INTRODUCTION TO BIOCONDUCTOR

The Bioconductor Project

Paula Andrea Martinez, PhD.
Data Scientist

Bioconductor

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

[1] Bioconductor (www.bioconductor.org)

What do we measure and why?

- **Structure**: elements, regions, size, order, relationships

- **Function**: expression, levels, regulation, phenotypes

How to install Bioconductor packages?

- Biconductor has its own repository, way to install packages, and each release is designed to work with a specific version of R.
    - For this course, you'll be using Bioconductor version 3.6.
- Bioconductor version 3.7 or earlier uses **BiocLite**:

```
source("https://bioconductor.org/biocLite.R")
biocLite("packageName")
```

- Bioconductor version 3.8 and later uses **BiocManager**:

```
if (!requireNamespace("BiocManager"))
    install.packages("BiocManager")
BiocManager::install()
```

Bioconductor version and package version

- BiocInstaller works for Bioconductor version 3.7 or earlier

```
# Check Bioconductor version (For versions <= 3.7)
BiocInstaller::biocVersion()
# or
biocVersion()

# Load a package
library(packageName)

# Check versions for reproducibility
sessionInfo()
# or
packageVersion("packageName")

# Check package updates (Bioconductor version <= 3.7)
BiocInstaller::biocValid()
# or
biocValid()
```

INTRODUCTION TO BIOCONDUCTOR

Let's practice!

INTRODUCTION TO BIOCONDUCTOR

The Role of S4 in Bioconductor

Paula Andrea Martinez, PhD.
Data Scientist

S3

Positive

- CRAN, simple but powerful
- Flexible and interactive
- Uses a generic function
- Functionality depends on the first argument
- Example: `plot()` and `methods(plot)`

Negative

- Bad at validating types and naming conventions (dot not dot?)
- Inheritance works, but depends on the input

S4

Positive

- Formal definition of classes
- Bioconductor reusability
- Has validation of types
- Naming conventions

Example: `mydescriptor <- new("GenomeDescription")`

Negative

- Complex structure compared to S3

Is it S4 or not?

Ask if an object is S4

```
isS4(mydescriptor)

[1] TRUE
```

`str` of S4 objects start with `Formal class`

```
str(mydescriptor)

Formal class 'GenomeDescription' [package "GenomeInfoDb"] with 7 slots
...
```

S4 class definition

A class describes a representation

- **name**
- **slots** (methods/fields)
- **contains** (inheritance definition)

Example

```
MyEpicProject <- setClass(# Define class name with UpperCamelCase
                          "MyEpicProject",
                          # Define slots, helpful for validation
                          slots = c(ini = "Date",
                                    end = "Date",
                                    milestone = "character"),
                          # Define inheritance
                          contains = "MyProject")
```

S4 Accesors

```
.S4methods(class = "GenomeDescription")
```

```
 [1] commonName       organism         provider         providerVersion
 [5] releaseDate      releaseName      seqinfo          seqnames
 [9] show             toString         bsgenomeName
```

```
showMethods(classes = "GenomeDescription", where = search())
```

Object summary

```
show(myDescriptor)
```

```
| organism:  ()
| provider:
| provider version:
| release date:
| release name:
| ---
| seqlengths:
```

INTRODUCTION TO BIOCONDUCTOR

Let's practice!

INTRODUCTION TO BIOCONDUCTOR

Introducing biology of genomic datasets

Paula Andrea Martinez, PhD.
Data Scientist

Chromosome
DNA helix
Nucleus
Cell
Base pairs
Sugar
phosphate
backbone
T
A
Thymine
Adenine
G
C
Guanine
Cytosine

Genome elements

- Genetic information DNA alphabet
- A set of chromosomes (highly variable number)
- Genes (carry heredity instructions)
  - coding and non-coding
- Proteins (responsible for specific functions)
  - DNA-to-RNA (transcription)
  - RNA-to-protein (translation)

Yeast

- A single cell microorganism
- The fungus that people love ❤
- Used for fermentation: beer, bread, kefir, kombucha, bioremediation, etc.
- Name: *Saccharomyces cerevisiae* or *S. cerevisiae*

Yeast genome

BSgenome annotation package

```
# load the package and store data into yeast
library(BSgenome.Scerevisiae.UCSC.sacCer3)
yeast <- BSgenome.Scerevisiae.UCSC.sacCer3

#interested in other genomes?
available.genomes()
```

Using accessors

```
# Chromosome number
length(yeast)

# Chromosome names
names(yeast)

# Sequence lengths
seqlengths(yeast)
```

Get sequences

S4 method for BSgenome

```
# S4 method getSeq() requires a BSgenome object
getSeq(yeast)

# Select chromosome sequence by name, one or many
getSeq(yeast, "chrM")

# Select start, end and or width
# end = 10, selects first 10 base pairs of each chromosome
getSeq(yeast, end = 10)
```

INTRODUCTION TO BIOCONDUCTOR

Let's practice!