# Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

## EXAME-STYLE-QUESTIONS

HS2020

## Bioinformatics - Exame-Style-Questions

# REPRODUCIBLE RESEARCH CODE / SCRIPTS

## Bioinformatics - Exame-Style-Questions

**Q1** Why would comments in a script improve reproducibility?

**Q2** What is the difference between "data analysis" and "data mining" in regard to reproducibility?

## Bioinformatics - Exame-Style-Questions

**Q1** **Why would comments in a script improve reproducibility?**

Comments help to better understand code. A better understanding helps you and other people to (re)use the code.

**Q2** **What is the difference between "data analysis" and "data mining" in regard to reproducibility?**

Data analysis has a focus to verify e.g. a hypothesis. For example, searching for viral DNA in a clinical sample. Data mining is looking for patterns in datasets to build a hypothesis. The probability of a "chance find" is high.

# Bioinformatics - Exame-Style-Questions

**Q3** What would be the reasons (name 3) to use R insead of Excel to work with tables?

**Q4** Why is it required to provide the version of a application in materials and methods?

## Bioinformatics - Exame-Style-Questions

**Q3** **What would be the reasons (name 3) to use R instead of Excel to work with tables?**

- I can share the R script with others.
- Troubleshooting or improvements are possible.
- Faster, once it is finished. I can also recycle to code for other similar projects.
- A "correct" script is more reliable.

**Q4** **Why is it required to provide the version of a application in materials and methods?**

There might be differences between different version of the same program. For example different default settings, bugs, or corrections. To reproduce the results one would need to know the version.

# Bioinformatics - Exame-Style-Questions

**Q5** What is a code style?

**Q6** What is the purpose of using different suffixes. For example file.fa for fasta and file.fq for fastq sequence files?

# Bioinformatics - Exame-Style-Questions

## Q5 What is a code style?

A collections of rules (e.g. variable naming) for writing code in a more standardised fashion. It helps to improve readability and so reproducibility.

## Q6 What is the purpose of using different suffixes. For example file**.fa for fasta and** file**.fq for fastq sequence files?**

Suffixed indicate the file format. It is always good pratice to veryfy the content and format of a file but and indicater can help to improve file search / selection.

## Bioinformatics - Exame-Style-Questions

TERMINAL

# Bioinformatics - Exame-Style-Questions

Q1 What do the three commands have in common and what are the differences?

```
$ cp file1.txt file2.txt
$ mv file1.txt > file2.txt
$ cat file1.txt > file2.txt
```

# Bioinformatics - Exame-Style-Questions

**A1**

```
$ cp file1.txt file2.txt
```

> Creates a copy of file1.txt called file2.txt but does not change the original file. Both files are identical in content.

```
$ mv file1.txt > file2.txt
```

> Renames (or moves) file1.txt to new, file2.txt.

```
$ cat file1.txt > file2.txt
```

> Reads the content of file1.txt and redirects the output to file2.txt. The source file is not change. This might take a while longer and careful with file that are not simple text format (e.g. pictures).

## Bioinformatics - Exame-Style-Questions

Q2

What are the similarities and what are the difference in these cat commands?

```
$ cat f1.txt f4.txt f2.txt > New.txt
$ cat f?.txt > New.txt
$ cat f*.txt > New.txt
```

## Bioinformatics - Exame-Style-Questions

A2

```
$ cat f1.txt f4.txt f2.txt > New.txt
```

> Reads text files f1, f4, and f3 and redirects content to a new file. The command concatenates the content of the three files into one file in the provided order.

```
$ cat f?.txt > New.txt
```

> Similar as before but it would concatenate all file that would match the search pattern (e.g. f3.txt, f5.txt, or ff.txt) in alphabethis order..

```
$ cat f*.txt > New.txt
```

> Same as above but it would concatenate all files that start with f and end with .txt e.g. f33.txt, file1.txt, or f.txt

## Bioinformatics - Exame-Style-Questions

Q3

How can I list all files starting with a capital A or B in a folder?

## Bioinformatics - Exame-Style-Questions

How can I list all files starting with a capital A or B in a folder?

```
$ ls -1 [AB]*

$ ls -1 [^C-Z]*
```

```
$ ls -1 | grep ^[AB]
```

## Bioinformatics - Exame-Style-Questions

Q4

Which command can I not use to count the sequence records in a fasta file?

```
$ grep ">" -c file.f*a
```

```
$ grep ">" file.fasta | wc -l
```

```
$ grep ">" -c file.fasta | wc -l
```

## Bioinformatics - **Exame-Style-Questions**

**A4**

Which command can I not use to count the sequence records in a fasta file?

```
$ grep ">" -c file.f*a
```

```
$ grep ">" file.fasta | wc -l
```

```
$ grep ">" -c file.fasta | wc -l
```

The reuslt would be one because it would count the lines of the grep count.

## Bioinformatics - Exame-Style-Questions

Can you make this cascade of commands work?

Q5

```
$ touch log.txt
$ echo -n "N(seq): " > log.txt
$ grep ">" -c seq.fa > log.txt
$ echo -n "n(lines): " > log.txt
$ wc -l seq.fa > log.txt
$ echo " " > log.txt
$ cat log.txt
```

## Bioinformatics - Exame-Style-Questions

Can you make this cascade of commands work?

