



R for bioinformatics (or anything else)

Cait Harrigan & Gabi Morgenshtern



Before we begin....

pollev.com/charrigan888

Please vote on what you want to get out of today's workshop



What is R? What is Rstudio?

R is a programming language - a set of rules for how to tell a computer to do something

RStudio is an interface for coding with R. You may see it called an “environment” or an “IDE”



The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains an R script with comments and code for downloading and processing data from STRING and GOSlim databases. The code includes comments in German and English, and R code to install and load the 'readr' package.
- Console:** Shows the R version (3.3.3), copyright information, and a disclaimer. It also displays the R startup message in German and English.
- Environment:** Shows the 'Global Environment' with a search bar and a message 'Environment is empty'.
- Files:** Shows a file explorer view of the 'scripts' directory, listing files such as '.gitignore', 'basics-intro.R', 'data', 'lil-bioinf.Rproj', 'plotting.R', 'preparingData.R', and 'README.md'.

```
# This script works through downloading, processing, and integrating information from the
# STRING and GOSlim databases on Saccharomyces cerevisiae, in preparation for visualizing
# a network of high-confidence yeast genes with the "mitotic cell cycle" GOSlim annotation.
#
# Adapted from Boris Steipe's material for BCH441
#
# Dataset Access:
#   STRING data source:
#   Download page: https://string-db.org/cgi/download.pl?species\_text=Saccharomyces\_cerevisiae
#   Data: (20.8 mb) https://string-db.org/download/protein.links.full.v11/4932.protein.links.full.v11.txt.gz
#
#   GOSlim data source:
#   Info page: http://www.geneontology.org/page/go-slim-and-subset-guide
#   Data: (3 mb) https://downloads.yeastgenome.org/curation/literature/go\_slim\_mapping.tab
#
22 ~ if (!require(readr, quietly = TRUE)) {
23   install.packages("readr")
24   library(readr)
25 }
26
25.2 (Top Level) >
```

R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

> |

The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

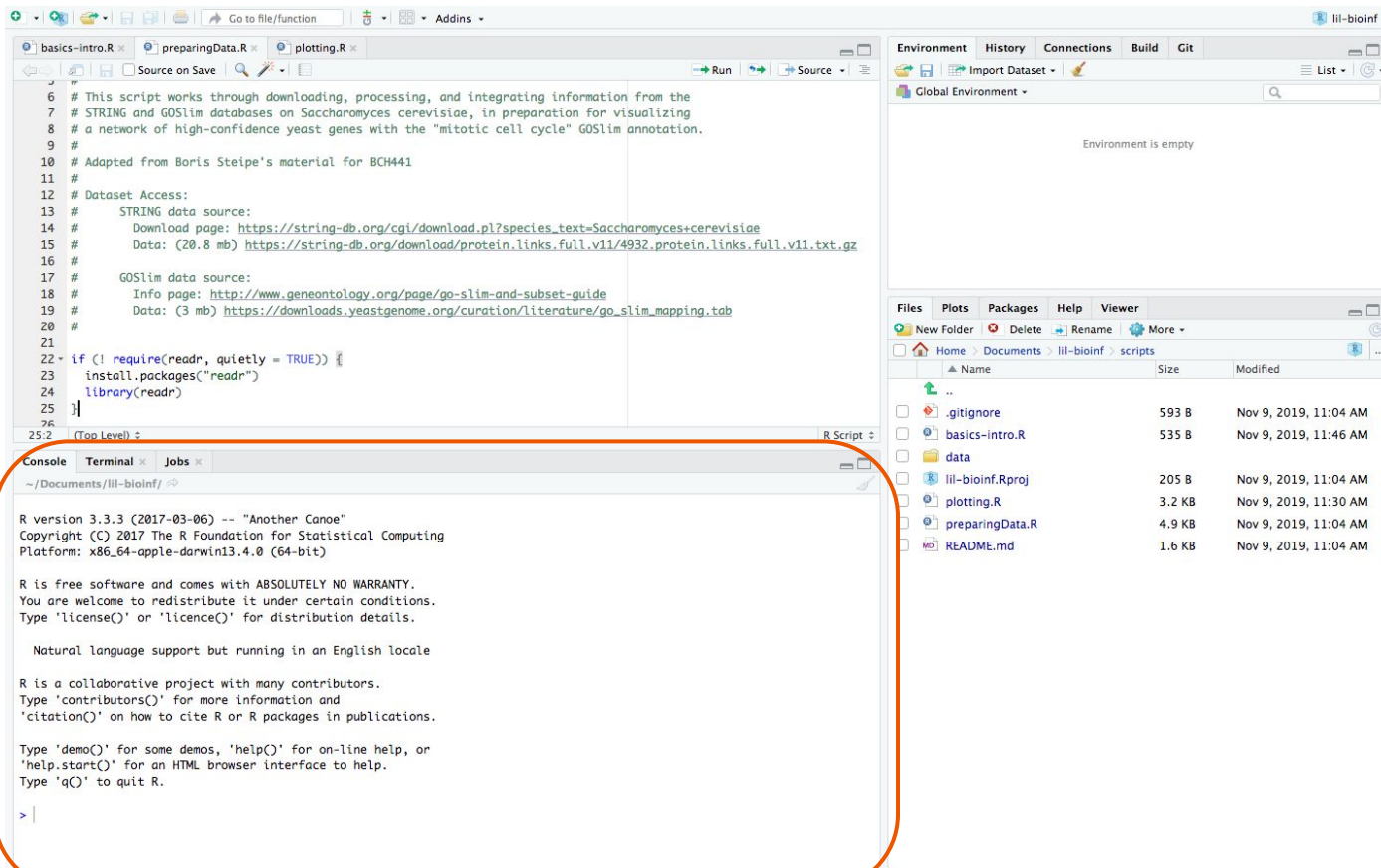
- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts



