

R-reproducible workflows

Half-day workshop

(after 2.5 days of intense R training!!)



Brendan Palmer, University College Cork
Adam Kane, University College Dublin
Enrico Pirotta, Washington State University

Putting the pieces together

- Project structure
 - Naming conventions
 - Scripted workflows
 - R Markdown
 - Reproducible research

What is reproducible research?

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

How is research presented?

Theses

Papers



Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses

Brendan A. O'Farrior,^a Daniel Schmidt-Martin,^a Zoya Dimitrova,^b Pavel Skums,^b Orla Crosbie,^c Elizabeth Kenny-Walsh,^c Liam J. Fanning^d

^aMolecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland; ^bDivision of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; ^cDepartment of Hepatology, Cork University Hospital, Cork, Ireland

ABSTRACT
Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically recognized as the major antigenic target of the humoral immune response. HVR1 is also known to undergo rapid sequence evolution during chronic infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 variants in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the predominant dominance of a single variant within a narrow sequence space.

IMPORTANCE
Hepatitis C virus is often asymptomatic, and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 variants in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the predominant dominance of a single variant within a narrow sequence space.

Hepatitis C virus (HCV) infection is a global health burden and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 variants in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the predominant dominance of a single variant within a narrow sequence space.

HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defines the major genotypes 1a and 1b and 67 subtypes within genotypes 1 and 5 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent capacity for de novo hypervariable recombination, is responsible for much of this variability. These mutations are located within the envelope glycoprotein E2. The greatest heterogeneity has been identified at the 27-amino-acid HVR1 (residues 456 to 482), located near the C-terminal end of the E2 glycoprotein (8). Recent studies indicated that the central region of E2 (residues 456 to 656) is globular and surprisingly compact, whereas the first 80 amino acids (including

HVR1) lack this structural rigidity (9). This observation is consistent with a hypothesis that a protein with reduced conformational plasticity in part accounts for high-density glycoprotein enhancement by scavenger receptor class B type I (SR-BI) interactions and is itself targeted by neutralizing antibodies (nAb) (10).

Mutational flexibility at HVR1 was characterized soon after the initial identification of HCV (8, 17). Rapid mutational change of HVR1 has been documented over weeks during the acute phase of infection (18). Subsequent HVR1 evolution is driven primarily by strong selective pressure with fixation of beneficial mutations (11, 18, 19). Reports examining samples collected over years to decades have documented the emergence of convergent HVR1

Received 25 November 2015; Accepted 22 December 2015
Accepted manuscript posted online 10 December 2015
Editorial decision received 10 December 2015
Editorial decision received 10 December 2015
Editor: M. S. Diamond
Associate Editor: L. J. Fanning
Editorial Reviewer: L. J. Fanning, H. H. Hwang, G. J. Liao
B.A.O. and D.S.M. contributed equally to this article.
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

Books



Talks

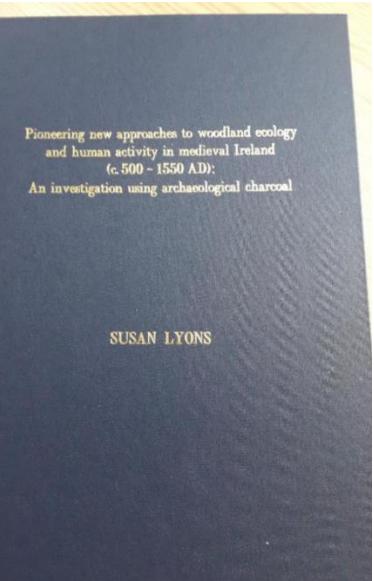


2318 J. Virol. 2016, Volume 90, Number 7

April 2016

Journal of Virology

Posters



But what does it look like under the bonnet?





I got your project and its problems galore

[Link to the video on YouTube](#)

Disclaimer



Jenny Bryan

@JennyBryan

Software engineer @rstudio, humane
#rstats, adjunct prof @UBC where I
created @STAT545, part of @opencsci

STAT
545

Home FAQ Syllabus Topics People

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R



Hadley Wickham

@hadleywickham

R, data, visualisation.