A5

```
$ touch log.txt
$ echo -n "N(seq): " >> log.txt
$ grep ">" -c seq.fa >> log.txt
$ echo -n "n(lines): " >> log.txt
$ wc -l seq.fa >> log.txt
$ echo " " >> log.txt
$ cat log.txt
```

> redirect output, overwrite if it already exists
>> redirect but add output to existing files

# REGULAR EXPRESSION

## Bioinformatics - Exame-Style-Questions

Q1 Below a multi-sequence fasta file with 3 sequences. Which lines would produce a hit with the following regex find term:

find: `^[^A-Z]`

1. >SeqA Firmicutes >80%
2. ATGTTGCCTGTCGACAAATGCTGTCGACAAATGC
3. >SeqB Protobacteria 95%
4. ATGTTGCCTTTCGACAGATGCTGTCGACAAATGC
5. >SeqC unknown bac
6. ATGTTGCCTTTCGACAGATGCTGTCGACAAATGC

## Bioinformatics - Exame-Style-Questions

Q1 Below a multi-sequence fasta file with 3 sequences. Which lines would produce a hit with the following regex find term:

find: `^>|^[^ACTG]`

Lines starting with ">" OR not starting with a captial A,C, T or G.

1. **>SeqA Firmicutes >80%**
2. ATGTTGCCTGTCGACAAATGCTGTCGACAAATGC
3. **>SeqB Protobacteria 95%**
4. ATGTTGCCTTTCGACAGATGCTGTCGACAAATGC
5. **>SeqC unknown bac**
6. ATGTTGCCTTTCGACAGATGCTGTCGACAAATGC

## Bioinformatics - Exame-Style-Questions

**Q2** Which RegEx of the 3 would turn the input (left) into the output (right):

Input:

A1 A2 A3

B1 B2 B3

C1 C2 C3

Output:

A2 A3 A1

B2 B3 B1

C1 C2 C3

1. Find: ([AB]\d)\s(\w+)\s(\w+) / Replace: $2 $3 $1
2. Find: (\w+)\s(\w+)\s(\w+) / Replace: $2 $3 $1
3. Find: (\d+)\s(\w+)\s(\w+) / Replace: $1 $2 $3

# Bioinformatics - Exame-Style-Questions

**Q2** Which RegEx of the 3 would turn the input (left) into the output (right):

| Input: | Output: |
|---|---|
| A1 A2 A3 | A2 A3 A1 |
| B1 B2 B3 | B2 B3 B1 |
| C1 C2 C3 | C1 C2 C3 |

Only the first two lines (A and B) need to be arranged but not the last (C).

1. **Find: ([AB]\d)\s(\w+)\s(\w+) / Replace: $2 $3 $1**
2. Find: (\w+)\s(\w+)\s(\w+) / Replace: $2 $3 $1
3. Find: (\d+)\s(\w+)\s(\w+) / Replace: $1 $2 $3

## Bioinformatics - Exame-Style-Questions

What would this RegEx term target?

```
[a-Z0-9._%+-]+@[a-Z0-9.-]+\.[a-Z]{2,4}
```

## What is the meaning of this RegEx term

```
[a-Z0-9._%+-]+@[a-Z0-9.-]+\.[a-Z]{2,4}
```

```
[a-Z0-9._%+-]+@[a-Z0-9.-]+\.[a-Z]{2,4}
```

A.Test@test.com

# Bioinformatics - Exame-Style-Questions

## Bioinformatics - Exame-Style-Questions

Q1 What is the difference between the two R code lines?

```
> distance::maxlength(data)

> maxlength(data)
```

Bioinformatics - **Exame-Style-Questions**

## Q1 What is the difference between the two R code lines?

```
> distance::maxlength(data)

> maxlength(data)
```

In the first example, I what to use function "maxlength" from the package "distance". The package needs to be installed but not loaded.

In the second example, I what to use the function maxlength on my data. The function name can be by multiple packages and I am sure I will use the correct one.

## Bioinformatics - Exame-Style-Questions

```
lb2kg <- function(lb) {
           # Convert pounds to kg
           kg = (lb / 2.2046)
 }
```

```
1. lb2kg(2.2046): 1
2. lb2kg(2.2046): error
3. lb2kg(2.2046):
4. lb2kg(2.2046): 2.2046
5. lb2kg(2.2046): kg
```

## Bioinformatics - Exame-Style-Questions

```
lb2kg <- function(lb) {
           # Convert pounds to kg
           kg = (lb / 2.2046)
 }
```

```
1. lb2kg(2.2046): 1
2. lb2kg(2.2046): error
3. lb2kg(2.2046): no output defined
4. lb2kg(2.2046): 2.2046
5. lb2kg(2.2046): kg
```

## Bioinformatics - Exame-Style-Questions

```
lb2kg <- function(lb) {
           # Convert pounds to kg
           kg = (lb / 2.2046)
 }
```

## Bioinformatics - Exame-Style-Questions

```
lb2kg <- function(lb) {
           # Convert pounds to kg
           kg = (lb / 2.2046)
 }
```

```
lb2kg <- function(lb) {
           # Convert pounds to kg
           kg = (lb / 2.2046)
           return(kg)
 }
```

## Bioinformatics - Exame-Style-Questions

```
hot <- function(x, type = "f2c") {
  switch(type,
         f2c = ((x-32)*5/9),
         c2f = (9/5*x+32),
         default = "Error")
         }
```

```
1. hot(32): Not working becasue type is missing.
2. hot(32): 32
3. hot(32):  0
4. hot(32): 89.6
5. hot(32): Error
```

## Bioinformatics - Exame-Style-Questions

```
hot <- function(x, type = "f2c") {
  switch(type,
         f2c = ((x-32)*5/9),
         c2f = (9/5*x+32),
         default = "Error")
         }
```

1. hot(32): Not working becasue type is missing.
2. hot(32): 32
3. **hot(32):  0**
4. hot(32): 89.6
5. hot(32): Error