The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts

The screenshot displays the RStudio interface with three main panels:

- Script Editor (Top Left):** Contains an R script with comments and code for downloading and processing data. The code includes comments about the STRING and GOSlim databases and the use of the readr package. The script is highlighted with an orange border.


```

6 # This script works through downloading, processing, and integrating information from the
7 # STRING and GOSlim databases on Saccharomyces cerevisiae, in preparation for visualizing
8 # a network of high-confidence yeast genes with the "mitotic cell cycle" GOSlim annotation.
9 #
10 # Adapted from Boris Steipe's material for BCH441
11 #
12 # Dataset Access:
13 #   STRING data source:
14 #   Download page: https://string-db.org/cgi/download.pl?species_text=Saccharomyces_cerevisiae
15 #   Data: (20.8 mb) https://string-db.org/download/protein.links.full.v11/4932.protein.links.full.v11.txt.gz
16 #
17 #   GOSlim data source:
18 #   Info page: http://www.geneontology.org/page/go-slim-and-subset-guide
19 #   Data: (3 mb) https://downloads.yeastgenome.org/curation/literature/go_slim_mapping.tab
20 #
21
22 if (!require(readr, quietly = TRUE)) {
23   install.packages("readr")
24   library(readr)
25 }
      
```
- Console (Bottom Left):** Shows the R version (3.3.3) and the R Foundation's copyright notice. It also displays the R license and the natural language support but running in an English locale.


```

R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

> |
      
```
- Files Panel (Bottom Right):** Shows a list of files in the current project directory. The files are:

Name	Size	Modified
..		
.gitignore	593 B	Nov 9, 2019, 11:04 AM
basics-intro.R	535 B	Nov 9, 2019, 11:46 AM
data		
lil-bioinf.Rproj	205 B	Nov 9, 2019, 11:04 AM
plotting.R	3.2 KB	Nov 9, 2019, 11:30 AM
preparingData.R	4.9 KB	Nov 9, 2019, 11:04 AM
README.md	1.6 KB	Nov 9, 2019, 11:04 AM

