University of Glasgow | College of Medical, Veterinary & Life Sciences

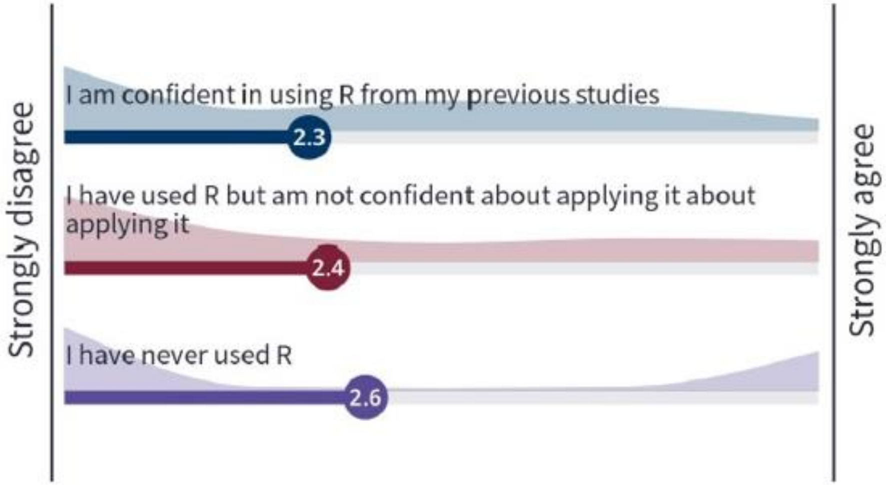
Biodiversity, One Health, and Veterinary Medicine

Key Research Skills

Intro to R: 1. Basics

Course Instructors:
Dr Roman Biek (Roman.Biek@glasgow.ac.uk)
Dr Simon Babayan (Simon.Babayan@glasgow.ac.uk)

2023 class


Strongly disagree
Strongly agree
I am confident in using R from my previous studies
2.3
I have used R but am not confident about applying it about applying it
2.4
I have never used R
2.6

What is R?

A computer language and environment for statistical computing and graphics

Related to the computing language S, which has been developed into a commercial product (S−PLUS)

=> R is free

Reasons for using R

1. **Highly flexible and versatile e.g.**
   - Data exploration and visualisation
   - Data manipulation
   - Descriptive statistics
   - Statistical testing
   - Parameter estimation
   - Building models
   - …

Reasons for using R

2. **Well supported and documented**
   - Constantly maintained and further developed by dedicated expert team
   - Large online user community providing information and advice
   - Extensive online help
   - Books, tutorials, courses etc

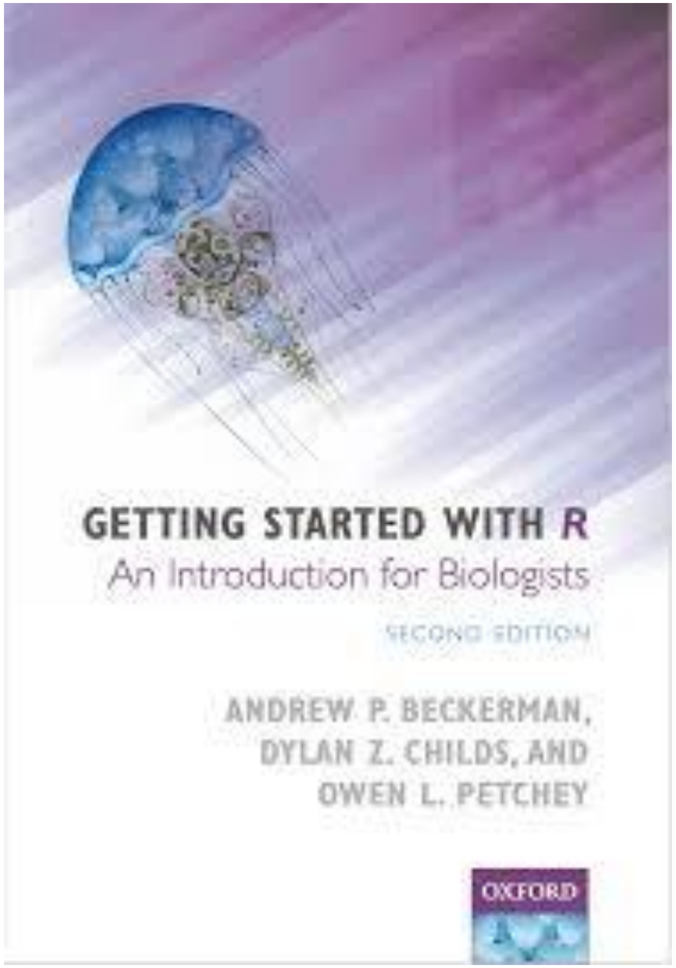
Reasons for using R

3. **Is the standard in the natural sciences**
    - Offers many state of the art tools e.g.
        - mixed effect models
        - spatial statistics
        - likelihood based approaches (Maximum Likelihood, Bayesian)
    - Powerful graphic capabilities

