GDC
Genetic
Diversity
Centre
Zurich

Genetic Diversity: Analysis

Warmup #2

Friday 25. June 2021

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Reproducible Research

GDA ▷ Warm Up ▷ RepRes

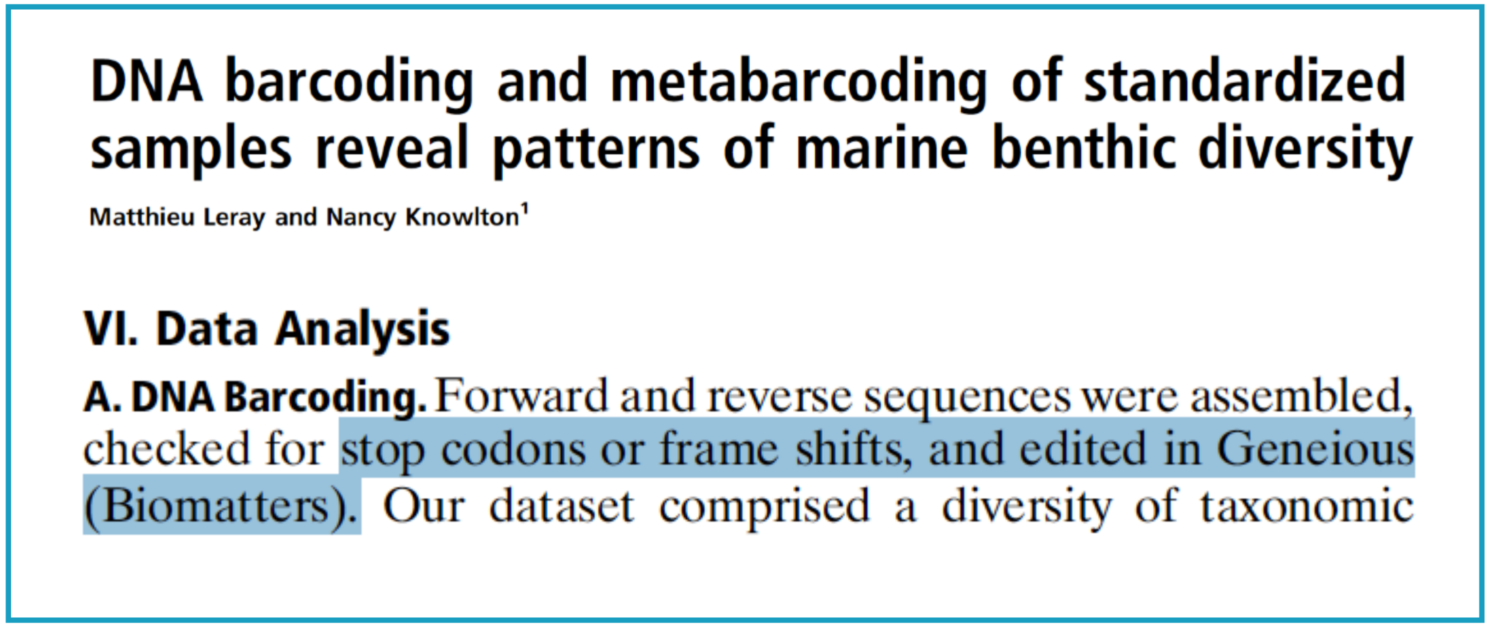
Q1

"Application with a GUI are convenient to use but difficult to document and reproduce".

Do you agree or diagree with this statement and can you explain why?

A1

GUI based applications are often balckboxes and it is often not clear what (default) parameters or settings have been used.



DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity

Matthieu Leray and Nancy Knowlton[1]

VI. Data Analysis

A. DNA Barcoding. Forward and reverse sequences were assembled, checked for stop codons or frame shifts, and edited in Geneious (Biomatters). Our dataset comprised a diversity of taxonomic

The vertebrate mtDNA code differs from the "Universal" code.

- AUA and AUG are both Met codons
- UGA codes for Trp and not a Stop codon
- AGA and AGG codons are read as Stops instead of Arg

Slightly different mtDNA codes are found in Drosophila and other invertebrate groups.