The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains an R script with comments and code for downloading and processing data from STRING and GOSlim databases. The code includes comments in German and English, and R code to install and load the 'readr' package.
- Environment:** A panel on the right showing the 'Global Environment' which is currently empty.
- Files:** A panel on the right showing a list of files in the 'scripts' directory, including '.gitignore', 'basics-intro.R', 'data', 'lil-bioinf.Rproj', 'plotting.R', 'preparingData.R', and 'README.md'.
- Console:** Displays the R version (3.3.3), copyright information, and a message about the English locale.

```

6 # This script works through downloading, processing, and integrating information from the
7 # STRING and GOSlim databases on Saccharomyces cerevisiae, in preparation for visualizing
8 # a network of high-confidence yeast genes with the "mitotic cell cycle" GOSlim annotation.
9 #
10 # Adapted from Boris Steipe's material for BCH441
11 #
12 # Dataset Access:
13 #   STRING data source:
14 #   Download page: https://string-db.org/cgi/download.pl?species\_text=Saccharomyces\_cerevisiae
15 #   Data: (20.8 mb) https://string-db.org/download/protein.links.full.v11/4932.protein.links.full.v11.txt.gz
16 #
17 #   GOSlim data source:
18 #   Info page: http://www.geneontology.org/page/go-slim-and-subset-guide
19 #   Data: (3 mb) https://downloads.yeastgenome.org/curation/literature/go\_slim\_mapping.tab
20 #
21
22 if (!require(readr, quietly = TRUE)) {
23   install.packages("readr")
24   library(readr)
25 }
  
```

R version 3.3.3 (2017-03-06) -- "Another Canoe"
 Copyright (C) 2017 The R Foundation for Statistical Computing
 Platform: x86_64-apple-darwin13.4.0 (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.

> |

The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains an R script with comments and code for downloading and processing data from STRING and GOSlim databases. The code includes comments in German and English, and R code to install and load the 'readr' package.
- Console:** Shows the R version (3.3.3), copyright information, and a disclaimer. It also displays the R startup message in German and English.
- Files Panel:** Lists files in the current directory, including .gitignore, basics-intro.R, data, lil-bioinf.Rproj, plotting.R, preparingData.R, and README.md.

```
# This script works through downloading, processing, and integrating information from the
# STRING and GOSlim databases on Saccharomyces cerevisiae, in preparation for visualizing
# a network of high-confidence yeast genes with the "mitotic cell cycle" GOSlim annotation.
#
# Adapted from Boris Steipe's material for BCH441
#
# Dataset Access:
#   STRING data source:
#   Download page: https://string-db.org/cgi/download.pl?species\_text=Saccharomyces\_cerevisiae
#   Data: (20.8 mb) https://string-db.org/download/protein.links.full.v11/4932.protein.links.full.v11.txt.gz
#
#   GOSlim data source:
#   Info page: http://www.geneontology.org/page/go-slim-and-subset-guide
#   Data: (3 mb) https://downloads.yeastgenome.org/curation/literature/go\_slim\_mapping.tab
#
22 ~ if (!require(readr, quietly = TRUE)) {
23   install.packages("readr")
24   library(readr)
25 }
26
25.2 (Top Level) >
```

R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

> |

Name	Size	Modified
..		
.gitignore	593 B	Nov 9, 2019, 11:04 AM
basics-intro.R	535 B	Nov 9, 2019, 11:46 AM
data		
lil-bioinf.Rproj	205 B	Nov 9, 2019, 11:04 AM
plotting.R	3.2 KB	Nov 9, 2019, 11:30 AM
preparingData.R	4.9 KB	Nov 9, 2019, 11:04 AM
README.md	1.6 KB	Nov 9, 2019, 11:04 AM

The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains an R script with comments and code for downloading and processing data from STRING and GOSlim databases. The code includes comments in German and English, and R code to install the 'readr' package and load it.
- Console:** Shows the R version (3.3.3), copyright information, and a disclaimer. It also displays the R startup message in German and English.
- Environment:** Shows the 'Global Environment' with a search bar and a list of objects. The 'lil-bioinf' package is highlighted in the top right corner.
- Files:** Shows a file explorer view of the 'lil-bioinf' directory, listing files such as '.gitignore', 'basics-intro.R', 'data', 'lil-bioinf.Rproj', 'plotting.R', 'preparingData.R', and 'README.md'.

