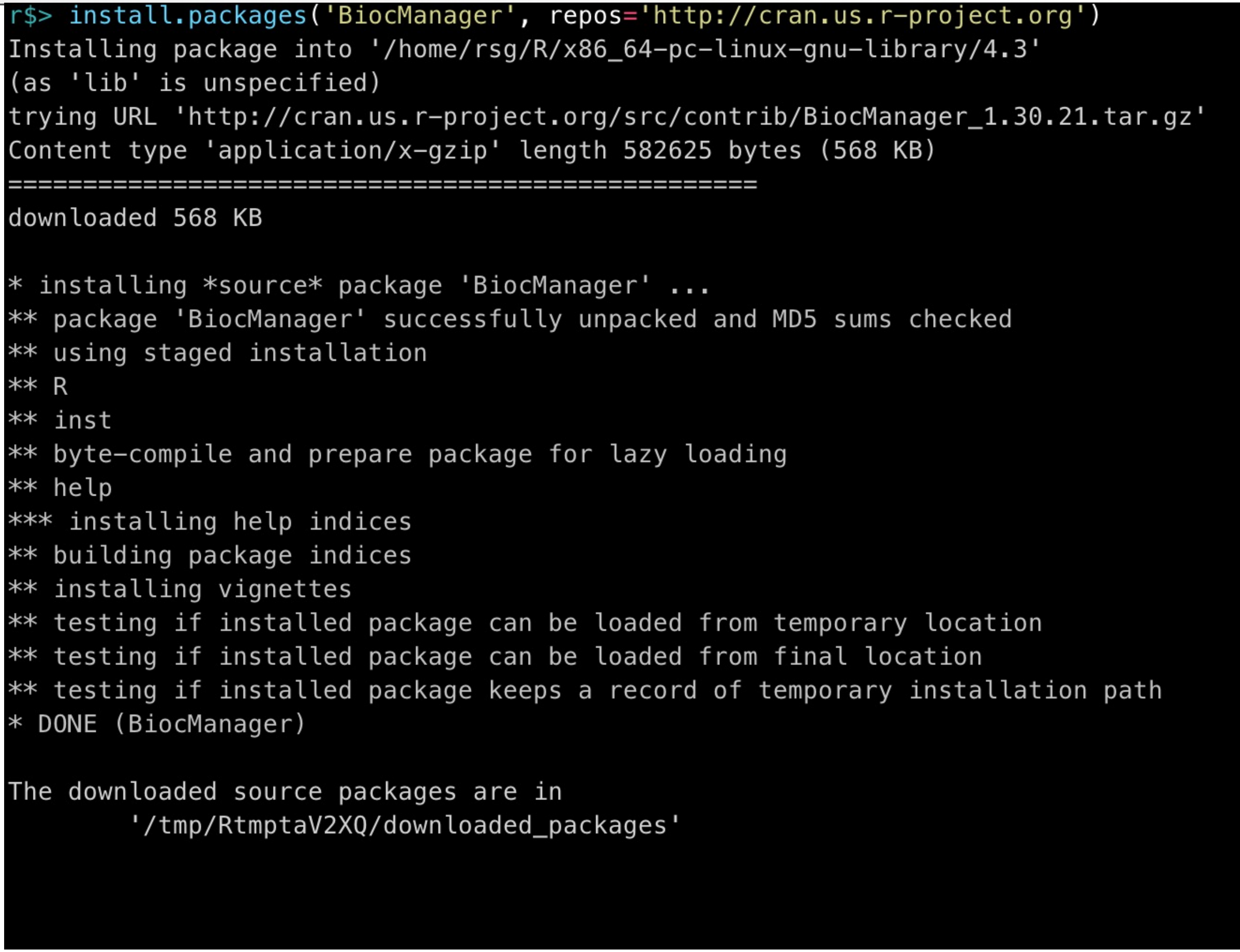
Bioconductor

The mission of the Bioconductor project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Scientific, Technical and Community Advisory Boards provide project oversight.

Bioconductor installation

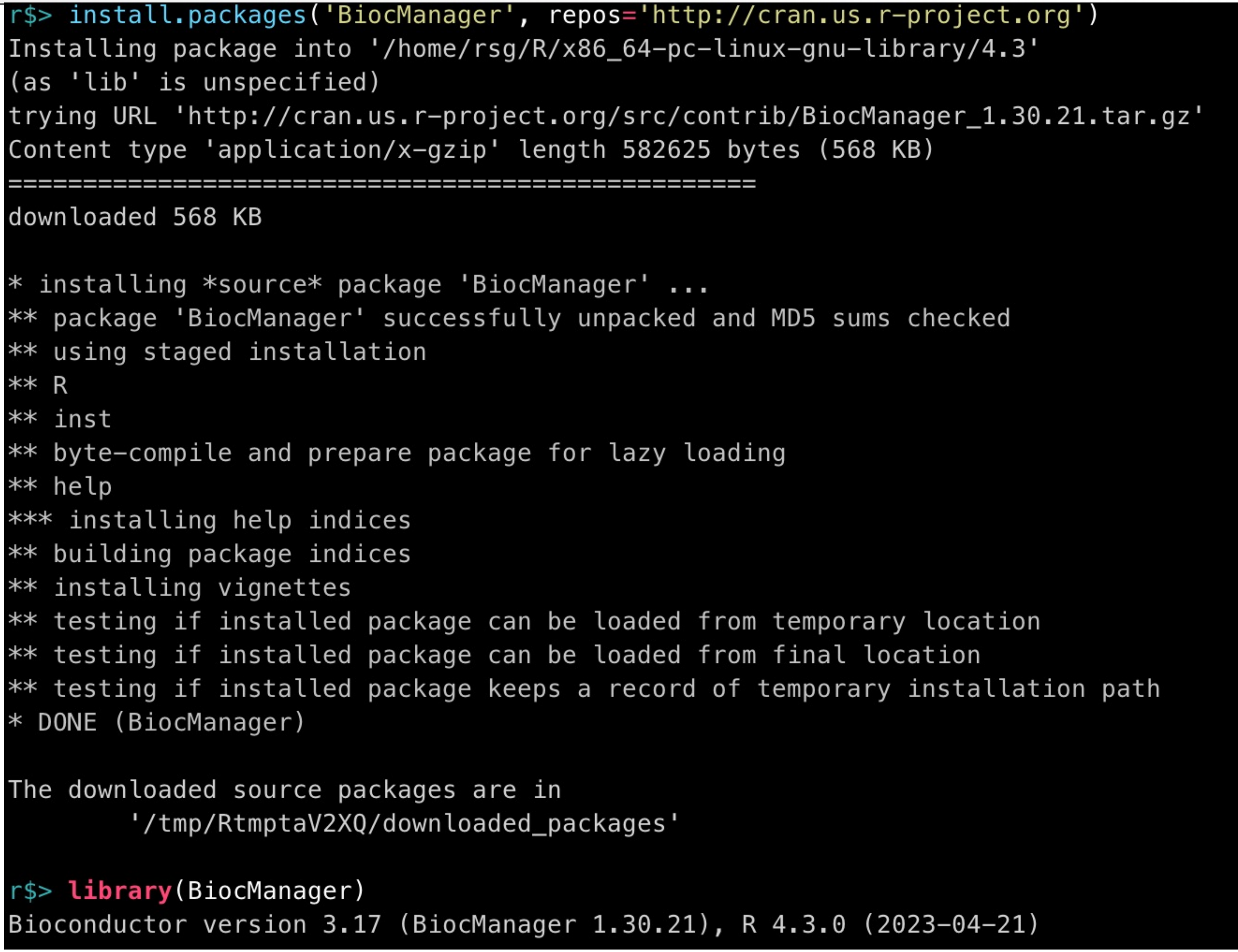
- As a package



```
r$> install.packages('BiocManager', repos='http://cran.us.r-project.org')
Installing package into '/home/rsg/R/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
trying URL 'http://cran.us.r-project.org/src/contrib/BiocManager_1.30.21.tar.gz'
Content type 'application/x-gzip' length 582625 bytes (568 KB)
==================================================
downloaded 568 KB

* installing *source* package 'BiocManager' ...
** package 'BiocManager' successfully unpacked and MD5 sums checked
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (BiocManager)

The downloaded source packages are in
        '/tmp/RtmptaV2XQ/downloaded_packages'
```