Material and Methods

In a first step, all paired-end raw reads were successfully merged using FLASh (version 1.2.9, Magoc and Salzberg 2011) with minimum overlap of 5nt and maximal mismatch ration of 0.8.

Supplementary Data

```
## (a) Merging overlapping paired-end reads
# -v Version (1.2.9)
# -m minimum overlap (default 10bp)
# -x max mismatch ration (default 0.25)

flash -m 5 -x 0.8 random_1000_R1.fq random_1000_R2.fq -o
merged | tee flash.log
```

You are part of the Covid broadcast team and you are responsible for the **daily statistics**. All laboratories conducting Covid testing have to send daily reports. It is your duty to collect these reports. The labs are using standardised **CSV files**. At the end of a day you have hundreds of CSV files and you need to add them to your existing table. Your team-leader recommends to use the spreadsheet editor (e.g. Microsoft Excel) to store the data because he already has create a summary report template. What do you think about it?

The Guardian

CovidNews: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

One lab had sent its daily test report to PHE (Public Health England) in the form of a **CSV file** – the simplest possible database format, just a list of values separated by commas. That report was then **loaded into Microsoft Excel**, and the new tests at the bottom were added to the main database.

But while CSV files can be any size, Microsoft Excel files can only be 1,048,576 rows long – or, in older versions which PHE may have still been using, a mere 65,536. When a CSV file longer than that is opened, the bottom rows get cut off and are no longer displayed. That means that, once the lab had performed more than a million tests, it was only a matter of time before its reports failed to be read by PHE.

Alex Hern *UK technology editor*

@alexhern

Tue 6 Oct 2020 08.21 BST

Microsoft’s spreadsheet software is one of the world’s most popular business tools, but it is regularly implicated in errors which can be costly, or even dangerous, because of the ease with which it can be used in situations it was not designed for.

In 2013, an Excel error at JPMorgan masked the loss of almost $6bn (£4.6bn), after a cell mistakenly divided by the sum of two interest rates, rather than the average. The news led James Kwak, a professor of law at the University of Connecticut, to warn that Excel is “incredibly fragile”.

“There is no way to trace where your data comes from, there’s no audit trail (so you can overtype numbers and not know it), and there’s no easy way to test spreadsheets, for starters. The biggest problem is that anyone can create Excel spreadsheets – badly. Because it’s so easy to use, the creation of even important spreadsheets is not restricted to people who understand programming and do it in a methodical, well-documented way,” Kwak wrote.

Errors from the spreadsheet software have even changed the very foundations of human genetics. The names of 27 genes have been changed over the past year by the Human Gene Nomenclature Committee, after Microsoft’s program continually misformatted them. The genes SEPT1 and MARCH1, for instance, have been changed to SEPTIN1 and MARCHF1 after they were repeatedly turned into dates, while symbols that were common words have been altered so that grammar tools didn’t autocorrect them: WARS is now WARS1, for instance.