Reasons for using R

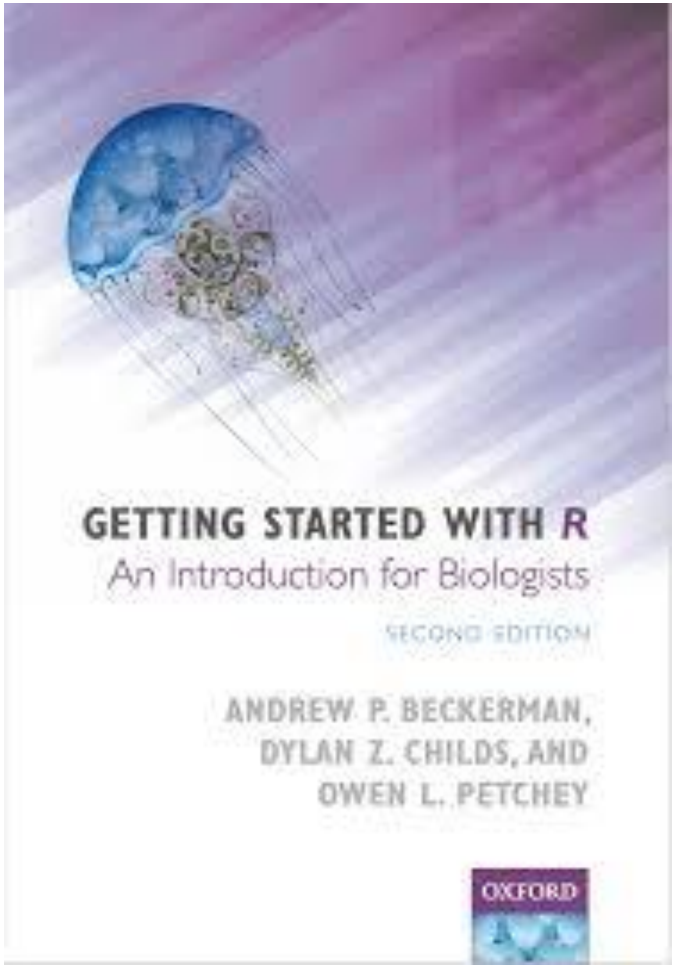
4. **Easy to produce a record of analysis**
   - Saving **scripts** makes analyses transparent and **reproducible** (*especially if well annotated*)
   - Easy to repeat analyses of updated datasets
   - Code can be shared with others and “recycled” in future analyses

Text book



- Gentle introduction, assumes very little background
- Focus is on making you proficient at dealing with data quickly rather than on being comprehensive
- 2nd edition thoroughly revised and expanded

Text book



- New edition focuses on newer packages e.g. dplyr, ggplot2 => some differences from the functions built into R by default ("base R")

- IMPORTANT: other KRS instructors likely to use base R in their exercises (see pointers at the end of class)

Online courses

codecademy

coursera

CODING CLUB

- A number of well-developed online course platforms to explore
- Most have free courses available especially for beginners, but more advanced course may require subscription

Course plan

Day 1: Basics – R syntax, importing data, scripts

Day 2: Data manipulation and visualisation

Day 3: Working with data

Day 4: Advancing your R skills

Day 5: Small group projects: data analysis exercise (formative assessment / no marks)

Day 6: Intro to Programming in R

Home assignment: Plotting graphs

What to hand in

- **Scripts/assignments**: always submit script (R script, or R markdown notebook)
- For Day 2 onwards, submit both the script (or R notebook) and the corresponding html report (i.e., only scripts for Chp 1&2, and script + html report for all other assignments).
- Usually due by start of next class (or earlier, check schedule)
- Scripts for Days 1, 2, and 4 are not marked (no individual feedback, only general), but need to be submitted and be complete to receive full **engagement mark**

Course mark

1. Complete R scripts for in class assignments (Days 1, 2, and 4)
2. R script for in class assignment Day 3*
3. Home assignment on plotting *

* each worth 10% of overall KRS mark; you will receive individual feedback only on these two and general class feedback on the rest.