Bioconductor installation

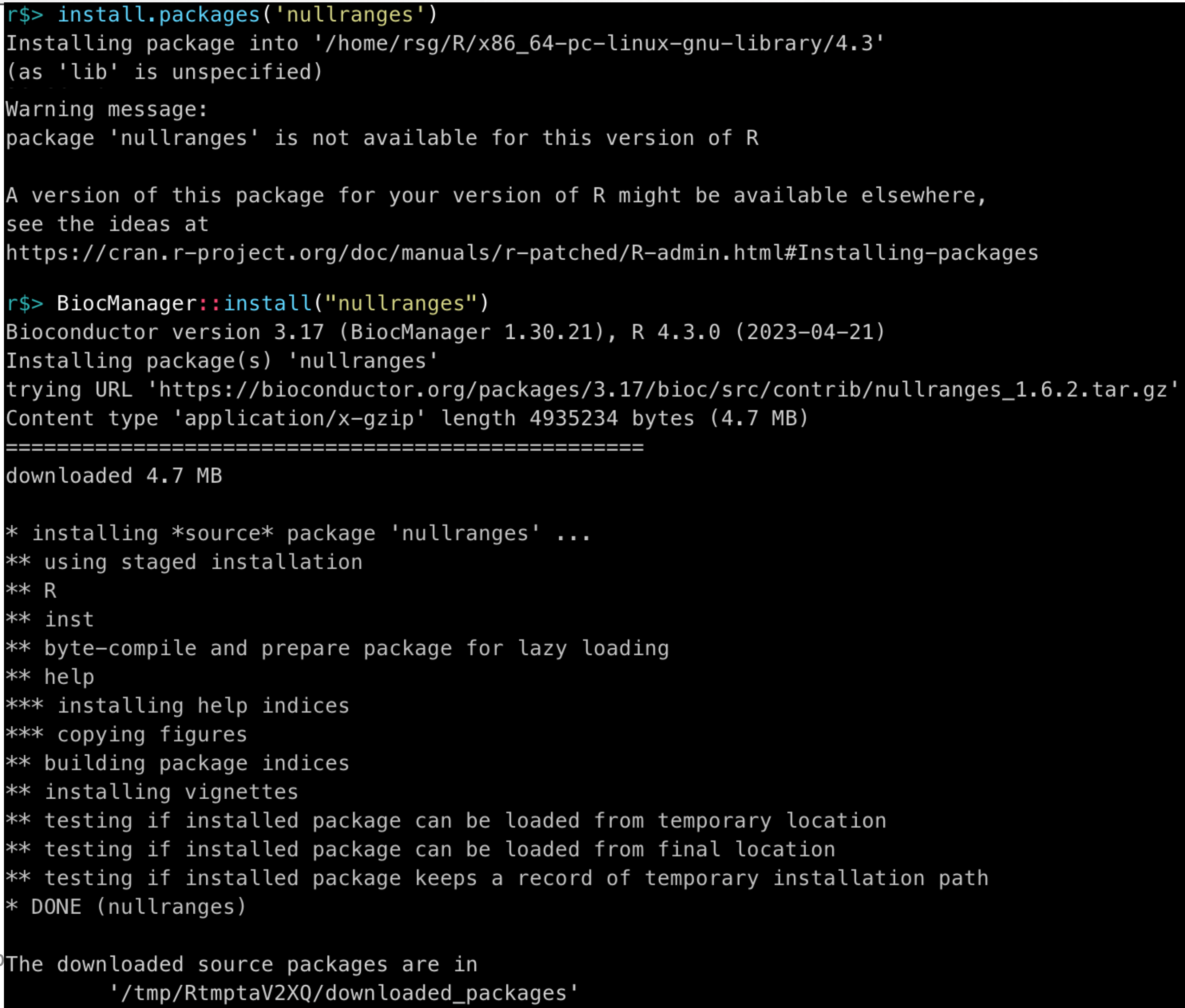
- As a package
- Integrated in R
- Bioconductor's version depends on your R version
- Some Bioc packages are restricted to a certain version!



```
r$> install.packages('BiocManager', repos='http://cran.us.r-project.org')
Installing package into '/home/rsg/R/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
trying URL 'http://cran.us.r-project.org/src/contrib/BiocManager_1.30.21.tar.gz'
Content type 'application/x-gzip' length 582625 bytes (568 KB)
==================================================
downloaded 568 KB

* installing *source* package 'BiocManager' ...
** package 'BiocManager' successfully unpacked and MD5 sums checked
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (BiocManager)

The downloaded source packages are in
        '/tmp/RtmptaV2XQ/downloaded_packages'

r$> library(BiocManager)
Bioconductor version 3.17 (BiocManager 1.30.21), R 4.3.0 (2023-04-21)
```

Bioconductor packages

- Bioconductor packages are on Bioconductor, not CRAN
- So you install them using Bioconductor’s BiocManager!



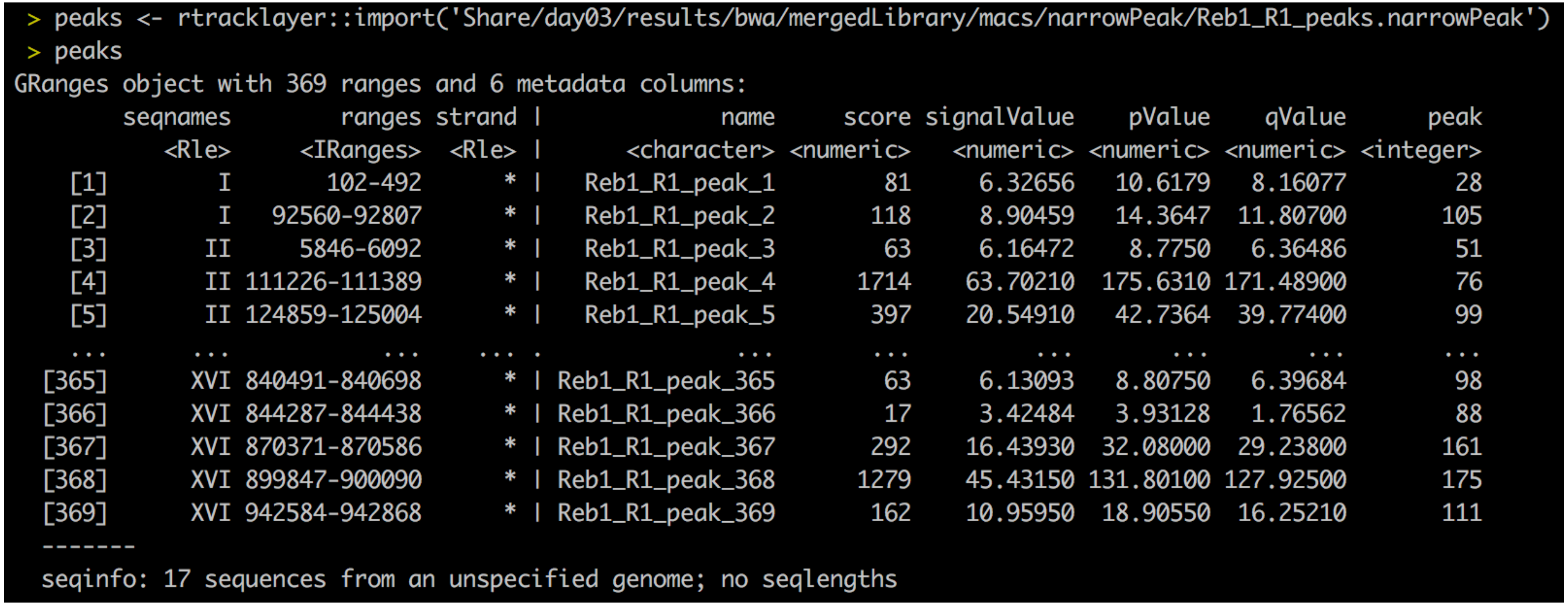
```
r$> install.packages('nullranges')
Installing package into '/home/rsg/R/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
Warning message:
package 'nullranges' is not available for this version of R

A version of this package for your version of R might be available elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages

r$> BiocManager::install("nullranges")
Bioconductor version 3.17 (BiocManager 1.30.21), R 4.3.0 (2023-04-21)
Installing package(s) 'nullranges'
trying URL 'https://bioconductor.org/packages/3.17/bioc/src/contrib/nullranges_1.6.2.tar.gz'
Content type 'application/x-gzip' length 4935234 bytes (4.7 MB)
==================================================
downloaded 4.7 MB

* installing *source* package 'nullranges' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
*** copying figures
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (nullranges)

The downloaded source packages are in
        '/tmp/RtmptaV2XQ/downloaded_packages'
```