The Guardian

Alex Hern *UK technology editor*

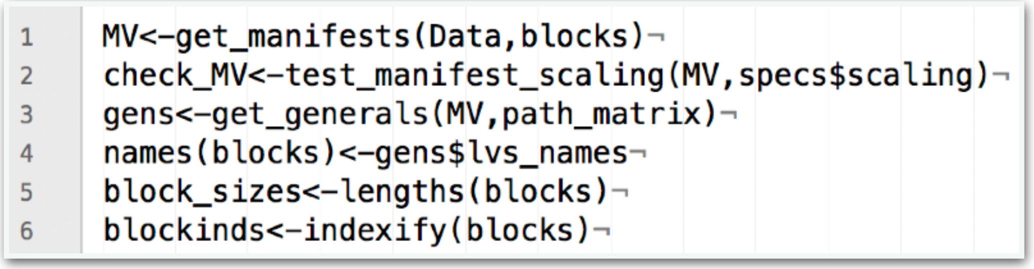
@alexhern

Tue 6 Oct 2020 08.21 BST

Q2

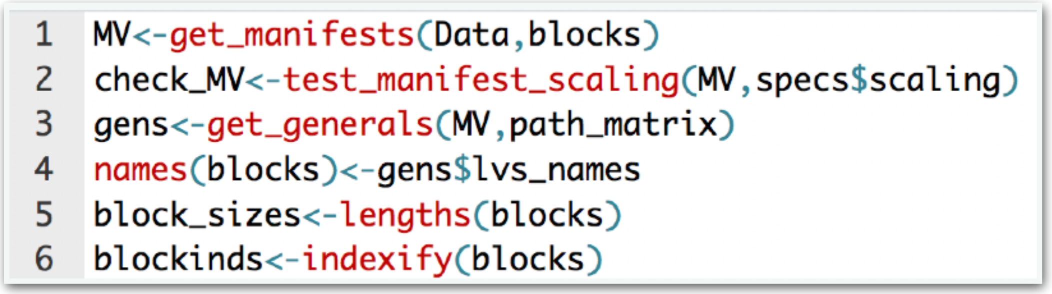
What can you do to increase reproducibility of your R script(s)?

Built-in Help: Example for RStudio



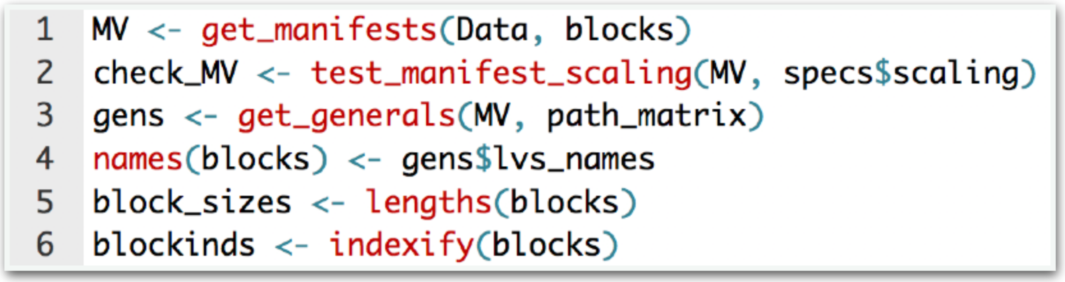
```
MV<-get_manifests(Data,blocks)¬
check_MV<-test_manifest_scaling(MV,specs$scaling)¬
gens<-get_generals(MV,path_matrix)¬
names(blocks)<-gens$lvs_names¬
block_sizes<-lengths(blocks)¬
blockinds<-indexify(blocks)¬
```

☑ Show syntax highlighting in console input



```
MV<-get_manifests(Data,blocks)
check_MV<-test_manifest_scaling(MV,specs$scaling)
gens<-get_generals(MV,path_matrix)
names(blocks)<-gens$lvs_names
block_sizes<-lengths(blocks)
blockinds<-indexify(blocks)
```

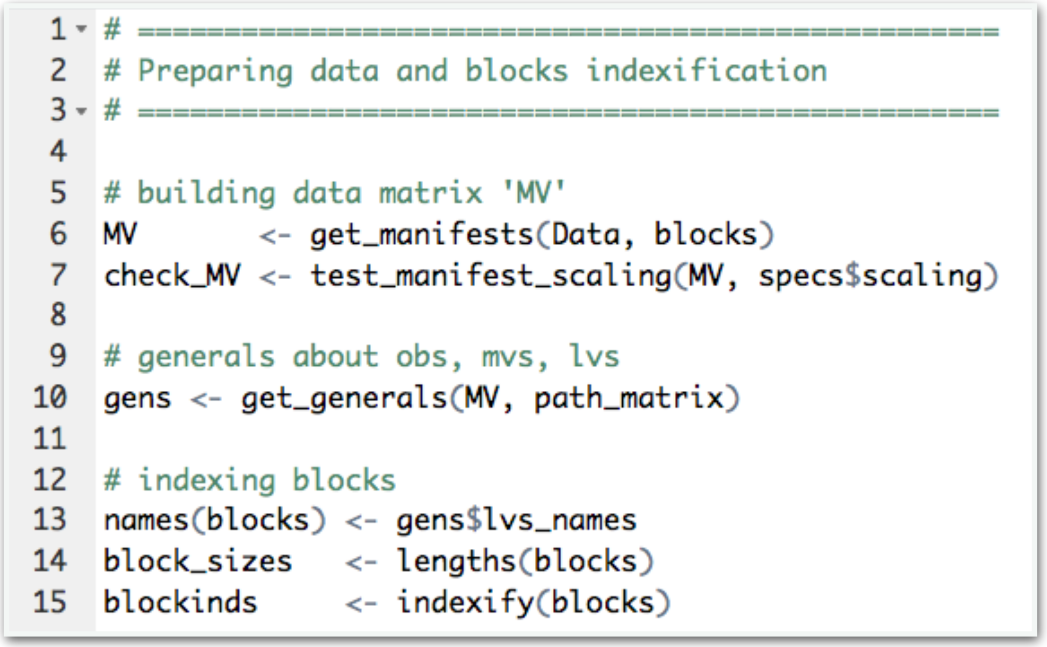
Reformat Code ⇧⌘A



```
MV <- get_manifests(Data, blocks)
check_MV <- test_manifest_scaling(MV, specs$scaling)
gens <- get_generals(MV, path_matrix)
names(blocks) <- gens$lvs_names
block_sizes <- lengths(blocks)
blockinds <- indexify(blocks)
```