The console:

- run R code

Scripts:

- write and save R code
- can be sourced to re-use useful code

Environment:

- what information is available to the console

Files:

- files on your machine

Project:

- a certain set-up for a group of scripts

Objects



Objects let us talk about the things we want our code to do. Objects are part of a useful paradigm that lets us turn abstract concepts into concrete actions.

Some examples of objects: a number, a string of characters, a list, or even a function.

Abstract concept



Concrete action

Objects



Objects let us talk about the things we want our code to do. Objects are part of a useful paradigm that lets us turn abstract concepts into concrete actions.

Some examples of objects: a number, a string of characters, a list, or even a function.

Abstract concept



Concrete action

What's the average speed
of these three snails?

$$\frac{0.013\text{m/s} + 0.008\text{m/s} + 0.011\text{m/s}}{3}$$

Objects



Objects let us talk about the things we want our code to do. Objects are part of a useful paradigm that lets us turn abstract concepts into concrete actions.

Some examples of objects: a number, a string of characters, a list, or even a function.

Abstract concept



Concrete action

What's the average speed of these three snails?

```
Console  Terminal x  Jobs x
~/Documents/lil-bioinf/
>
> (0.013 + 0.008 + 0.011) / 3
[1] 0.01066667
> |
```

Variables



We can store objects in variables. They act as a shorthand name for the object that they refer to.

We use the assignment operator `<-` to set the value of a variable

We can manipulate variables in the same way that we did the objects they refer to.

```
Console Terminal x Jobs x
~/Documents/lil-bioinf/ ↗
> num_snails <- 3
> num_snails
[1] 3
```

```
Console Terminal x Jobs x
~/Documents/lil-bioinf/ ↗
>
> num_snails <- 3
> speed_snail1 <- 0.013
> speed_snail2 <- 0.008
> speed_snail3 <- 0.011
> (speed_snail1 + speed_snail2 + speed_snail3) / num_snails
[1] 0.01066667
```

Functions

We use functions to write reusable code. They let us make modifications to objects, or get information about them.

A function takes input parameters, and returns output.



```
Console Terminal x Jobs x
~/Documents/lil-bioinf/ ↗
> fastest_snail <- function(speeds){
+
+   snail_num <- which(speeds == max(speeds))
+   message <- sprintf("The fastest is snail number %s", snail_num)
+
+   return(message)
+ }
>
> fastest_snail(snail_speeds)
[1] "The fastest is snail number 1"
> |
```

Function definition

Function call - the **input** is the object `snail_speeds` and the **return value** is a message about the snails

Let's dive in!

github.com/harrig12/lil-bioinf

The screenshot shows the GitHub repository page for `harrig12 / lil-bioinf`. The repository is described as an "R workshop for bioinformatics (and more!)" and is linked to `https://harrig12.github.io/lil-bioinf`. It has 10 commits, 4 branches, 0 packages, 0 releases, 1 environment, and 2 contributors. The repository is currently on the `master` branch. A recent commit by `gabmorg` is shown, titled "Merge branch 'master' of https://github.com/harrig12/lil-bioinf". The commit details show a list of files added or modified, including `data`, `.gitignore`, `README.md`, `lil-bioinf.Rproj`, `plotting.R`, and `preparingData.R`. The `README.md` file is highlighted, showing the title "R workshop for bioinformatics (and more!)" and a link to the repository's archive.

harrig12 / lil-bioinf

Unwatch 2 ★ Star 0 🍴 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 1 Wiki Security Insights Settings

R workshop for bioinformatics (and more!) <https://harrig12.github.io/lil-bioinf> Edit

Manage topics

10 commits 4 branches 0 packages 0 releases 1 environment 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download -

gabmorg Merge branch 'master' of https://github.com/harrig12/lil-bioinf

File	Commit Message	Time
data	added STRING RDS file, and code for prepping it and scCCnet	10 days ago
.gitignore	init	10 days ago
README.md	added STRING RDS file, and code for prepping it and scCCnet	9 days ago
lil-bioinf.Rproj	init	9 days ago
plotting.R	rm significance testing	9 days ago
preparingData.R	added STRING RDS file, and code for prepping it and scCCnet	9 days ago

Clone with HTTPS Use SSH

Use Git or checkout with SVN using the web URL.