Bioconductor essentials

- **GRanges** (through GenomicRanges package)
- **XNAStrings** (through Biostrings)
- **Import/export** from/to common genomic files (through BiocIO and rtracklayer packages)

GRanges

- **Workhorse class of Bioconductor**
- Used to describe <u>genomic intervals</u>

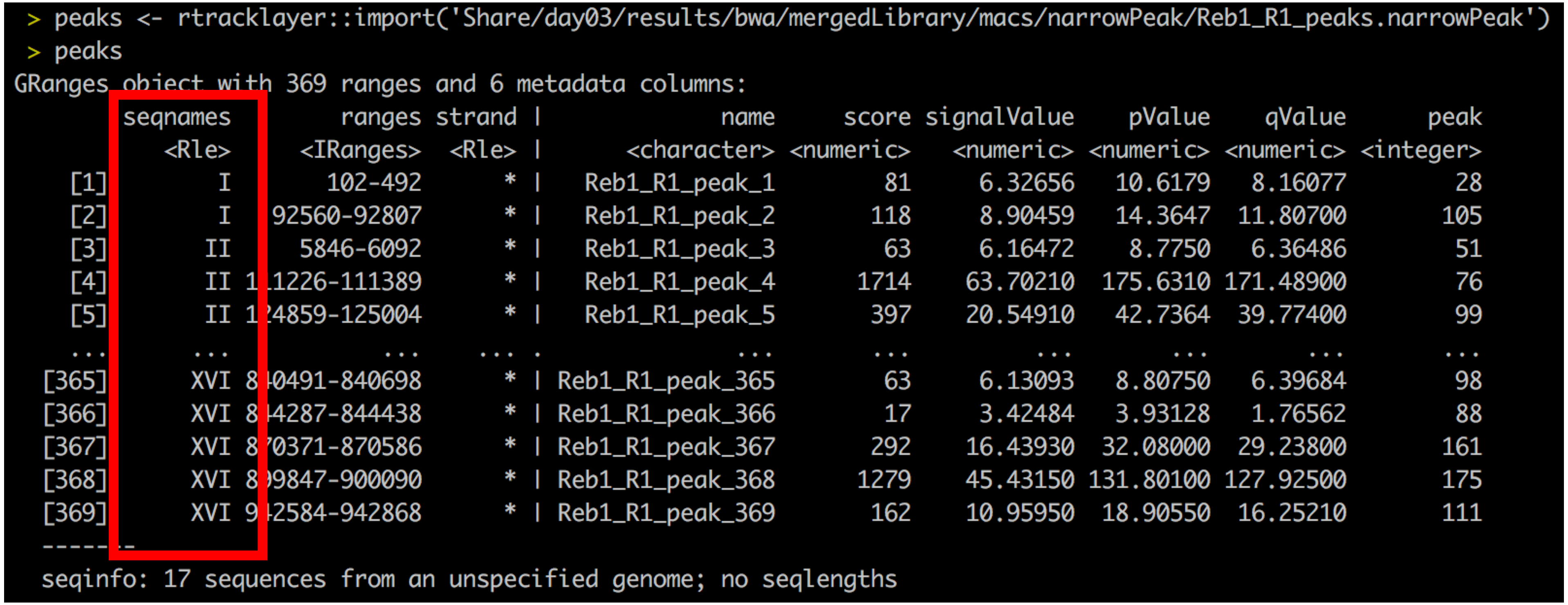


```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |             name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |      <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .              ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- Workhorse class of Bioconductor
- Used to describe <u>genomic intervals</u>

`seqnames(x)` -> chromosome names

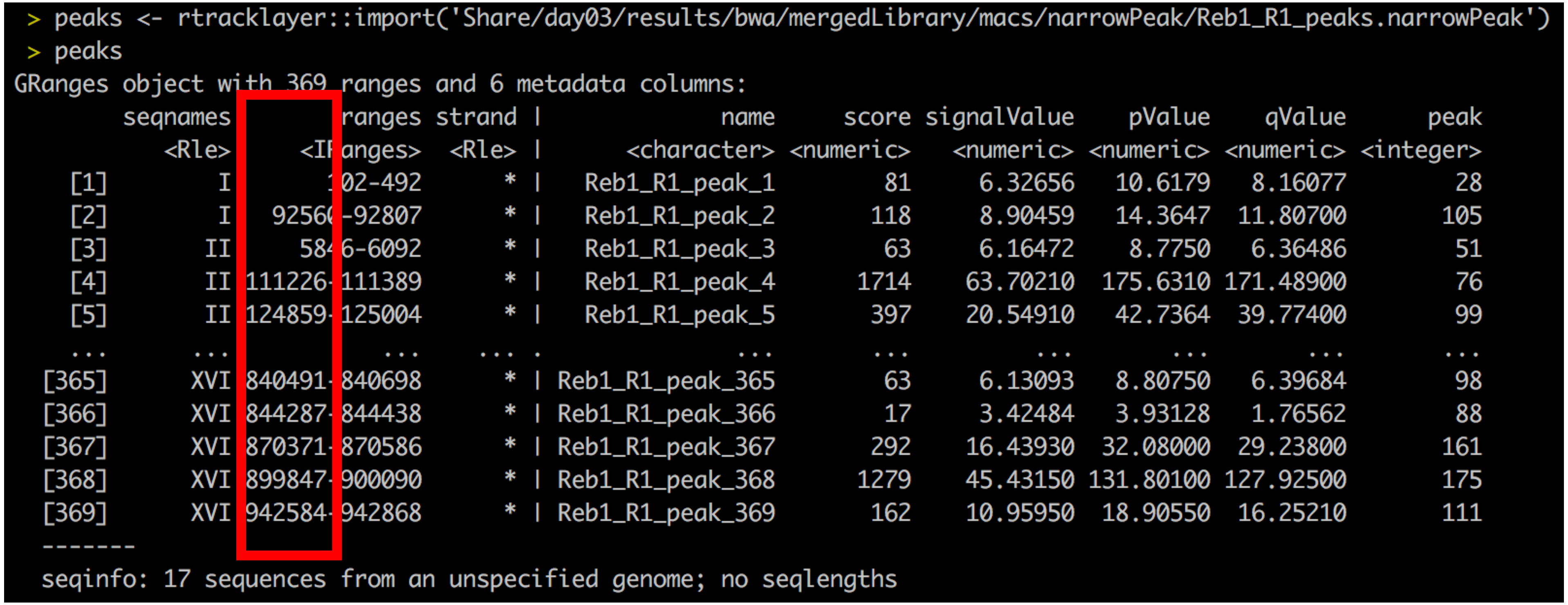


```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |            name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |     <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .             ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- **Workhorse class of Bioconductor**
- Used to describe <u>genomic intervals</u>