#

Structure and comment your script



```
# ===================================================
# Preparing data and blocks indexification
# ===================================================

# building data matrix 'MV'
MV       <- get_manifests(Data, blocks)
check_MV <- test_manifest_scaling(MV, specs$scaling)

# generals about obs, mvs, lvs
gens <- get_generals(MV, path_matrix)

# indexing blocks
names(blocks) <- gens$lvs_names
block_sizes   <- lengths(blocks)
blockinds     <- indexify(blocks)
```

GDA ▷ Warm Up ▷ RepRes

Q3

Is it possible to make "random" processes reproducible? In other words, what can I do to make random objects reproducible for example in R?

```
> sample(1:100, 3, replace = TRUE)
# 55 28 33
```

GDA ▷ Warm Up ▷ RepRes

```
# Generate random number(s)
> sample(1:100, 3, replace = TRUE)
# 55 28 33
> sample(1:100, 3, replace = TRUE)
# 34 92 26
```

```
# Repeatable Random Numbers
> set.seed(210625)
> sample(1:100, 3, replace = TRUE)
[1] 71 51 88
```

Non-Parametric Resampling

```
bootstrap(data, mean)

Call:
bootstrap(data = CLEC, statistic = mean, seed = 19)
Replications: 10000

Summary Statistics:
     Observed       SE     Mean        Bias
mean 16.50913 3.96866 16.54073 0.03160109
```