⌚ Houston, TX

🔗 hadley.nz

HADLEY WICKHAM

TEACHING CODE PERSONAL

I also teach in person workshops from time-to-time; see the [RStudio workshops page](#) for more details.

CODE

Most of my work is in the form of open source R code, which you can find on [my github](#). You can roughly divide my work into three categories: tools for data science, tools for data import, and software engineering tools.

DATA SCIENCE

- [ggplot2](#) for visualising data.
- [dplyr](#) for manipulating data.
- [tidyverse](#) for tidying data.
- [stringr](#) for working with strings.
- [lubridate](#) for working with date/times.

DATA IMPORT

- [readr](#) for reading .csv and fwf files.
- [readxl](#) for reading .xls and .xlsx files.
- [haven](#) for SAS, SPSS, and Stata files.
- [httr](#) for talking to web APIs.
- [rvest](#) for scraping websites.
- [xml2](#) for importing XML files.

SOFTWARE ENGINEERING

- [devtools](#) for general package development.
- [roxygen2](#) for in-line documentation.
- [testthat](#) for unit testing

VARIANCE EXPLAINED



David Robinson

@drob

Chief Data Scientist at [@DataCamp](#),
#rstats fan/evangelist

⌚ New York, NY

🔗 varianceexplained.org

ABOUT ME POSTS LEARN R TEXT MINING IN R INTRODUCTION TO EMPIRICAL BAYES

This is the homepage and blog of David Robinson, Chief Data Scientist at DataCamp. For more about me, [see here](#).



David Robinson

Chief Data Scientist at
DataCamp, works in R and
Python.

- ✉ Email
- ⌚ Twitter
- ⌚ Github
- 🌐 Stack Overflow

Recent Posts

Exploring college major and income: a live data analysis in R
A live screencast of an exploratory data analysis from the Tidy Tuesday series. This one explores college major and income data from 538.

October 16, 2018

Who wrote the anti-Trump New York Times op-ed? Using tidytext to find document similarity
An analysis of an anonymous op-ed in the New York Times, using document similarity metrics to match it to Twitter accounts.

September 06, 2018

Scientific debt
Introducing an analogy to 'technical debt' for data scientists.

May 10, 2018

R projects

R projects

- Here's one I made earlier.....

The screenshot shows a GitHub repository page. At the top, there is a navigation bar with links for Pull requests, Issues, Marketplace, and Explore. Below the navigation bar, the repository name is displayed as `bapalmer / project-structure-March-19`. To the right of the repository name are buttons for Watch (0), Star (0), and Fork (0). Below the repository name, there is a horizontal menu with links for Code, Issues (0), Pull requests (0), Projects (0), Wiki, Insights, and Settings. The 'Code' link is highlighted with an orange border. Below the menu, a message states "No description, website, or topics provided." On the right side of this message is an "Edit" button. Below this, there is a section for managing topics with a "Manage topics" link. At the bottom of the page, there are statistics: 2 commits, 1 branch, 0 releases, 1 contributor, and MIT license. There are also buttons for Branch: master ▾, New pull request, Create new file, Upload files, Find File, and Clone or download ▾. A red arrow points to the "Clone or download" button.