Marking criteria

COMPLETENESS: Is the assignment complete? Is the script correct in terms of code (no error messages) and the graphical output it produces?

STRUCTURE: Is the script neat and logically structured? Are there redundant or unnecessary parts?

ANNOTATIONS: Are the explanations meaningful and informative? Do they reveal correct technical understanding?

Marking rubrics: 0-6 points for each category, maximum of 18 in total

Keeping good records of your work

- Hand in your assignments on time
- Revise your scripts as you go along according to feedback given
- If you stay on top of this, you will learn more quickly and will have a useful resource you can go back to in the future

Intended Learning Outcomes

By the end of this course, students should be able to:

- Use R effectively in the R Studio environment
- Install R packages/libraries
- Import and manipulate data
- Summarise data and produce descriptive statistics
- Plot data and produce professional looking graphs and reports
- Acquire technical help in the use R from literature and online sources
- Produce permanent and informative records of their work in the form of annotated scripts
- Know how to use some of the key features of R including basic statistical analyses.

🚧 Using Large Language Models

A few key considerations:

1. Accuracy and Reliability

- **Error-Prone Outputs:** AI models can generate incorrect or misleading information. Always verify the outputs with trusted sources.
- **Lack of Context:** LLMs may not understand the full context of your query, leading to partially accurate or irrelevant responses.

2. Understanding and Learning

- **Superficial Understanding:** Relying too heavily on AI tools can impede deep learning and understanding of concepts, as students might not engage fully with the problem-solving process.
- **Skill Development:** Fundamental skills in data analysis and coding are essential. Students should practice independently to build a solid foundation.

3. Ethical Considerations

- **Plagiarism Risks:** Unattributed use of AI-generated content can result in academic misconduct. Always credit any assistance appropriately and ensure your work is your own.
- **Misuse of Data:** Ensure compliance with data privacy laws and ethical guidelines when handling sensitive information in analysis.

4. Analytical Soundness

- **Lack of Critical Thinking:** AI tools do not replace critical thinking and analytical skills necessary for accurate data interpretation and problem-solving.
- **Debugging and Validation:** Code and data analysis outputs must be thoroughly tested and validated. AI cannot guarantee error-free solutions.