`start(x)` -> interval start



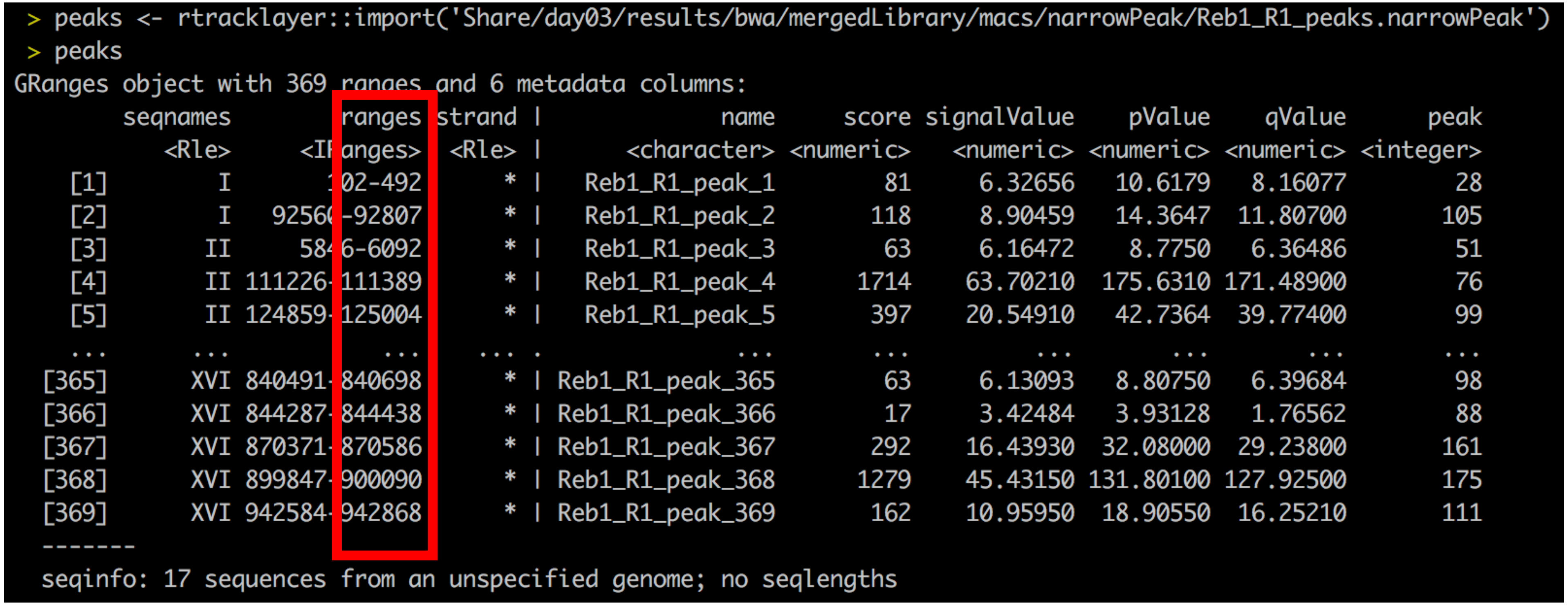
```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |            name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |     <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .             ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- **Workhorse class of Bioconductor**

- Used to describe <u>genomic intervals</u>

`end(x)` -> interval end



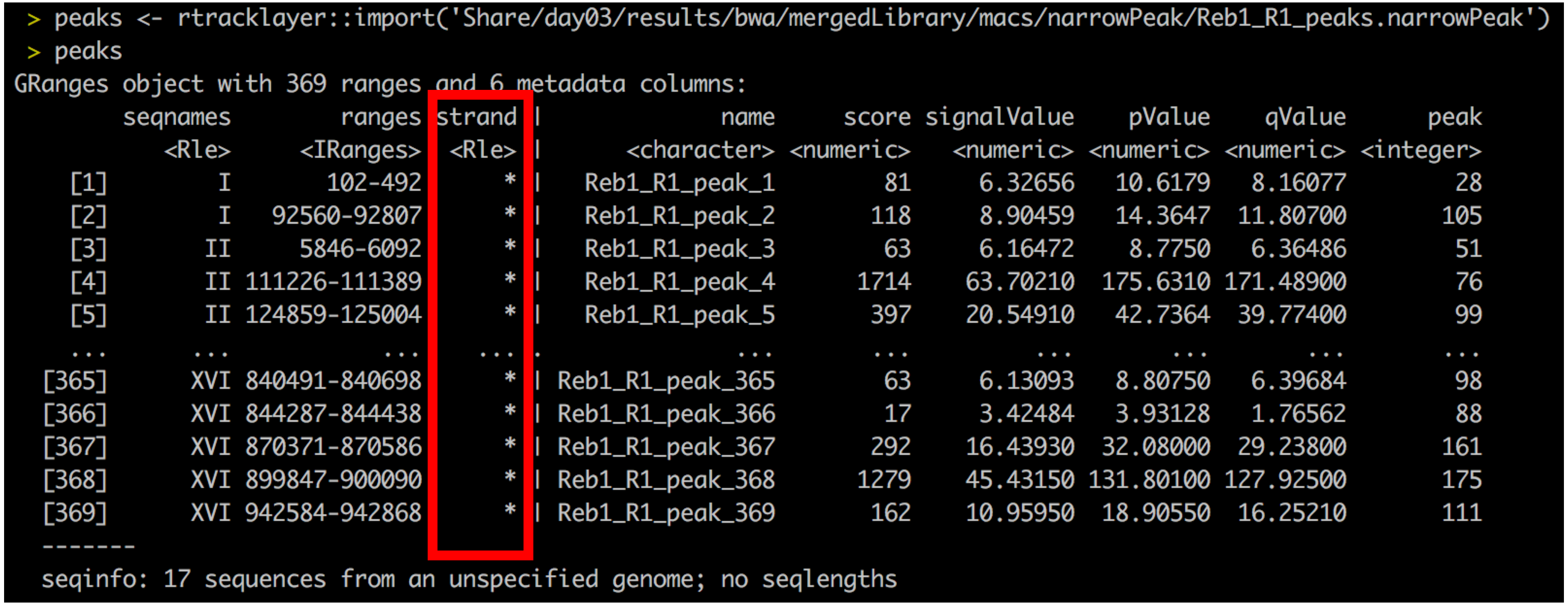
```
 > peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
 > peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |            name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |     <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .              ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- Workhorse class of Bioconductor

- Used to describe <u>genomic intervals</u>

`strand(x)`

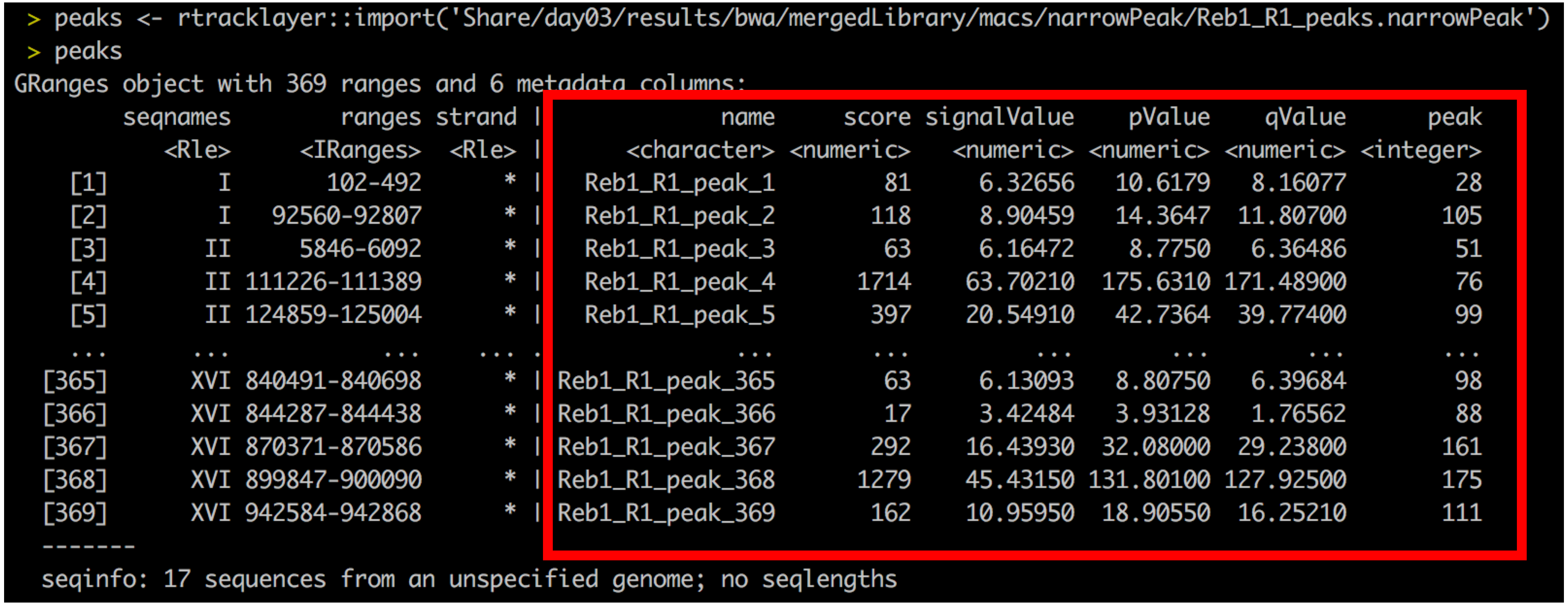


```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames          ranges strand |              name     score signalValue    pValue    qValue      peak
           <Rle>       <IRanges>  <Rle> |       <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I         102-492      * |    Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I     92560-92807      * |    Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II       5846-6092      * |    Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II   111226-111389      * |    Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II   124859-125004      * |    Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...             ...    ... .               ...       ...         ...       ...       ...       ...
  [365]      XVI   840491-840698      * |  Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI   844287-844438      * |  Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI   870371-870586      * |  Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI   899847-900090      * |  Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI   942584-942868      * |  Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- **Workhorse class of Bioconductor**
- Used to describe <u>genomic intervals</u>

`mcols(x)` -> all metadata



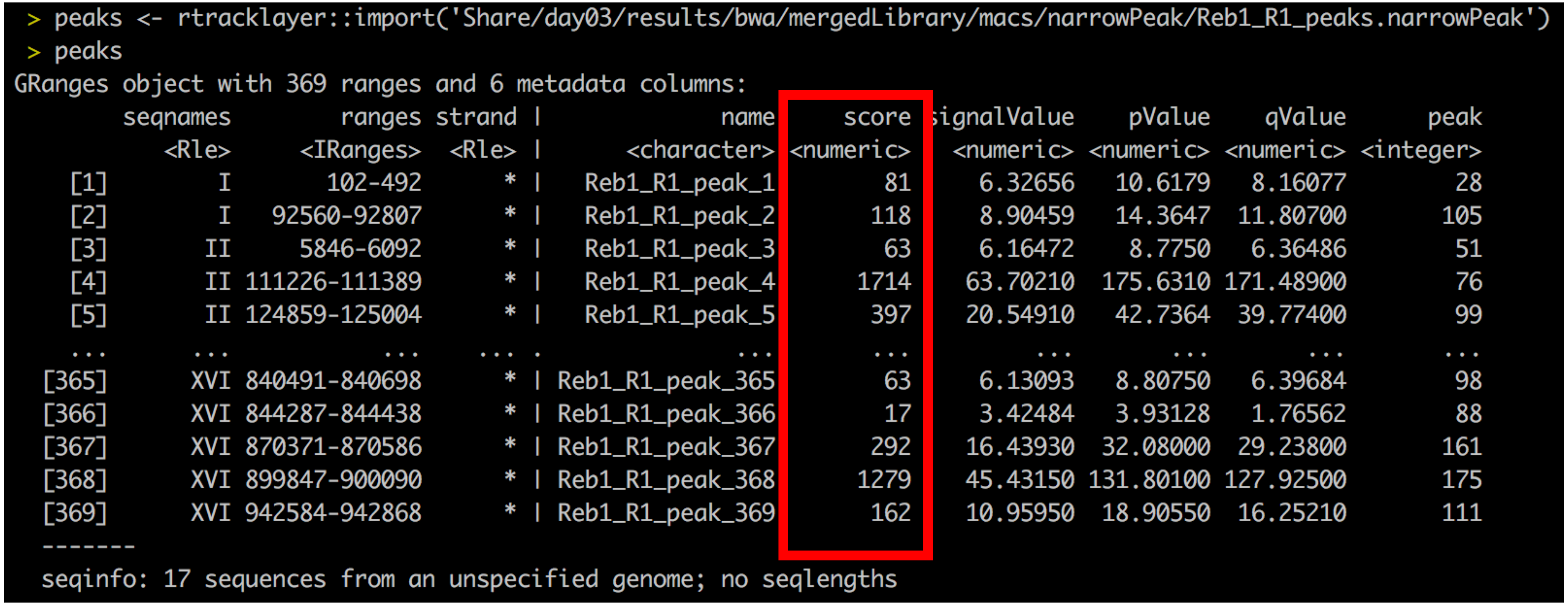
```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames          ranges strand |            name     score signalValue    pValue    qValue      peak
           <Rle>       <IRanges>  <Rle> |     <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I         102-492      * |  Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I     92560-92807      * |  Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II       5846-6092      * |  Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II   111226-111389      * |  Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II   124859-125004      * |  Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...             ...    ... .             ...       ...         ...       ...       ...       ...
  [365]      XVI   840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI   844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI   870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI   899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI   942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- Workhorse class of Bioconductor