Search or jump to... / Pull requests Issues Marketplace Explore

bapalmer / project-structure-March-19

Watch 0 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

No description, website, or topics provided. Edit

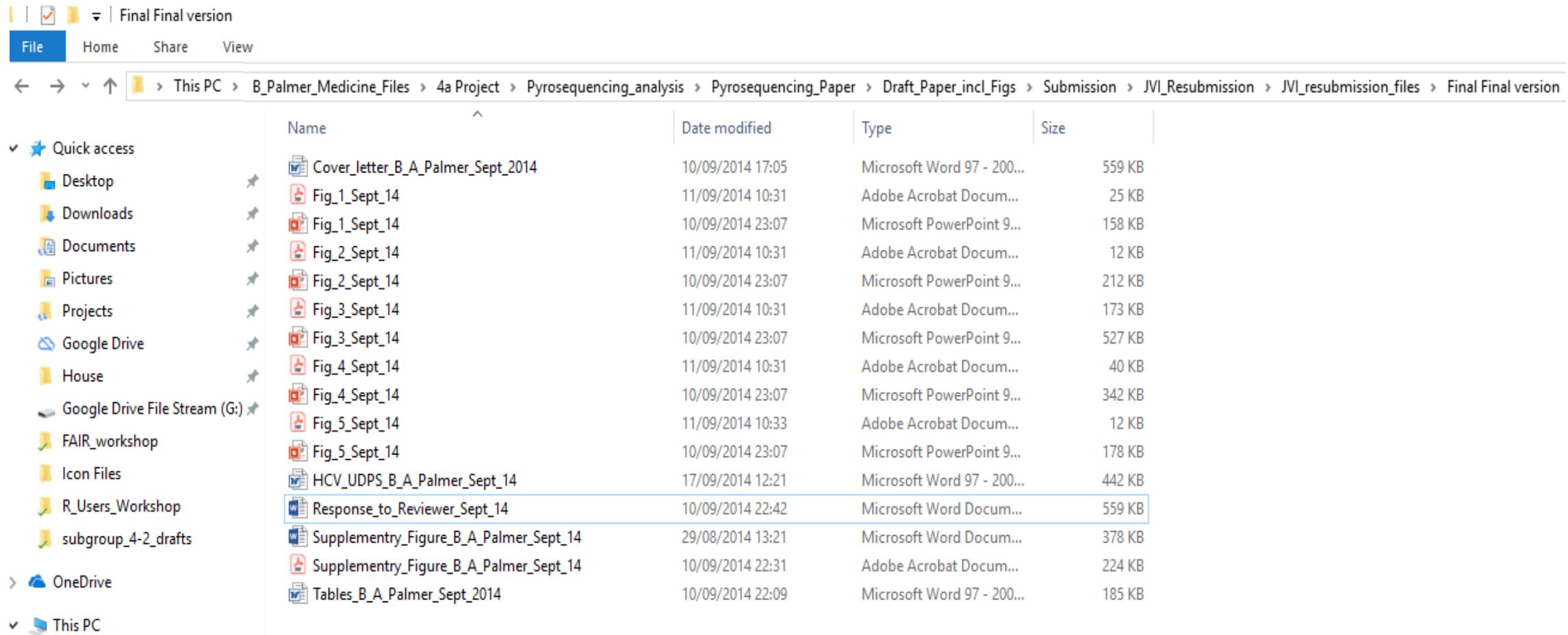
Manage topics

2 commits 1 branch 0 releases 1 contributor MIT

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

Still haven't found what I'm looking for

- Help your future-self

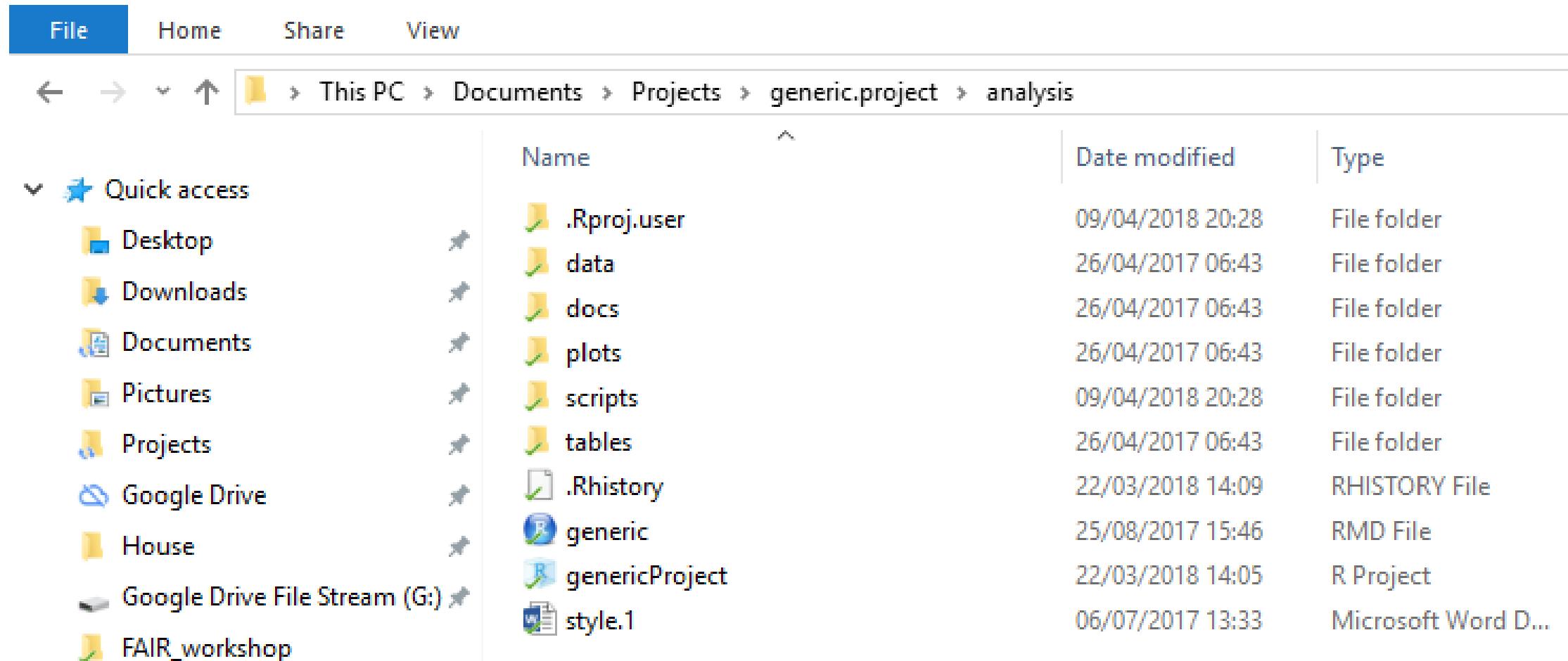


The screenshot shows a Windows File Explorer window with the following details:

- Path:** This PC > B_Palmer_Medicine_Files > 4a Project > Pyrosequencing_analysis > Pyrosequencing_Paper > Draft_Paper_incl_Figs > Submission > JVI_Resubmission > JVI_resubmission_files > Final Final version
- File Type:** Microsoft Word 97 - 2003 Document
- File List:** The table below lists the files found in the folder.

Name	Date modified	Type	Size
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05	Microsoft Word 97 - 200...	559 KB
Fig_1_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	25 KB
Fig_1_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	158 KB
Fig_2_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	12 KB
Fig_2_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	212 KB
Fig_3_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	173 KB
Fig_3_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	527 KB
Fig_4_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	40 KB