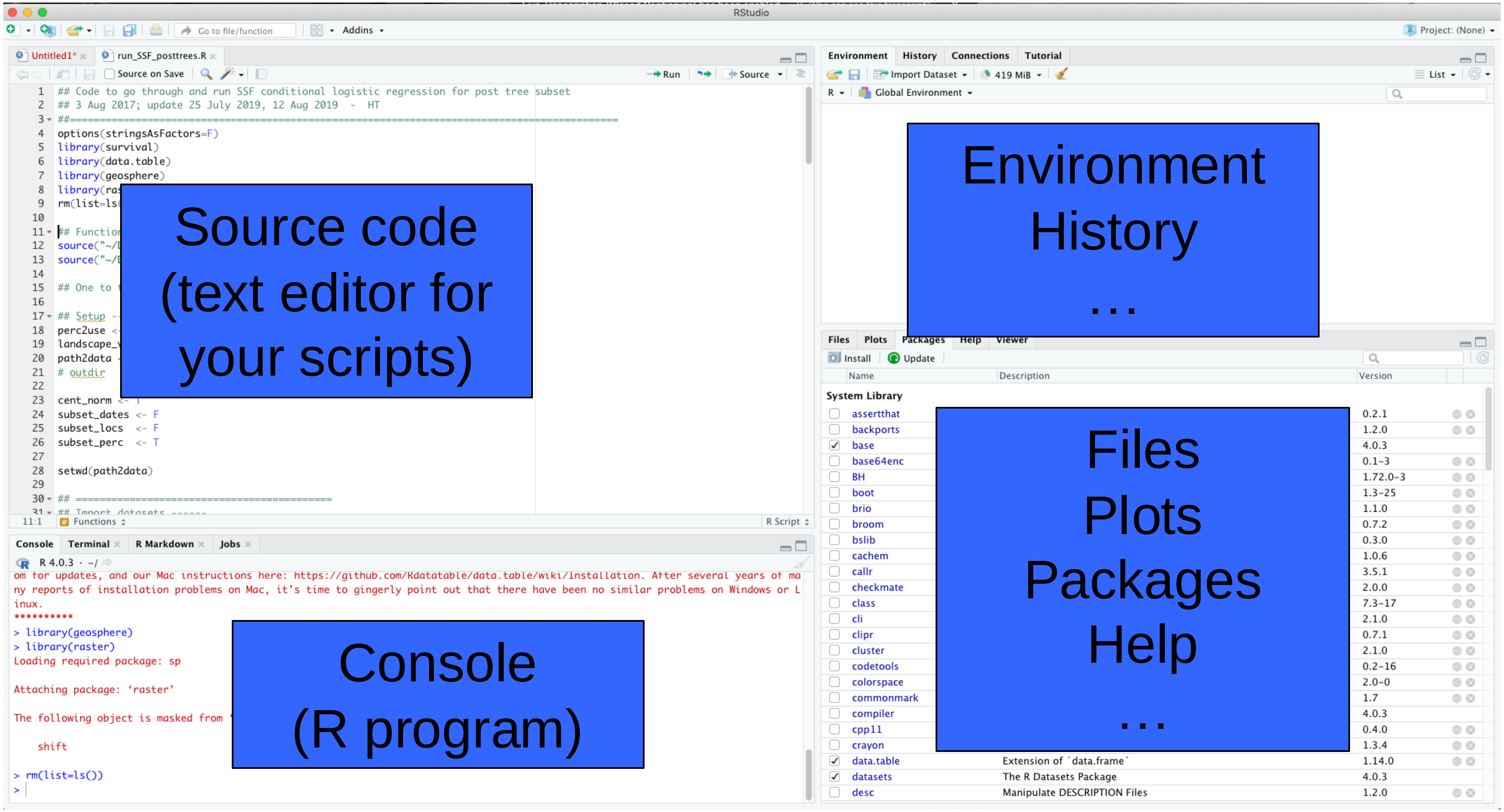
Recommendations

- **Supplement, Don’t Replace:** Use AI as a supplementary tool rather than a primary source for solving problems.
- **Seek Feedback:** Cross-check AI-generated outputs with peers, mentors, and authoritative resources.
- **Continuous Learning:** Engage with educational material and practice independently to strengthen your understanding and skills.

[Adapted from ChatGPT 4o output with prompt: *Urge caution to students on the use of ChatGPT and other LLMs for data analysis and programming*]

Using RStudio

- Provides an easier user interface that runs R for you (still need to have R installed too – R and Rstudio are distinct)
- Range of useful tools and capabilities:
  - Good text editor with syntax highlighting
  - Visualises workspace, plots, packages, data
  - No need to switch between programs (e.g. text editor and R)
- Also used in the textbook



Source code (text editor for your scripts)

```
## Code to go through and run SSF conditional logistic regression for post tree subset
## 3 Aug 2017; update 25 July 2019, 12 Aug 2019  -  HT
##=============================================================================
options(stringsAsFactors=F)
library(survival)
library(data.table)
library(geosphere)
library(ras
rm(list=ls(

## Functio
source("~/D
source("~/D

## One to 

## Setup -
perc2use <
landscape_v
path2data <
# outdir

cent_norm <- T
subset_dates <- F
subset_locs  <- F
subset_perc  <- T

setwd(path2data)

## ========================================
## Import datasets ------
```

Environment

History

…

Console

(R program)

```
om for updates, and our Mac instructions here: https://github.com/Rdatatable/data.table/wiki/Installation. After several years of ma
ny reports of installation problems on Mac, it's time to gingerly point out that there have been no similar problems on Windows or L
inux.
**********
> library(geosphere)
> library(raster)
Loading required package: sp

Attaching package: ‘raster’

The following object is masked from '

    shift

> rm(list=ls())
> 
```