<https://github.com/harrig12/lil-bioinf>

Open in Desktop Download ZIP

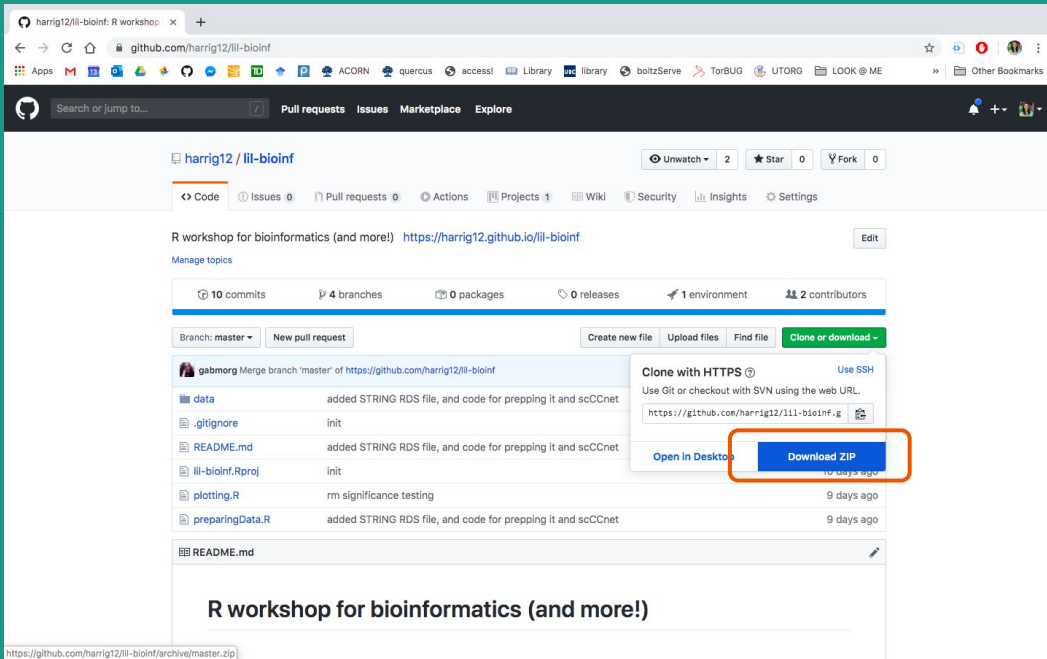
README.md

R workshop for bioinformatics (and more!)

<https://github.com/harrig12/lil-bioinf/archive/master.zip>

Let's dive in!

github.com/harrig12/lil-bioinf



harrig12 / lil-bioinf

Unwatch 2 Star 0 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 1 Wiki Security Insights Settings

R workshop for bioinformatics (and more!) <https://harrig12.github.io/lil-bioinf> Edit

Manage topics

10 commits 4 branches 0 packages 0 releases 1 environment 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download -

gabmorg Merge branch 'master' of <https://github.com/harrig12/lil-bioinf>

- data added STRING RDS file, and code for prepping it and scCNet
- .gitignore init
- README.md added STRING RDS file, and code for prepping it and scCNet
- lil-bioinf.Rproj init
- plotting.R rm significance testing 9 days ago
- preparingData.R added STRING RDS file, and code for prepping it and scCNet 9 days ago

README.md

R workshop for bioinformatics (and more!)

<https://github.com/harrig12/lil-bioinf/archive/master.zip>

Scripts



Save code for reusability.

For example - we saw the basics-intro.R script

Scripts can be executed with “run” and this will *sequentially* step through all the code in the script.

This can be done without explicitly opening an R session (ex. from the command line). Running the script will start a fresh one, different from your console session.