Fig_4_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	342 KB
Fig_5_Sept_14	11/09/2014 10:33	Adobe Acrobat Docum...	12 KB
Fig_5_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	178 KB
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21	Microsoft Word 97 - 200...	442 KB
Response_to_Reviewer_Sept_14	10/09/2014 22:42	Microsoft Word Docum...	559 KB
Supplementary_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21	Microsoft Word Docum...	378 KB
Supplementary_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31	Adobe Acrobat Docum...	224 KB
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09	Microsoft Word 97 - 200...	185 KB

Step 1: Define a generic project structure



The screenshot shows a Windows File Explorer window with the following details:

- File** tab is selected.
- Address Bar:** This PC > Documents > Projects > generic.project > analysis
- Left Sidebar (Quick access):** Desktop, Downloads, Documents, Pictures, Projects, Google Drive, House, Google Drive File Stream (G:), FAIR_workshop.
- Right Panel (File List):**

Name	Date modified	Type
.Rproj.user	09/04/2018 20:28	File folder
data	26/04/2017 06:43	File folder
docs	26/04/2017 06:43	File folder
plots	26/04/2017 06:43	File folder
scripts	09/04/2018 20:28	File folder
tables	26/04/2017 06:43	File folder
.Rhistory	22/03/2018 14:09	RHISTORY File
generic	25/08/2017 15:46	RMD File
genericProject	22/03/2018 14:05	R Project
style.1	06/07/2017 13:33	Microsoft Word D...