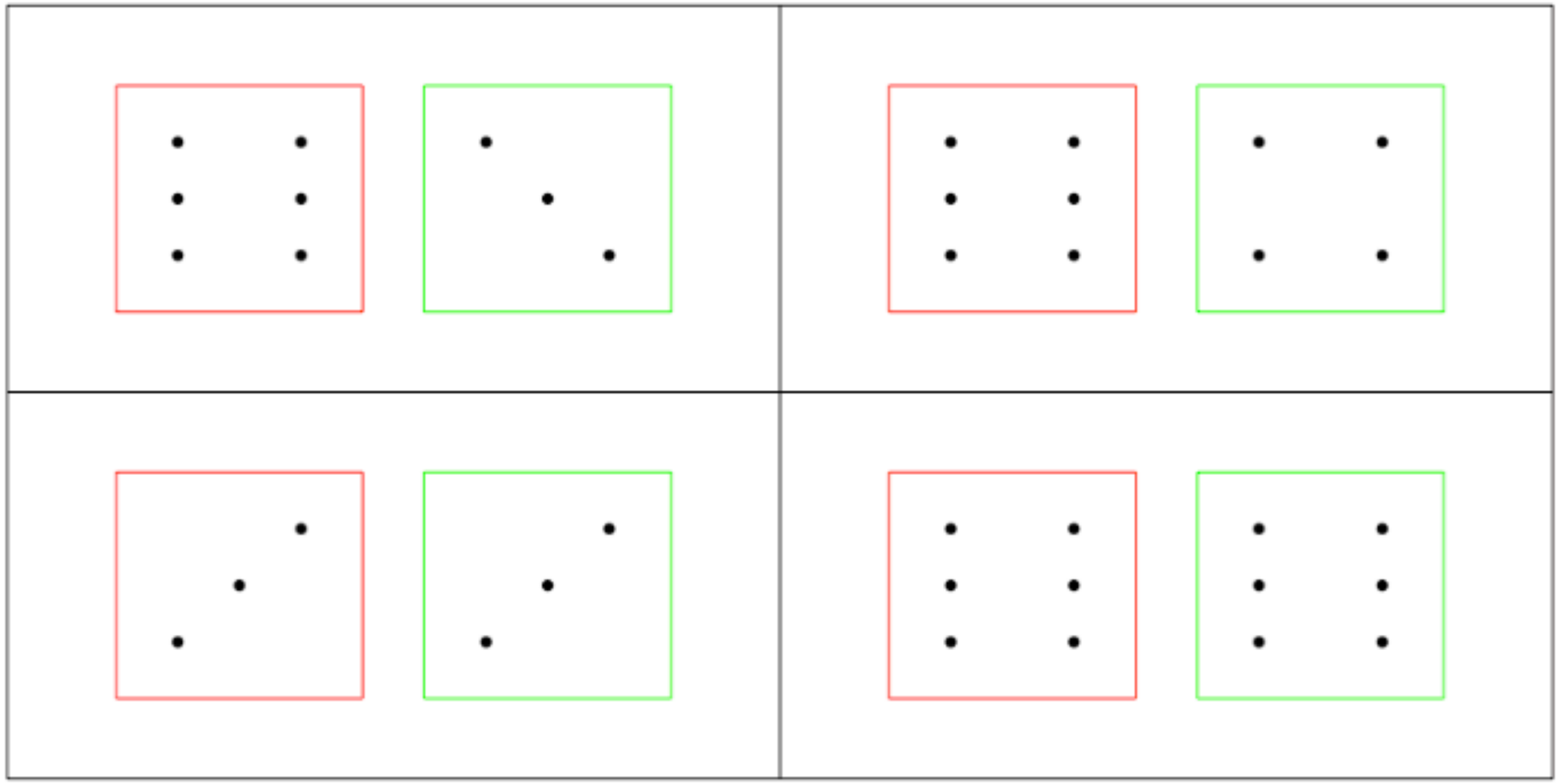
```
bootstrap(data, mean)

Call:
bootstrap(data = CLEC, statistic = mean)
Replications: 10000

Summary Statistics:
     Observed       SE     Mean          Bias
mean 16.50913 3.985572 16.47043 -0.03869548
```

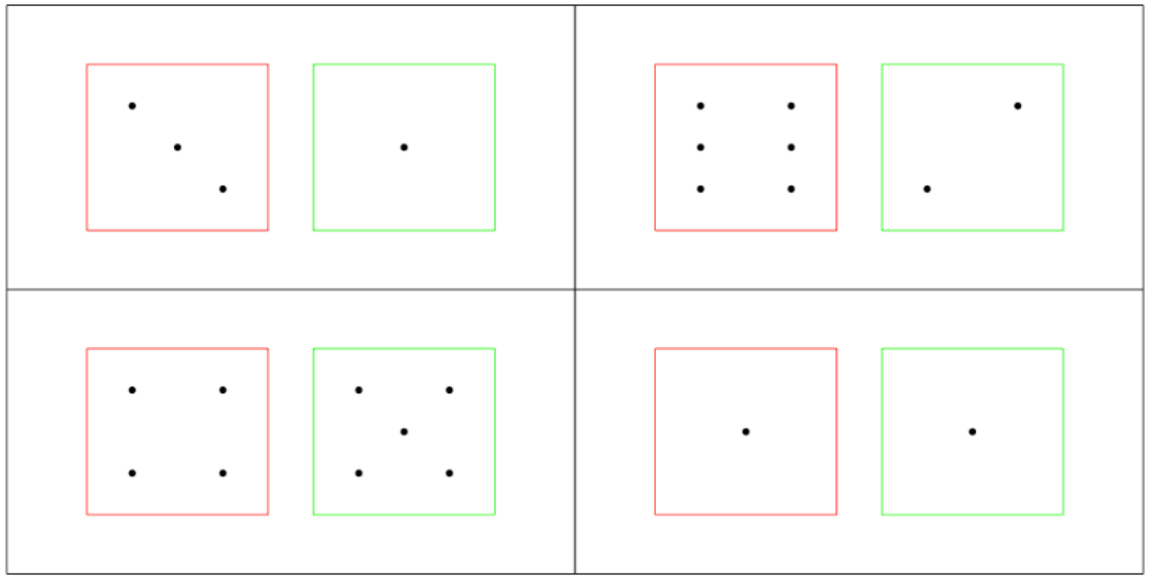
Non-Parametric Resampling

```
bootstrap(data, mean, seed = 210625)

Call:
bootstrap(data = CLEC, statistic = mean, seed = 191029)
Replications: 10000

Summary Statistics:
     Observed       SE     Mean        Bias
mean 16.50913 3.957816 16.47054 -0.03859261
```