- Used to describe <u>genomic intervals</u>

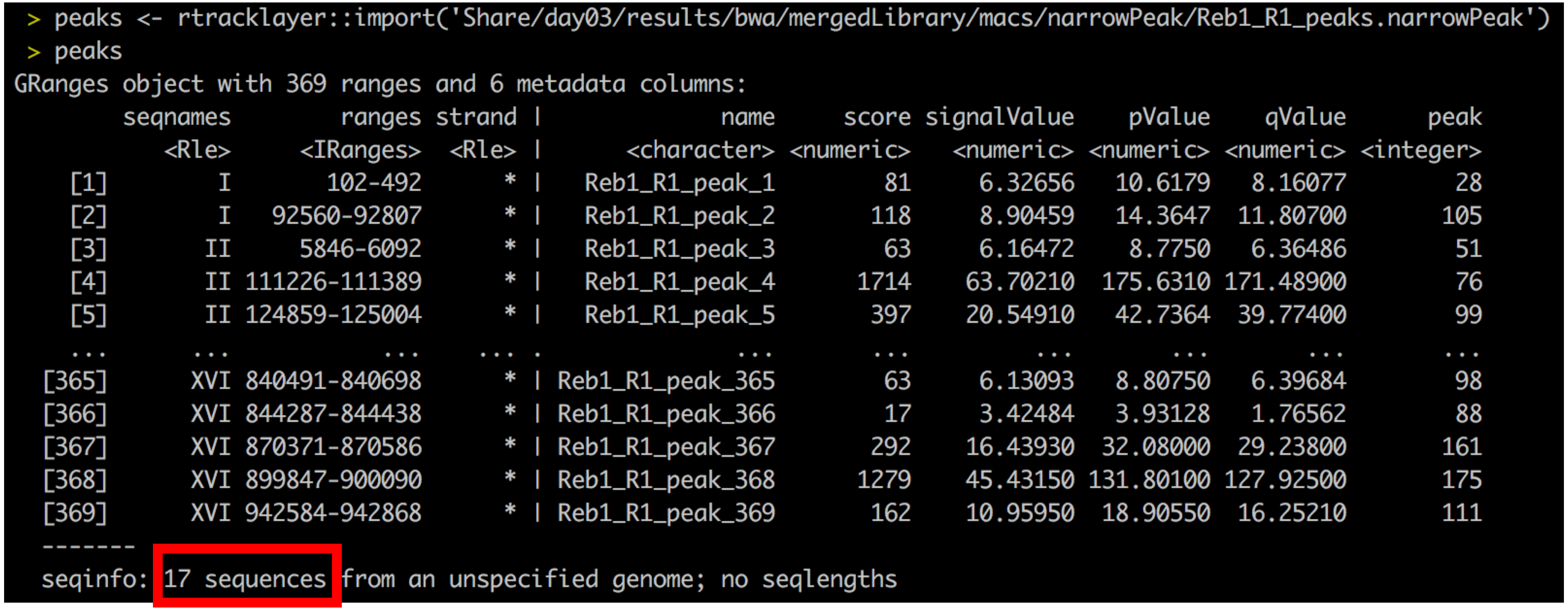
`x$score`
`mcols(x)$score` -> specific metadata



```
> peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
> peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |              name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |       <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .               ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

GRanges

- Workhorse class of Bioconductor
- Used to describe <u>genomic intervals</u>



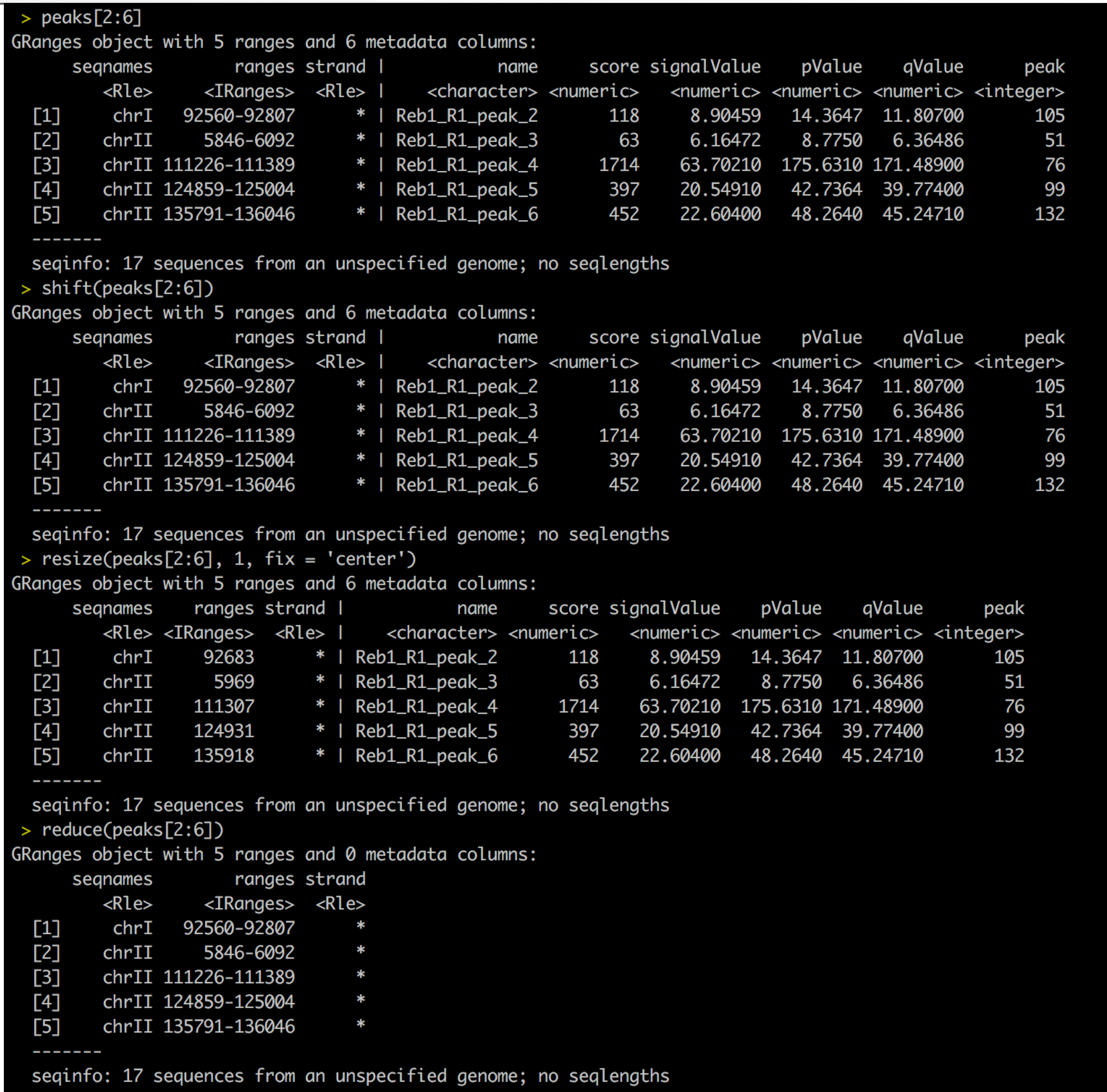
```
 > peaks <- rtracklayer::import('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak')
 > peaks
GRanges object with 369 ranges and 6 metadata columns:
        seqnames        ranges strand |             name     score signalValue    pValue    qValue      peak
           <Rle>     <IRanges>  <Rle> |      <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I       102-492      * |   Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]        I   92560-92807      * |   Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
    [3]       II     5846-6092      * |   Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [4]       II 111226-111389      * |   Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [5]       II 124859-125004      * |   Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    ...      ...           ...    ... .              ...       ...         ...       ...       ...       ...
  [365]      XVI 840491-840698      * | Reb1_R1_peak_365        63     6.13093   8.80750   6.39684        98
  [366]      XVI 844287-844438      * | Reb1_R1_peak_366        17     3.42484   3.93128   1.76562        88
  [367]      XVI 870371-870586      * | Reb1_R1_peak_367       292    16.43930  32.08000  29.23800       161
  [368]      XVI 899847-900090      * | Reb1_R1_peak_368      1279    45.43150 131.80100 127.92500       175
  [369]      XVI 942584-942868      * | Reb1_R1_peak_369       162    10.95950  18.90550  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