Step 2: Give your files informative names

- Make your file names machine readable, human readable and work with default ordering

The screenshot shows a Windows File Explorer window. The ribbon at the top has 'File' selected. The address bar shows the path: This PC > Documents > Projects > 18.04.27-WP3_Feeding_Trial > analysis > data. On the left, there's a 'Quick access' sidebar with links to Desktop, Downloads, Documents, Pictures, and Projects. The main area displays a list of files in the 'data' folder, sorted by name. The columns are 'Name' and 'Date modified'. The files listed are: raw_files (modified 27/09/2018 11:54), master_database (modified 12/06/2018 12:22), nutritics_food_level (modified 09/05/2018 14:25), nutritics_grouped (modified 09/05/2018 14:27), and nutritics_grouped_reduced (modified 14/05/2018 01:14). The 'nutritics_grouped_reduced' file is highlighted.

Name	Date modified
raw_files	27/09/2018 11:54
master_database	12/06/2018 12:22
nutritics_food_level	09/05/2018 14:25
nutritics_grouped	09/05/2018 14:27
nutritics_grouped_reduced	14/05/2018 01:14

Step 3: Everything in its right place

NO

Name
All unique 4a amino acid Sequences (B-N).fas
All unique 4a amino acid Sequences (B-N).meg
All_AA_haplotypes.meg
All_AA_haplotypes_with_clonal_sequences.meg
BS100_AA_with_clones
BS100_AA_with_clones.nwk
BS1000_AA_pyro&clones
BS1000_AA_pyro&clones.nwk
BS1000_AA_pyro_only
BS1000_AA_pyro_only.nwk
BS1000_Uncle_Clonal_AA
BS1000_Uncle_Clonal_AA.nwk
BS1000_Uncle_Pyro_AA
BS1000_Uncle_Pyro_AA.nwk

Yes

Documents > Projects > 18.04.27-WP3_Feeding_Trial > analysis > scripts	
Name	Date modified
01_data_import_and_tidying_master_file	02/10/2018 18:51
02_data_import_and_tidying_nutritics_grouped	19/10/2018 19:47
03_figures	17/10/2018 16:40
04_tables	22/05/2018 12:26
05_study_overview	19/10/2018 23:06
functions	13/05/2018 23:13

Outline a file naming convention

Machine readable:

- Inherent order
- Avoid spaces
- Avoid punctuation
- Remove case-sensitivity

Human readable:

- Contains info on content
- Avoid spaces
- Avoid punctuation
- Remove case sensitivity

Metadata:

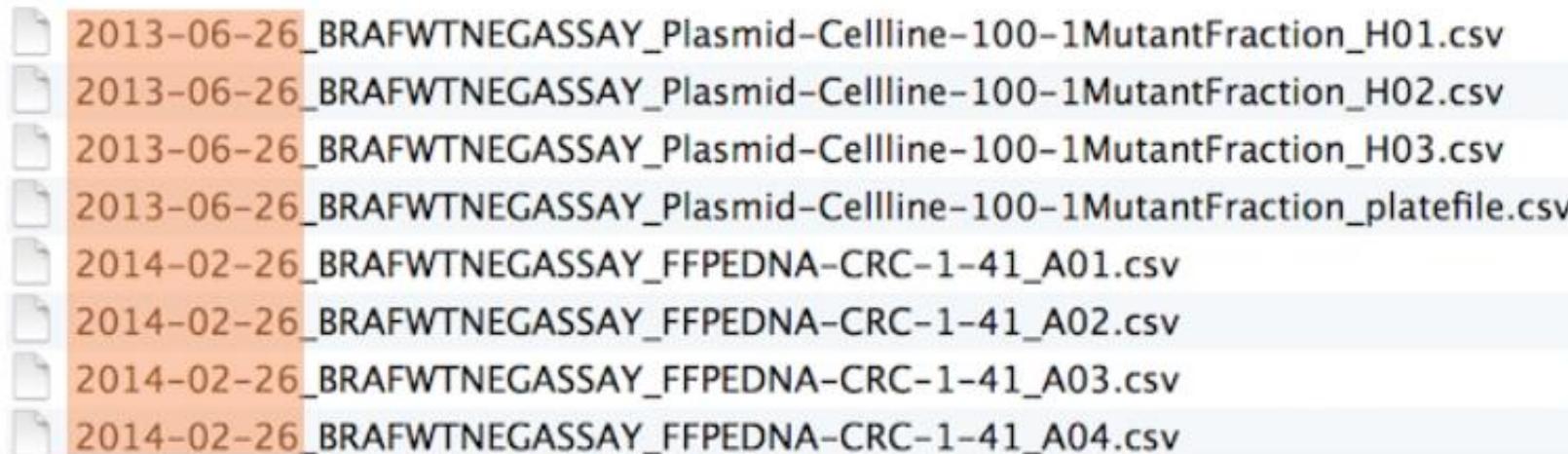
Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

