The Bioconductor Project

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

datacamp

Bioconductor

[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?

- Bioconductor has its own repository and way to install packages

```
# Install BiocManager
install.packages("BiocManager")

# Install the GenomicRanges package
BiocManager::install("GenomicRanges")
```

Bioconductor version and package version

```
# Load BiocManager
library(BiocManager)


# Check Bioconductor version
version()
```

Checking package versions

```
# Load GenomicRanges
library(GenomicRanges)


# Check versions for reproducibility
sessionInfo()
```

Package updates

```
# Check for package updates
valid()
```

Let's practice!

INTRODUCTION TO BIOCONDUCTOR IN R

The Role of S4 in Bioconductor

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

datacamp

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)
```

```
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)

```
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")
```

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

```
 [1] coerce          commonName      organism        provider
 [5] providerVersion releaseDate     releaseName     seqinfo
 [9] seqnames        show            toString        bsgenomeName
see '?methods' for accessing help and source code
```

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

Object summary

```
show(myDescriptor)
```

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

Let's practice!

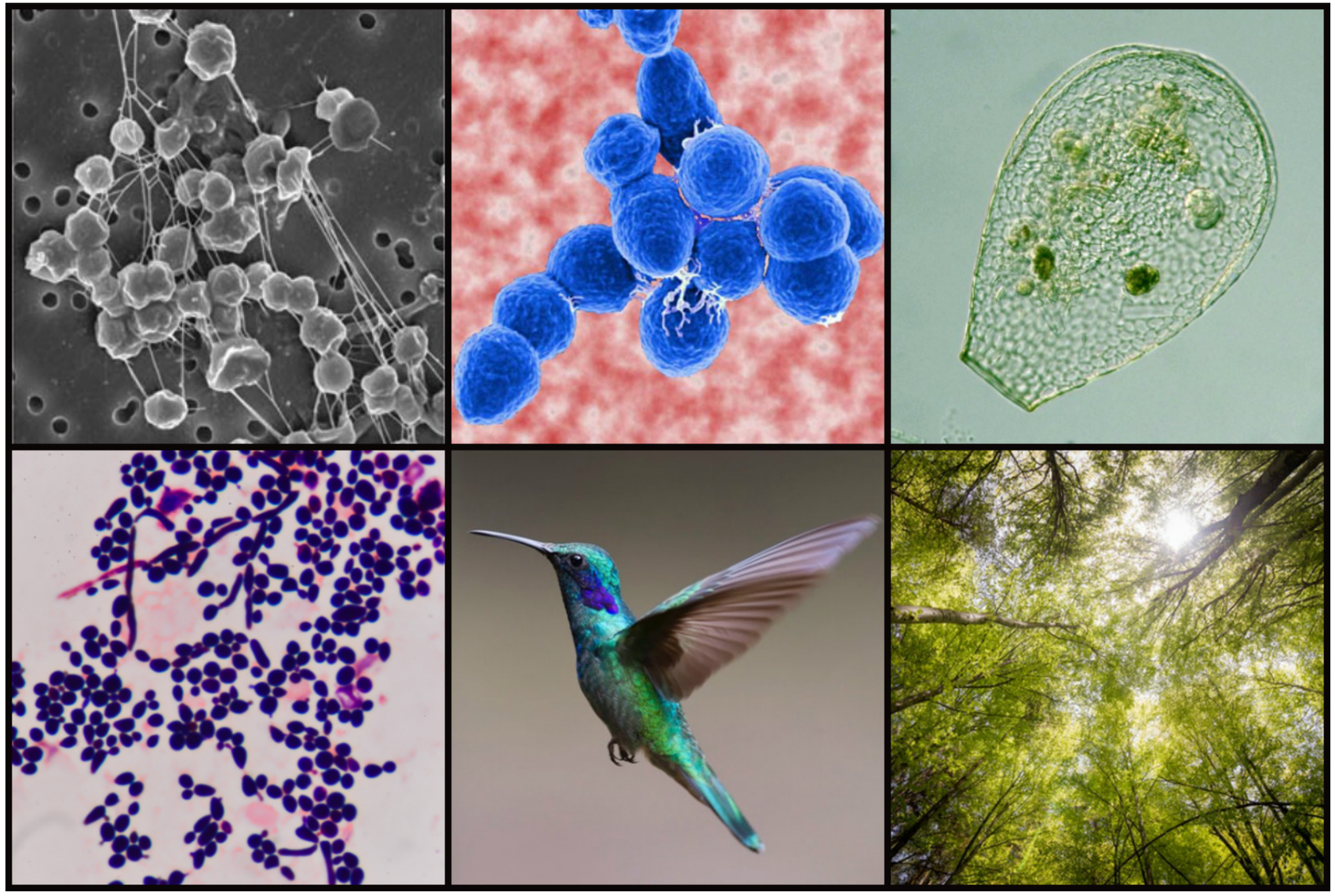
INTRODUCTION TO BIOCONDUCTOR IN R

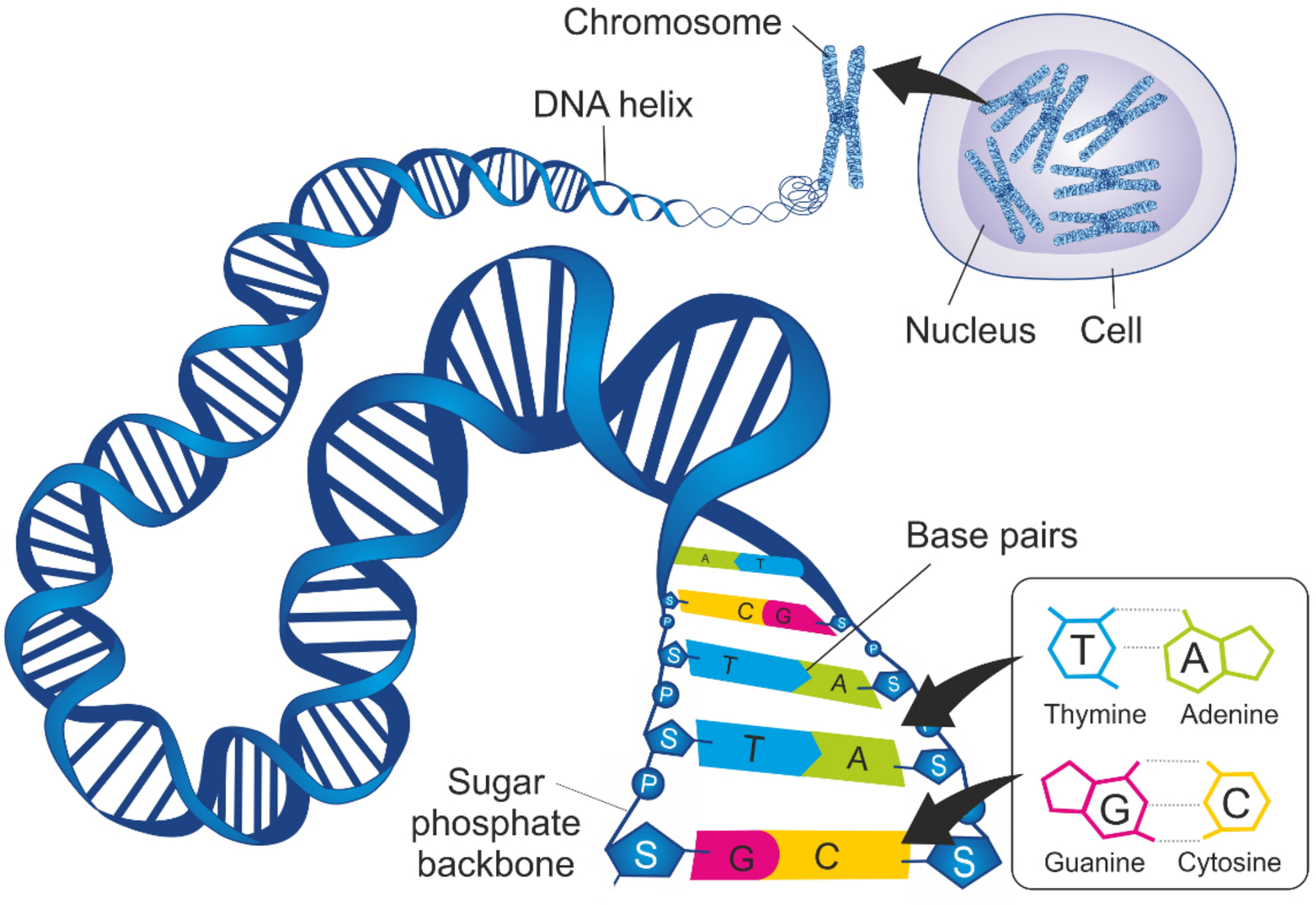
Introducing biology of genomic datasets

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

datacamp




Chromosome
DNA helix
Nucleus
Cell
Base pairs
Sugar
phosphate
backbone
Thymine
Adenine
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*

BSgenome annotation package

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

# Other available genomes
available.genomes()
```

```
"BSgenome.Alyrata.JGI.v1"
"BSgenome.Amellifera.BeeBase.assembly4"
"BSgenome.Amellifera.NCBI.AmelHAv3.1"
"BSgenome.Amellifera.UCSC.apiMel2"
"BSgenome.Amellifera.UCSC.apiMel2.masked"
...
```

```
length(yeast)
```

```
17
```

```
names(yeast)
```

```
"chrI"    "chrII"   "chrIII"  "chrIV"   "chrV"    "chrVI"   "chrVII"
"chrVIII" "chrIX"   "chrX"    "chrXI"   "chrXII"  "chrXIII" "chrXIV"
"chrXV"   "chrXVI"  "chrM"
```

```
seqlengths(yeast)
```

```
chrI    chrII   chrIII  chrIV    chrV    chrVI   chrVII   chrVIII  chrIX   chrX
230218  813184  316620  1531933  576874  270161  1090940  562643   439888  745751
chrXI  chrXII   chrXIII  chrXIV   chrXV    chrXVI   chrM
666816 1078177  924431   784333   1091291  948066   85779
```

Get sequences

- `getSeq()` : S4 method for BSgenome

```
# Select entire genomic sequence
getSeq(yeast)

# Select sequence from chromosome M
getSeq(yeast, "chrM")

# Select first 10 base pairs
getSeq(yeast, end = 10)
```

Let's practice!

INTRODUCTION TO BIOCONDUCTOR IN R