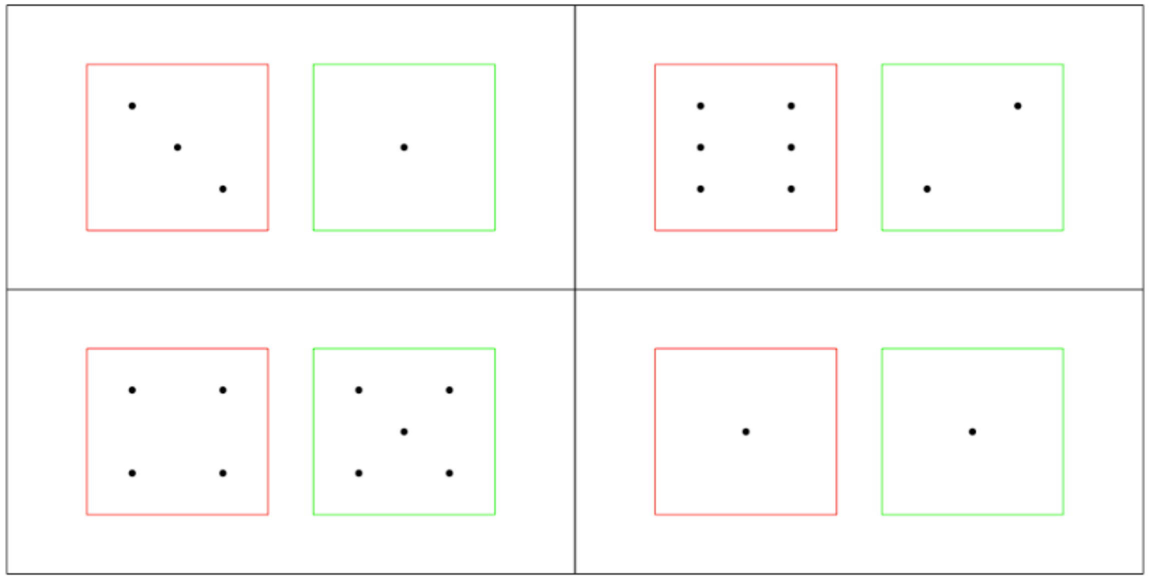
```
bootstrap(data, mean, seed = 210625)

Call:
bootstrap(data = CLEC, statistic = mean, seed = 191029)
Replications: 10000

Summary Statistics:
     Observed       SE     Mean        Bias
mean 16.50913 3.957816 16.47054 -0.03859261
```

```
TeachingDemos::dice(4, 2, plot.it = TRUE)
```





```
set.seed(123)
dice(4, 2, plot.it = TRUE)
```



```
set.seed(123)
dice(4, 2, plot.it = TRUE)
```

Q4 Why would you need comments in your code and what is the connection to reproducibility?

Good developers write good code;
great ones also write good comments.

Everything that increases the readability of your code increases reproducibility!

A Quick Recap

What can you do to make your research reproducible?

1 **Get Organised**
Develop a simple style and cultivate organisational habits. Ask questions!

2 **Avoid applications with GUIs**
and use terminal command instead.

3 **Write Reports**
Precise description of the workflow including versions and parameters.

4 **Publish Your Script / Code**
You can publish your polished script together with your other files.

A Quick Recap

What can you do to make your research reproducible?

5

Code Style

Learn and use a common style.

6

#

Comment your code and do it generously.

7

Code Editor

Make use of the different features like syntax highlighting, code folding or auto-completion.

8

Version Control

For bigger or collaborative projects use version control and /or cloud based solutions.