Files

Plots

Packages

Help

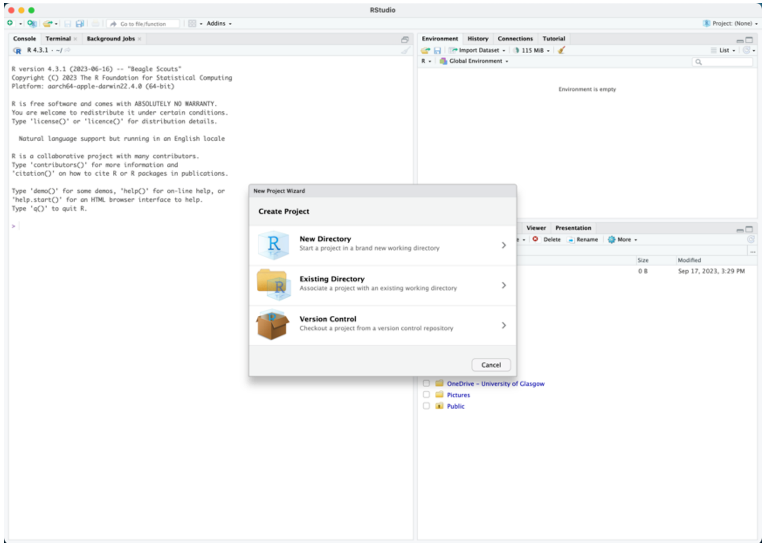
…

System Library

	Name	Description	Version
□	assertthat		0.2.1
□	backports		1.2.0
☑	base		4.0.3
□	base64enc		0.1-3
□	BH		1.72.0-3
□	boot		1.3-25
□	brio		1.1.0
□	broom		0.7.2
□	bslib		0.3.0
□	cachem		1.0.6
□	callr		3.5.1
□	checkmate		2.0.0
□	class		7.3-17
□	cli		2.1.0
□	clipr		0.7.1
□	cluster		2.1.0
□	codetools		0.2-16
□	colorspace		2.0-0
□	commonmark		1.7
□	compiler		4.0.3
□	cpp11		0.4.0
□	crayon		1.3.4
☑	data.table	Extension of `data.frame`	1.14.0
☑	datasets	The R Datasets Package	4.0.3
□	desc	Manipulate DESCRIPTION Files	1.2.0

Open RStudio and prepare your workspace

1. In your file explorer, create a new directory called Intro_to_R
2. Launch Rstudio
3. Start a new Project by going to File > New Project
4. Then select New Directory > New Project
5. Give it a directory name, e.g. “Day1”, as subdirectory of “Intro_to_R”
6. Start a new R script by going to File > New file > R script



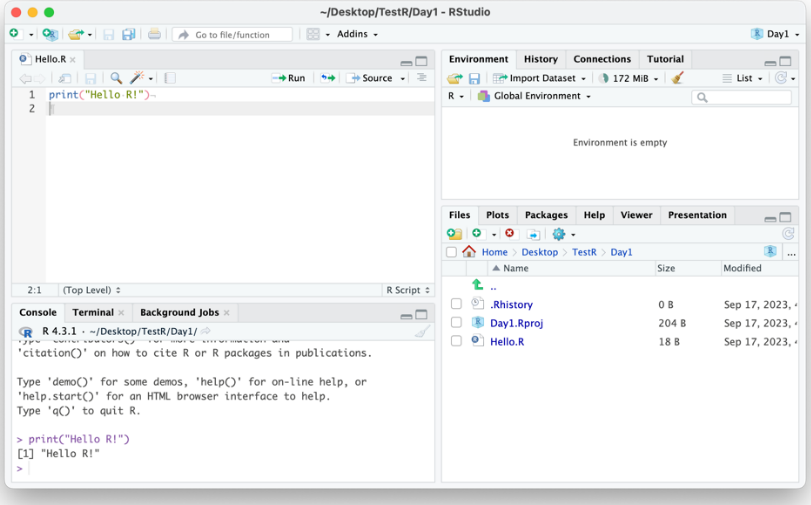
```
R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin22.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

New Project Wizard

Create Project

- **New Directory** Start a project in a brand new working directory
- **Existing Directory** Associate a project with an existing working directory
- **Version Control** Checkout a project from a version control repository

Name	Size	Modified
	0 B	Sep 17, 2023, 3:29 PM
OneDrive – University of Glasgow		
Pictures		
Public		