`seqinfo(x)` -> genome information

Granges operators

Action functions

- `…[…] (to subset)`
- `shift()`
- `resize()`
- `reduce()`
- `coverage()`
- …

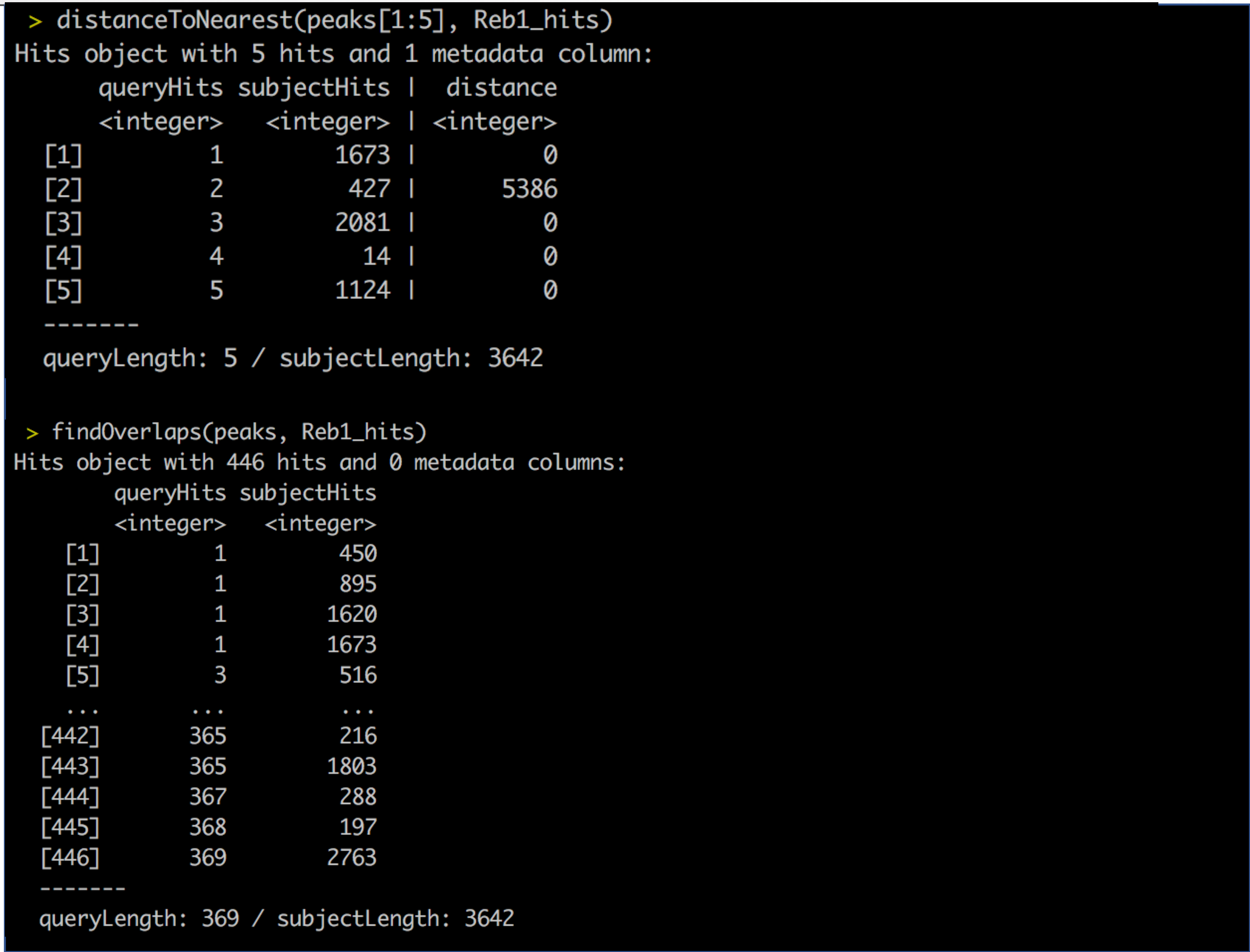


```
 > peaks[2:6]
GRanges object with 5 ranges and 6 metadata columns:
      seqnames          ranges strand |           name     score signalValue    pValue    qValue      peak
         <Rle>       <IRanges>  <Rle> |    <character> <numeric>   <numeric> <numeric> <numeric> <integer>
  [1]     chrI   92560-92807      * | Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
  [2]    chrII     5846-6092      * | Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
  [3]    chrII 111226-111389      * | Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
  [4]    chrII 124859-125004      * | Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
  [5]    chrII 135791-136046      * | Reb1_R1_peak_6       452    22.60400   48.2640  45.24710       132
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
 > shift(peaks[2:6])
GRanges object with 5 ranges and 6 metadata columns:
      seqnames          ranges strand |           name     score signalValue    pValue    qValue      peak
         <Rle>       <IRanges>  <Rle> |    <character> <numeric>   <numeric> <numeric> <numeric> <integer>
  [1]     chrI   92560-92807      * | Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
  [2]    chrII     5846-6092      * | Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
  [3]    chrII 111226-111389      * | Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
  [4]    chrII 124859-125004      * | Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
  [5]    chrII 135791-136046      * | Reb1_R1_peak_6       452    22.60400   48.2640  45.24710       132
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
 > resize(peaks[2:6], 1, fix = 'center')
GRanges object with 5 ranges and 6 metadata columns:
      seqnames    ranges strand |           name     score signalValue    pValue    qValue      peak
         <Rle> <IRanges>  <Rle> |    <character> <numeric>   <numeric> <numeric> <numeric> <integer>
  [1]     chrI     92683      * | Reb1_R1_peak_2       118     8.90459   14.3647  11.80700       105