Markdown

GDA ▷ Warm Up ▷ Markdown

Q5

Why would you write a markdown report if you can use a word processor like Microsoft Word?

GDA ▷ Warm Up ▷ Markdown

A5

Don’t get me started …

GDA ▷ Warm Up ▷ Markdown

Q6 Why would a markdown report help to improve reproducibility?

GDA ▷ Warm Up ▷ Markdown

A6

In a markdown report, you can easily combine text, figures, tables and code. It helps to divide "code junk" into better readable "code chunks" with context.

Regular Expressions

Q7

What are Regular Expressions and for what purpose would you use it?

GDA ▷ Warm Up ▷ RegEx

A7

A **regular expression** (**regex** or **regexp** for short) is a special **text string for describing a search pattern**. You can think of regular expressions as wildcards on steroids.

A regular expression "**engine**" is a piece of software that can process regular expressions, trying to match the pattern to the given string. Usually, the engine is part of a larger application and you do not access the engine directly.

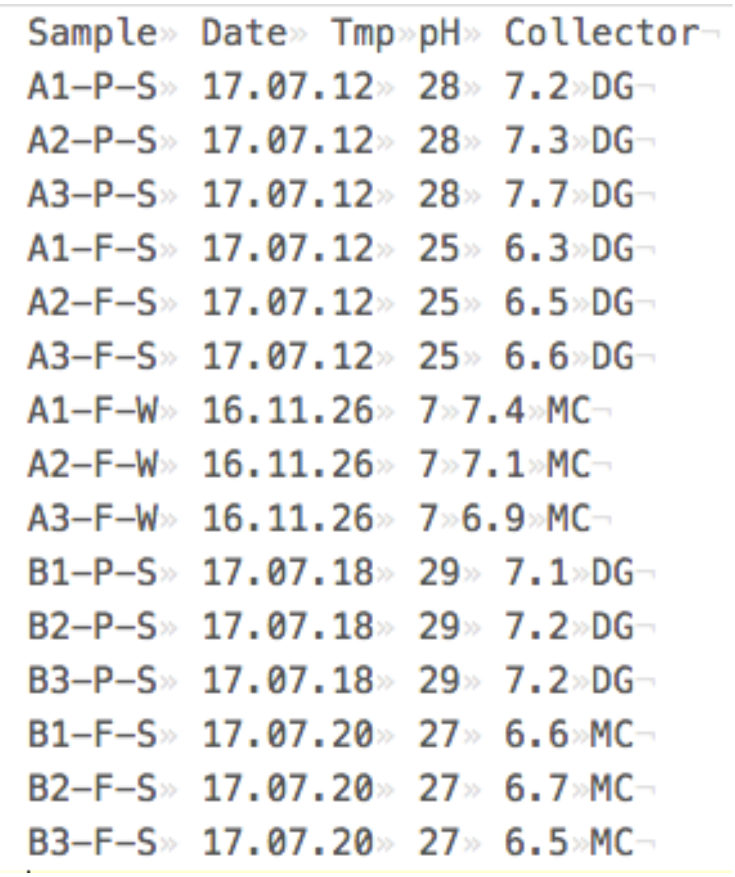
GDA ▷ Warm Up ▷ RegEx

Mus musclus	M. musclus
Agalma elegans	A. elegans
Frillagalma vitiazi	F. vitiazi
Cordagalma tottoni	C. tottoni
Shortia galacifolia	S. galacifolia