```r
print("Hello R!")
```

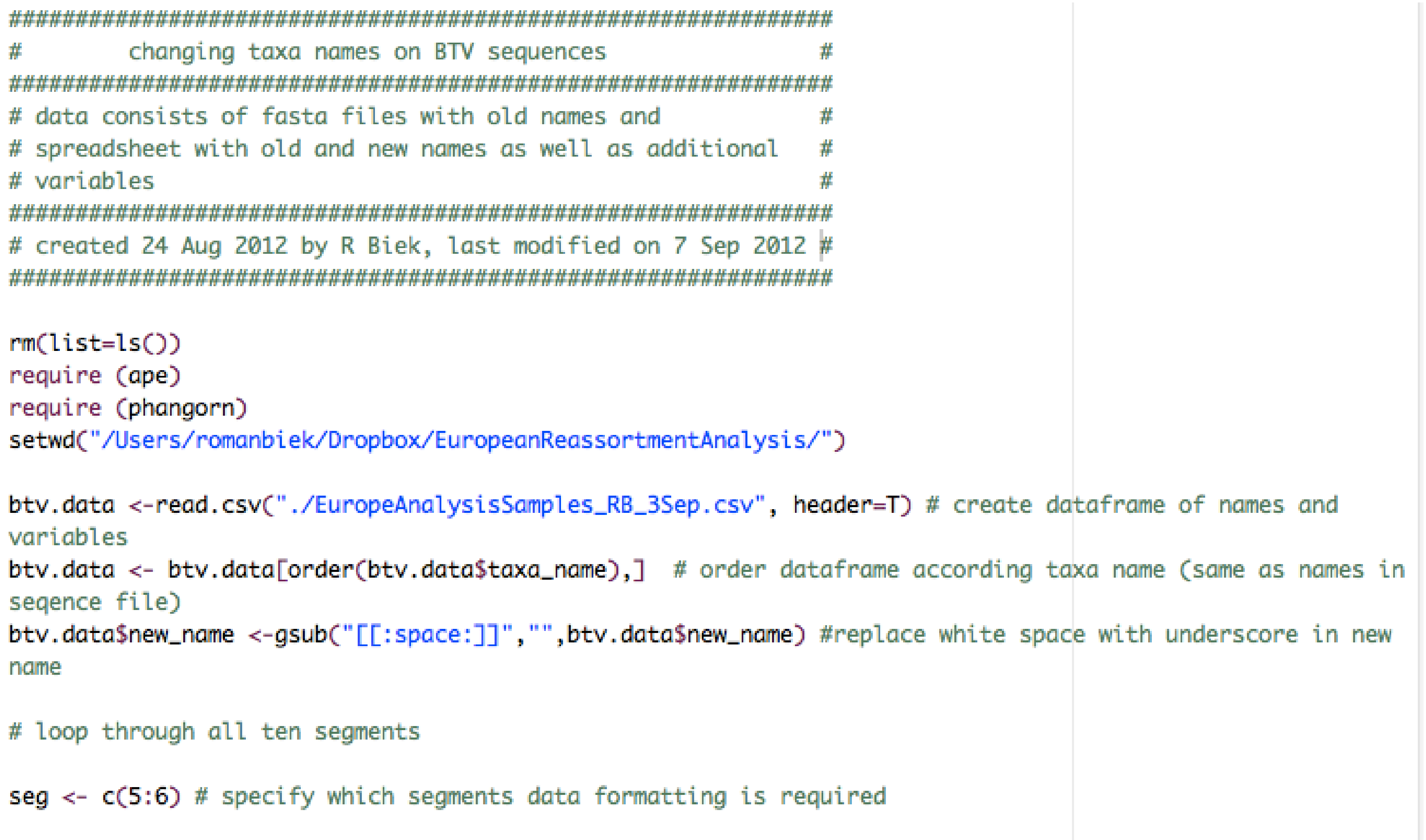
```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> print("Hello R!")
[1] "Hello R!"
>
```

Name	Size	Modified
..		
.Rhistory	0 B	Sep 17, 2023,
Day1.Rproj	204 B	Sep 17, 2023,
Hello.R	18 B	Sep 17, 2023,

Annotating scripts

- Anything after # symbol is ignored by R (until the next line) => use it to add comments to your code
- Critically important for remembering what your script does and why
- Start with informative header (date, author, what the purpose of the script is) and continue commenting throughout

Script example



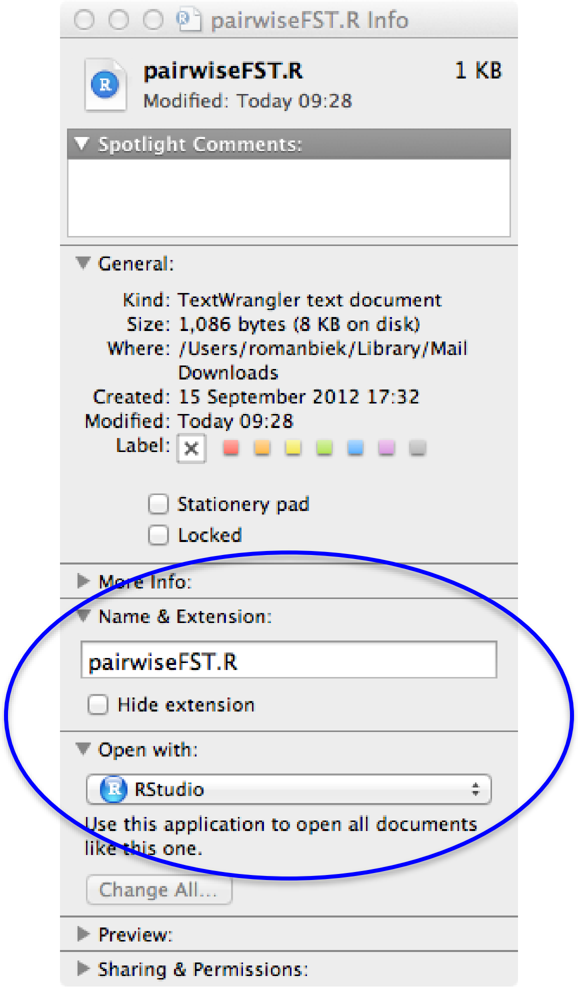
```
##############################################################
#        changing taxa names on BTV sequences                #
##############################################################
# data consists of fasta files with old names and            #
# spreadsheet with old and new names as well as additional   #
# variables                                                  #
##############################################################
# created 24 Aug 2012 by R Biek, last modified on 7 Sep 2012 #
##############################################################