  [2]    chrII      5969      * | Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
  [3]    chrII    111307      * | Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
  [4]    chrII    124931      * | Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
  [5]    chrII    135918      * | Reb1_R1_peak_6       452    22.60400   48.2640  45.24710       132
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
 > reduce(peaks[2:6])
GRanges object with 5 ranges and 0 metadata columns:
      seqnames          ranges strand
         <Rle>       <IRanges>  <Rle>
  [1]     chrI   92560-92807      *
  [2]    chrII     5846-6092      *
  [3]    chrII 111226-111389      *
  [4]    chrII 124859-125004      *
  [5]    chrII 135791-136046      *
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

Granges operators

Comparison functions

- `%over%`
- `distance()`
- `distanceToNearest()`
- `findOverlaps()`
- `subsetByOverlaps()`

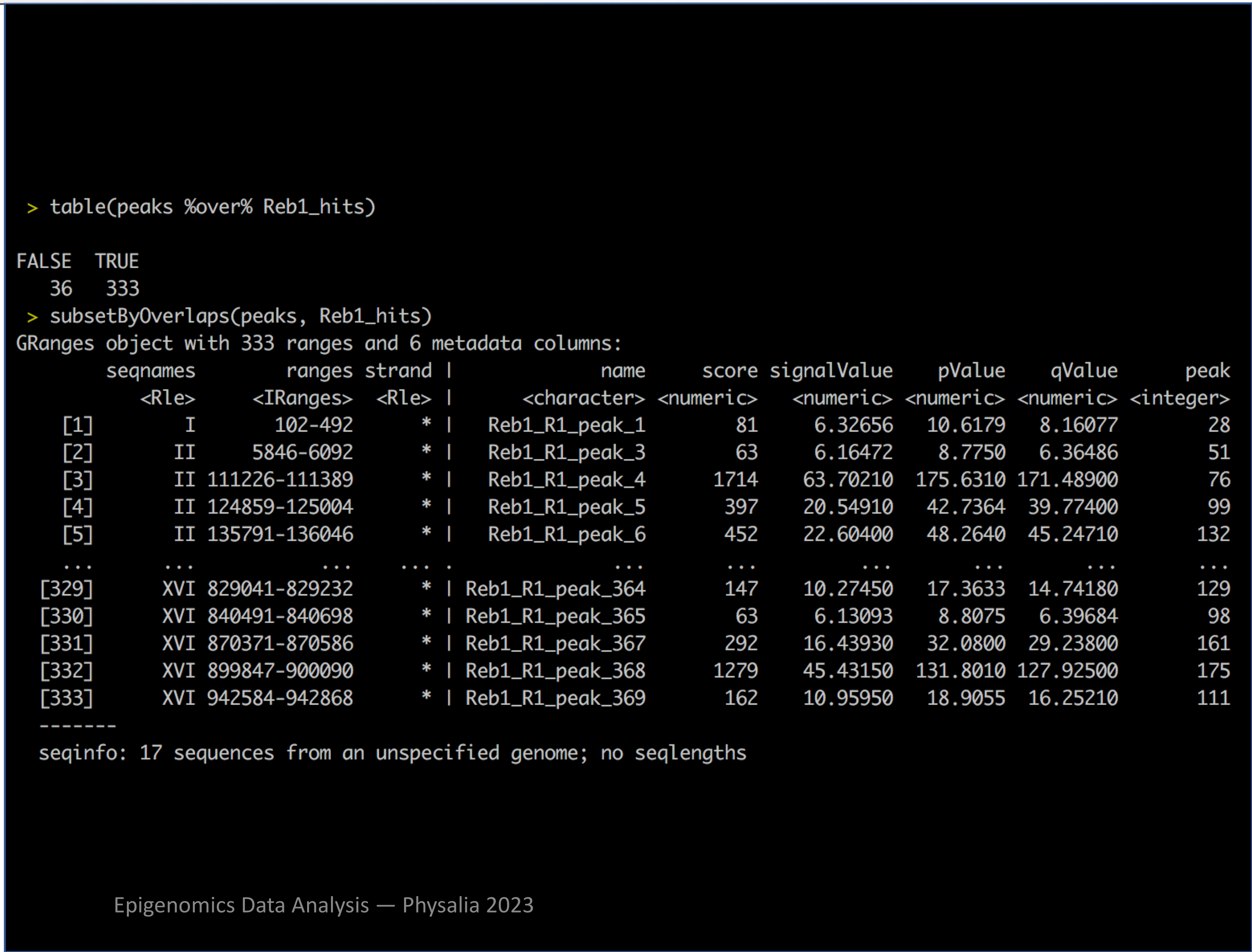


```
> distanceToNearest(peaks[1:5], Reb1_hits)
Hits object with 5 hits and 1 metadata column:
      queryHits subjectHits |  distance
      <integer>   <integer> | <integer>
  [1]         1        1673 |         0
  [2]         2         427 |      5386
  [3]         3        2081 |         0
  [4]         4          14 |         0
  [5]         5        1124 |         0
  -------
  queryLength: 5 / subjectLength: 3642

> findOverlaps(peaks, Reb1_hits)
Hits object with 446 hits and 0 metadata columns:
        queryHits subjectHits
        <integer>   <integer>
    [1]         1         450
    [2]         1         895
    [3]         1        1620
    [4]         1        1673
    [5]         3         516
    ...       ...         ...
  [442]       365         216
  [443]       365        1803
  [444]       367         288
  [445]       368         197
  [446]       369        2763
  -------
  queryLength: 369 / subjectLength: 3642
```

Granges operators

Comparison functions

- `%over%`
- `distance()`
- `distanceToNearest()`
- `findOverlaps()`
- `subsetByOverlaps()`



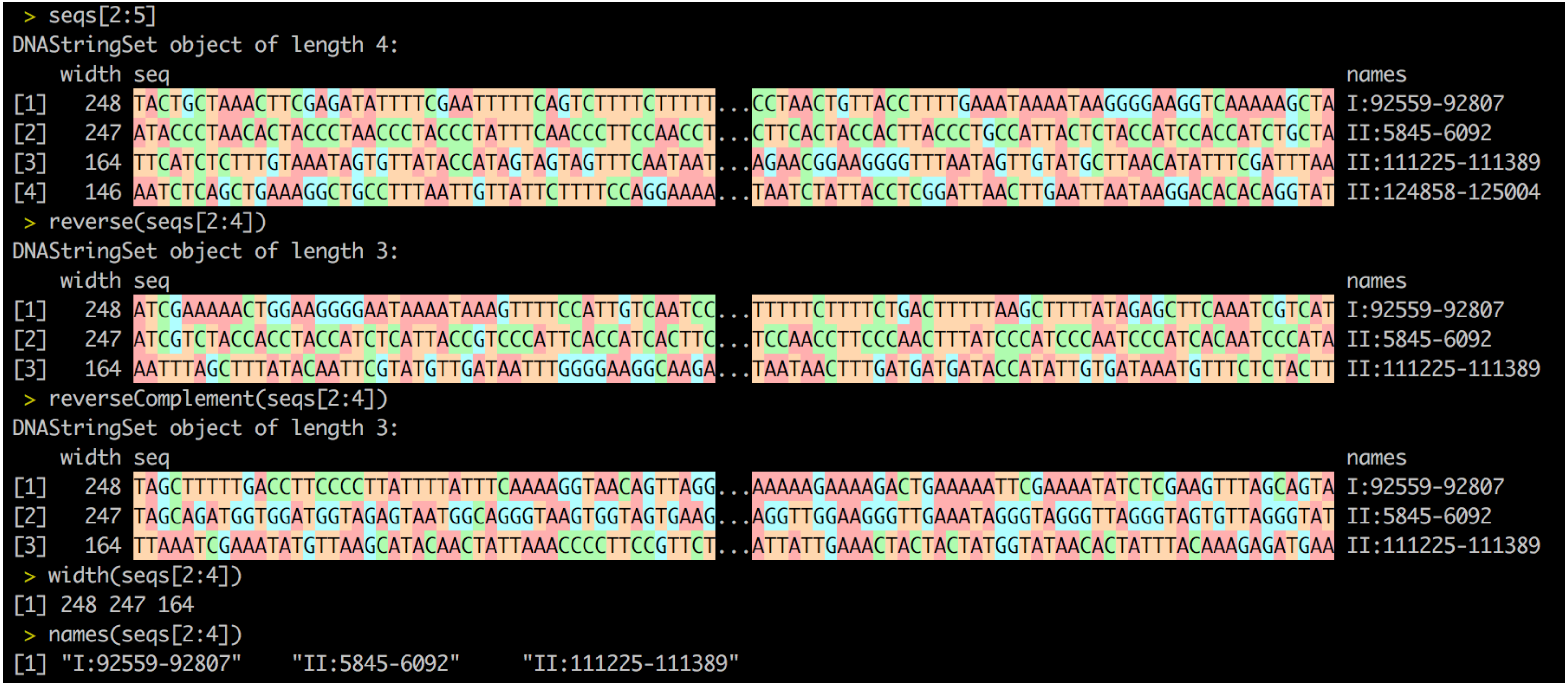
```
 > table(peaks %over% Reb1_hits)