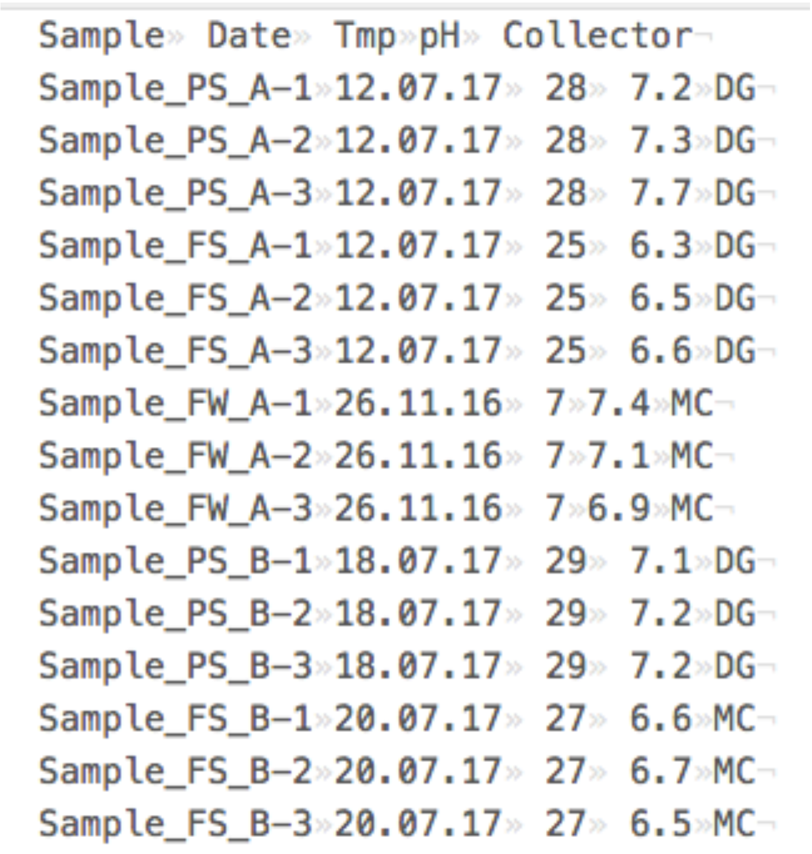
Find: `(\w)\w+ (\w+)`

Replace: `$1. $2`

GDA ▷ Warm Up ▷ RegEx



```
Sample» Date» Tmp»pH» Collector¬
A1-P-S» 17.07.12» 28» 7.2»DG¬
A2-P-S» 17.07.12» 28» 7.3»DG¬
A3-P-S» 17.07.12» 28» 7.7»DG¬
A1-F-S» 17.07.12» 25» 6.3»DG¬
A2-F-S» 17.07.12» 25» 6.5»DG¬
A3-F-S» 17.07.12» 25» 6.6»DG¬
A1-F-W» 16.11.26» 7»7.4»MC¬
A2-F-W» 16.11.26» 7»7.1»MC¬
A3-F-W» 16.11.26» 7»6.9»MC¬
B1-P-S» 17.07.18» 29» 7.1»DG¬
B2-P-S» 17.07.18» 29» 7.2»DG¬
B3-P-S» 17.07.18» 29» 7.2»DG¬
B1-F-S» 17.07.20» 27» 6.6»MC¬
B2-F-S» 17.07.20» 27» 6.7»MC¬
B3-F-S» 17.07.20» 27» 6.5»MC¬
```



```
Sample» Date» Tmp»pH» Collector¬
Sample_PS_A-1»12.07.17» 28» 7.2»DG¬
Sample_PS_A-2»12.07.17» 28» 7.3»DG¬
Sample_PS_A-3»12.07.17» 28» 7.7»DG¬
Sample_FS_A-1»12.07.17» 25» 6.3»DG¬
Sample_FS_A-2»12.07.17» 25» 6.5»DG¬
Sample_FS_A-3»12.07.17» 25» 6.6»DG¬
Sample_FW_A-1»26.11.16» 7»7.4»MC¬
Sample_FW_A-2»26.11.16» 7»7.1»MC¬
Sample_FW_A-3»26.11.16» 7»6.9»MC¬
Sample_PS_B-1»18.07.17» 29» 7.1»DG¬
Sample_PS_B-2»18.07.17» 29» 7.2»DG¬
Sample_PS_B-3»18.07.17» 29» 7.2»DG¬
Sample_FS_B-1»20.07.17» 27» 6.6»MC¬
Sample_FS_B-2»20.07.17» 27» 6.7»MC¬
Sample_FS_B-3»20.07.17» 27» 6.5»MC¬
```

find: **(\w)(\d)-(\w)-(\w)**\t(\d{2}).(\d{2}).(\d+)\t(\d+)\t(\d.\d)\t(\w+)

replace: **Sample_$3$4_$1-$2**\t$7.$6.$5\t$8\t$9\t$10

GDA ▷ Warm Up ▷ RegEx

Genetic Diversity Centre Zurich

Q8 What is this Regular Expression looking for?

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

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