rm(list=ls())
require (ape)
require (phangorn)
setwd("/Users/romanbiek/Dropbox/EuropeanReassortmentAnalysis/")

btv.data <-read.csv("./EuropeAnalysisSamples_RB_3Sep.csv", header=T) # create dataframe of names and
variables
btv.data <- btv.data[order(btv.data$taxa_name),]  # order dataframe according taxa name (same as names in
seqence file)
btv.data$new_name <-gsub("[[:space:]]","",btv.data$new_name) #replace white space with underscore in new
name

# loop through all ten segments

seg <- c(5:6) # specify which segments data formatting is required
```

File naming

Start with course name	IntroR
Add chapter or session name	Day1
Add your name or ID number	223572
End with file extension	.R or .Rmd
Use underscores instead of spaces	

e.g. IntroR_Day1_223572.R

Saving scripts



pairwiseFST.R Info

pairwiseFST.R 1 KB
Modified: Today 09:28

Spotlight Comments:

General:

Kind: TextWrangler text document
Size: 1,086 bytes (8 KB on disk)
Where: /Users/romanbiek/Library/Mail Downloads
Created: 15 September 2012 17:32
Modified: Today 09:28
Label: ☒

□ Stationery pad
□ Locked

More Info:

Name & Extension:

pairwiseFST.R

□ Hide extension

Open with:

RStudio

Use this application to open all documents like this one.

Preview:

Sharing & Permissions:

- Save with .R (or .r) extension
- Need to tell your computer which program to use for such files, i.e. R Studio (not R!)
- Usually no need to save your workspace (say ‘no’ when R Studio asks you)

On a Mac, use Finder-> File-> Get Info

Notes on filing system

- Create a logical system of nested folders

e.g. `~/Documents/Msc/KRS/Intro_to_R/Day1`

(note: good habit to avoid spaces in names)

- Once you have a system, stick to it!

- Avoid creating lots of folders on your desktop (use links or shortcuts instead)

R packages

- Also called ‘libraries’ in R
- Provide additional functions not included in the R base package
- Need to be installed only once
  - May require other packages, therefore always ‘Install dependencies’
  - Should be updated occasionally
- Need to be called exactly once in every script in which they will be used (usually at the start of the script), with `library(<name>)`

Installing packages

- Can be done within R Studio
- Go to R Studio *Tools > Install Packages* or in the console type

  ```
  install.packages("<name>", dep=TRUE)
  ```

  ⚠ note the use of quotes around the package name

- Try this out now and install the package *`ggplot2`*

If you are having problems let us know after the lecture

A few helpful things in R Studio

- Choosing directories and path
- Setting the working directory
- Using command history
- Using tab to auto-complete names

Naming objects in R