FALSE  TRUE
   36   333
 > subsetByOverlaps(peaks, Reb1_hits)
GRanges object with 333 ranges and 6 metadata columns:
        seqnames          ranges strand |                name     score signalValue    pValue    qValue      peak
           <Rle>       <IRanges>  <Rle> |         <character> <numeric>   <numeric> <numeric> <numeric> <integer>
    [1]        I         102-492      * |     Reb1_R1_peak_1        81     6.32656   10.6179   8.16077        28
    [2]       II       5846-6092      * |     Reb1_R1_peak_3        63     6.16472    8.7750   6.36486        51
    [3]       II   111226-111389      * |     Reb1_R1_peak_4      1714    63.70210  175.6310 171.48900        76
    [4]       II   124859-125004      * |     Reb1_R1_peak_5       397    20.54910   42.7364  39.77400        99
    [5]       II   135791-136046      * |     Reb1_R1_peak_6       452    22.60400   48.2640  45.24710       132
    ...      ...             ...    ... .                 ...       ...         ...       ...       ...       ...
  [329]      XVI   829041-829232      * |   Reb1_R1_peak_364       147    10.27450   17.3633  14.74180       129
  [330]      XVI   840491-840698      * |   Reb1_R1_peak_365        63     6.13093    8.8075   6.39684        98
  [331]      XVI   870371-870586      * |   Reb1_R1_peak_367       292    16.43930   32.0800  29.23800       161
  [332]      XVI   899847-900090      * |   Reb1_R1_peak_368      1279    45.43150  131.8010 127.92500       175
  [333]      XVI   942584-942868      * |   Reb1_R1_peak_369       162    10.95950   18.9055  16.25210       111
  -------
  seqinfo: 17 sequences from an unspecified genome; no seqlengths
```

Biostrings

Biostrings in R

```
> seqs <- Biostrings::readDNAStringSet('Share/day03/results/bwa/mergedLibrary/macs/narrowPeak/Reb1_R1_peaks.narrowPeak.fa')
> seqs
DNAStringSet object of length 369:
      width seq                                                                                               names
  [1]   391 CCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTT...ATATACCATCTCAAACTTACCCTACTCTCAGATTCCACTTCACTCCA I:101-492
  [2]   248 TACTGCTAAACTTCGAGATATTTTCGAATTTTTCAGTCTTTTCTTTT...CTAACTGTTACCTTTTGAAATAAAATAAGGGGAAGGTCAAAAAGCTA I:92559-92807
  [3]   247 ATACCCTAACACTACCCTAACCCTACCCTATTTCAACCCTTCCAACC...TTCACTACCACTTACCCTGCCATTACTCTACCATCCACCATCTGCTA II:5845-6092
  [4]   164 TTCATCTCTTTGTAAATAGTGTTATACCATAGTAGTAGTTTCAATAA...GAACGGAAGGGGTTTAATAGTTGTATGCTTAACATATTTCGATTTAA II:111225-111389
  [5]   146 AATCTCAGCTGAAAGGCTGCCTTTAATTGTTATTCTTTTCCAGGAAA...AATCTATTACCTCGGATTAACTTGAATTAATAAGGACACACAGGTAT II:124858-125004
  ...   ... ...
[365]   208 ACTTACTGGTCTTTAGCACACGACGACCGTACTTTGCACGTGGCTGC...TTCAGACCCACACAAAATCCGCGTAGCCGAGATTGCTTATGTATGTT XVI:840490-840698
[366]   152 AAGGGGTATGTTCCTCAGCATTATCTGAAGGTACTCCTCTAAATTTT...ATAATATCAGGTAAAGAAATTGTTGGAATAAAAATCCACTATCGTCT XVI:844286-844438
[367]   216 AGGAAAAAAGGAAAAAGCAAAAAATATCGATTTTTATGACTTACAAA...TACCCGCATATTATCGGGAAACAGAAGCCATGTTAGAGTGATTTCCA XVI:870370-870586
[368]   244 TAGTCGTCGCAAGCGACAAATCTCAACTGACAGTAAATAACGGTGAT...TTCTTGTTCCACCTCTTTTCCCCAACATATATGAACATGAGATGGTA XVI:899846-900090
[369]   285 TGGGTGAATGGCACAGGGTATAGACCGCTGAGGCAAGTGCCGTGCAT...GAAGCGTGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTT XVI:942583-942868
```

Biostrings



```
 > seqs[2:5]
DNAStringSet object of length 4:
    width seq                                                                                                 names
[1]   248 TACTGCTAAACTTCGAGATATTTTCGAATTTTTCAGTCTTTTCTTTTT...CCTAACTGTTACCTTTTGAAATAAAATAAGGGGAAGGTCAAAAAGCTA I:92559-92807
[2]   247 ATACCCTAACACTACCCTAACCCTACCCTATTTCAACCCTTCCAACCT...CTTCACTACCACTTACCCTGCCATTACTCTACCATCCACCATCTGCTA II:5845-6092
[3]   164 TTCATCTCTTTGTAAATAGTGTTATACCATAGTAGTAGTTTCAATAAT...AGAACGGAAGGGGTTTAATAGTTGTATGCTTAACATATTTCGATTTAA II:111225-111389
[4]   146 AATCTCAGCTGAAAGGCTGCCTTTAATTGTTATTCTTTTCCAGGAAAA...TAATCTATTACCTCGGATTAACTTGAATTAATAAGGACACACAGGTAT II:124858-125004
 > reverse(seqs[2:4])
DNAStringSet object of length 3:
    width seq                                                                                                 names
[1]   248 ATCGAAAAACTGGAAGGGGAATAAAATAAAGTTTTCCATTGTCAATCC...TTTTTCTTTTCTGACTTTTTAAGCTTTTATAGAGCTTCAAATCGTCAT I:92559-92807
[2]   247 ATCGTCTACCACCTACCATCTCATTACCGTCCCATTCACCATCACTTC...TCCAACCTTCCCAACTTTATCCCATCCCAATCCCATCACAATCCCATA II:5845-6092
[3]   164 AATTTAGCTTTATACAATTCGTATGTTGATAATTTGGGGAAGGCAAGA...TAATAACTTTGATGATGATACCATATTGTGATAAATGTTTCTCTACTT II:111225-111389
 > reverseComplement(seqs[2:4])
DNAStringSet object of length 3:
    width seq                                                                                                 names
[1]   248 TAGCTTTTTGACCTTCCCCTTATTTTATTTCAAAAGGTAACAGTTAGG...AAAAAGAAAAGACTGAAAAATTCGAAAATATCTCGAAGTTTAGCAGTA I:92559-92807
[2]   247 TAGCAGATGGTGGATGGTAGAGTAATGGCAGGGTAAGTGGTAGTGAAG...AGGTTGGAAGGGTTGAAATAGGGTAGGGTTAGGGTAGTGTTAGGGTAT II:5845-6092
[3]   164 TTAAATCGAAATATGTTAAGCATACAACTATTAAACCCCTTCCGTTCT...ATTATTGAAACTACTACTATGGTATAACACTATTTACAAAGAGATGAA II:111225-111389
 > width(seqs[2:4])
[1] 248 247 164
 > names(seqs[2:4])
[1] "I:92559-92807"    "II:5845-6092"     "II:111225-111389"
```