Get more functionality



CRAN and Bioconductor are both great places to get *maintained* R packages. After you install a new package, load it at the start of your scripts or in the console with the `library()` function

Data structures

To hold and manipulate data, we tend to use particular objects. [Dataframes](#) are great for their flexibility and versatility. You may also see data in lists, matrices, etc.

Certain packages implement their own data structures with “slots” for different pieces of information.

For now, we'll focus on dataframes.

Dataframes can be constructed with named or un-named columns.

If you try to add data that is mismatched in length to other columns in the dataframe, you might get an error.

```
Console Terminal x Jobs x
~/Documents/lil-bioinf/
> snail_data <- data.frame(speed = snail_speeds, shell_colour = snail_shells_factor)
> print(snail_data)
  speed shell_colour
1 0.013         brown
2 0.008         orange
3 0.011         brown
```

```
Console Terminal x Jobs x
~/Documents/lil-bioinf/
> data.frame(1:5, c("one", "two"))
Error in data.frame(1:5, c("one", "two")) :
  arguments imply differing number of rows: 5, 2
```

Useful functions with dataframes



Try each of these functions out on [snail_data](#). What does each output?

Function	Output
<code>head()</code>	
<code>dim()</code>	
<code>colnames()</code>	
<code>summary()</code>	
<code>table()</code>	
<code>plot()</code>	

Useful functions with dataframes



Try each of these functions out on [snail_data](#). What does each output?

Function	Output
<code>head()</code>	A preview of the dataframe
<code>dim()</code>	Dimensions of the dataframe
<code>colnames()</code>	Column names of the dataframe
<code>summary()</code>	Summary of each column
<code>table()</code>	Counts of matching entries
<code>plot()</code>	A plot

Accessing and subsetting dataframes



Subsetting is an important part of data manipulation!

For columns, we call this “selecting”. For rows, we call this “filtering” or “subsetting”.

In base R, there are a few ways to do this. In the wild, you may see the following forms:

```
snail_data$speed
```

```
snail_data[["speed"]]
```

```
snail_data[,1]
```

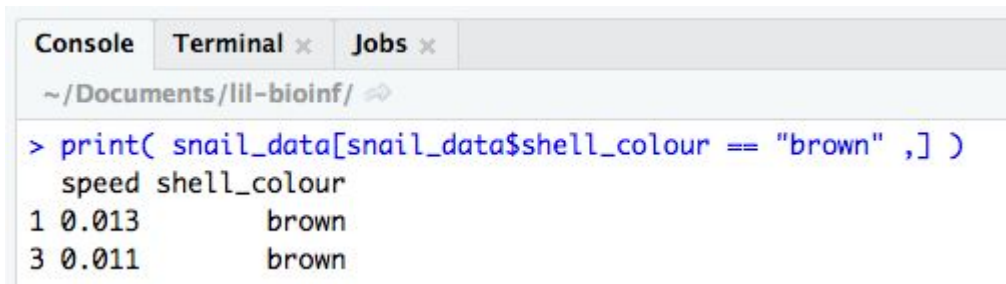
You may find it easier to use `subset()`

Accessing and subsetting dataframes

Subsetting is an important part of data manipulation!

For columns, we call this “selecting”. For rows, we call this “filtering” or “subsetting”.

A great way filter rows is to use boolean vectors, that record whether a desired property is true or not



The screenshot shows a R console window with tabs for 'Console', 'Terminal', and 'Jobs'. The current directory is ~/Documents/lil-bioinf/. The command executed is `print(snail_data[snail_data$shell_colour == "brown",])`, which filters the 'snail_data' data frame for rows where 'shell_colour' is 'brown'. The output shows two rows with their indices, 'speed' values, and 'shell_colour' values.

```
> print( snail_data[snail_data$shell_colour == "brown" ,] )
  speed shell_colour
1 0.013         brown
3 0.011         brown
```

Exit survey

pollev.com/charrigan888

Thanks!

These slides and workshop materials are available at harrig12.github.io/lil-bioinf