```
compensation <- read.csv("compensation.csv")
```

- Names can consist of any alphanumeric character as well as "." and "_" (but no spaces!)
- Can't start with number (i.e. "3_comp.dat")
- Case sensitive (so Comp_dat ≠ comp_dat)

Importing data as part of your script

- Move the data (.csv) file into the same folder where your R script is saved, i.e. your “Day1” Project folder
- Make sure your working directory to this folder by typing `getwd()` into your console; it should print a file path ending in “…/Day1”. If not, menu *Session -> Set Working Directory -> To Source File Location*
- With this approach, there is no need to specify the full path to the data file

```
compensation <- read.csv(“compensation.csv”)
```

When to stick with base R

Important for other KRS sessions

1) **Data import**

- Stick with R’s built-in **`read.csv()`** function
- Different from `read_csv()` in package readr, which does some (sometimes bad) automated formatting
- When using “import dataset” button in R Studio stick with first option ‘from text (base)’

2) **Plotting graphs**

- For quick plotting, use base function `plot()`

When to stick with base R

3) **Looking inside data frames**

you should know

- The use of the $ sign to select columns e.g.

```
Compensation$Root
```

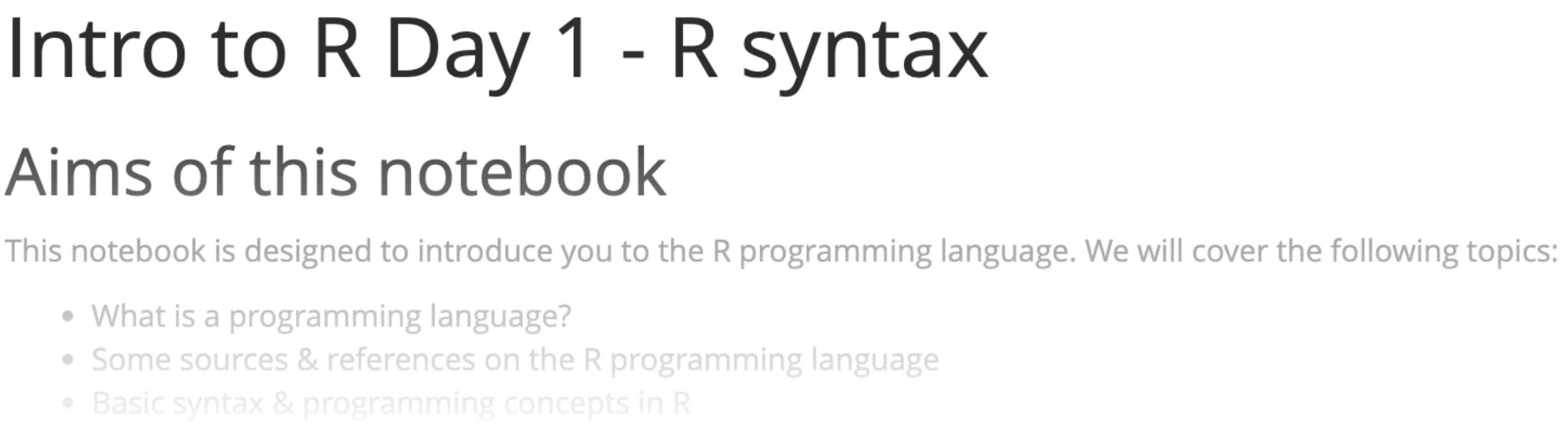
- The use of square brackets for indexing e.g.

```
Compensation[1,”Root”]
```

returns the value of the first row in the column ‘Root’

=> indexing works by *[rows, columns]*

Hands-on: getting used to R's syntax and basic features [Not assessed!]



Intro to R Day 1 - R syntax

Aims of this notebook

This notebook is designed to introduce you to the R programming language. We will cover the following topics:

- What is a programming language?
- Some sources & references on the R programming language
- Basic syntax & programming concepts in R

Working through chapters 1 & 2 in “Getting started with R”

- You should have R and R Studio installed so can start at chapter 1.4
- When you get to chapter 2, **start a new script** (in general for this course, create one folder per session and one script per chapter)
- Datasets for the book are available on Moodle (under ‘R resources’)

Home assignments

- Finish working through chapters 1 & 2
- You should hand in two R scripts, one for each chapter (do not hand in the R Syntax Demo!)
- Upload both scripts by next Thursday 3rd, using the Moodle upload portal
- Need to complete short Day 1 quiz (not assessed) before you can upload your scripts

Break(s)

- Take a break now, or whenever you need one.