01_marshall-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r

Outline a file naming convention

Chronological order:



A screenshot of a file list from a computer interface. The files are listed vertically with small icons next to each. The first seven files have their names highlighted with a light orange color. The files are:

- 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
- 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
- 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
- 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
- 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
- 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
- 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
- 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv

Logical order:

```
01_marshall-data.r  
02_pre-dea-filtering.r  
03_dea-with-limma-voom.r  
04_explore-dea-results.r  
90_limma-model-term-name-fiasco.r  
helper01_load-counts.r  
helper02_load-exp-des.r  
helper03_load-focus-statinf.r  
helper04_extract-and-tidy.r
```

Joined up thinking

- The R scripts you generate should be human readable
 - Annotate the code
 - Break up the scripts into dedicated tasks
 - Interlink with other within project scripts

```
1 # Data ----
2 # Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
3 # 1. fgf23_data => FGF23 readings from study centres 01-03
4 # 2. food_level_data => Food diary entries
5 # 3. grouped_data => Dialysis and nondialysis diary entries by component
6 # 4. k_data => Serum potassium
7 # 5. master_data_clean => all the clean master file data if required
8 # 6. p_data => Serum phosphate
9 # 7. pth_data => Parathyroid hormone readings
10 # 8. pulses_nuts_data
11
12 source("scripts/01_data_import_and_tidying_master_file.R")
```

Is too much choice good or bad?

Blue Horizon SW 6497	Sky High SW 6504	Snowdrop SW 6511	Ski Slope SW 6518	Rarified Air SW 6525
Byte Blue SW 6498	Atmospheric SW 6505	Balmy SW 6512	Hinting Blue SW 6519	Icelandic SW 6526
Stream SW 6499	Vast Sky SW 6506	Take Five SW 6513	Honest Blue SW 6520	Blissful Blue SW 6527
Open Seas SW 6500	Resolute Blue SW 6507	Respite SW 6514	Notable Hue SW 6521	Cosmos SW 6528
Manitou Blue SW 6501	Secure Blue SW 6508	Leisure Blue SW 6515	Sporty Blue SW 6522	Scanda SW 6529
Loch Blue SW 6502	Georgian Bay SW 6509	Down Pour SW 6516	Denim SW 6523	Revel Blue SW 6530
Bosporus SW 6503	Loyal Blue SW 6510	Regatta SW 6517	Commodore SW 6524	Indigo SW 6531

Inconsistent function names, inconsistent syntax

- R is a very versatile language
 - Sometimes it can be too versatile
 - Do you want to use.....
 - Names or colnames
 - row.names or rownames
 - rowSums or rowsum
 - Sys.time, system.time
- Is it written as.....
 - newobject or new.Object
 - x = 5 or x <- 5
 - mapping=aes(x,y) or mapping = aes(x, y)

Variable selection

```
summary(starwars$name)
```

```
summary(starwars$"name")
```

```
summary(starwars ["name"] )
```

```
summary(starwars [, "name"] )
```

```
summary(starwars[1])
```

```
summary(starwars[, 1])
```

```
summary(starwars[[1]])
```

- Open the script 01_pm_too_much_choice.R

Writing clearer code

- Annotation
- Object names
 - should use only lowercase letters, numbers, and “_”
- Spacing
 - Put a space before and after =
 - Put a space after a ,
 - Operators should be surrounded by spaces e.g. ==, <-, +
- For a more complete list visit
 - <http://style.tidyverse.org/syntax.html>
- Open the script 02_pm_good_habits.R

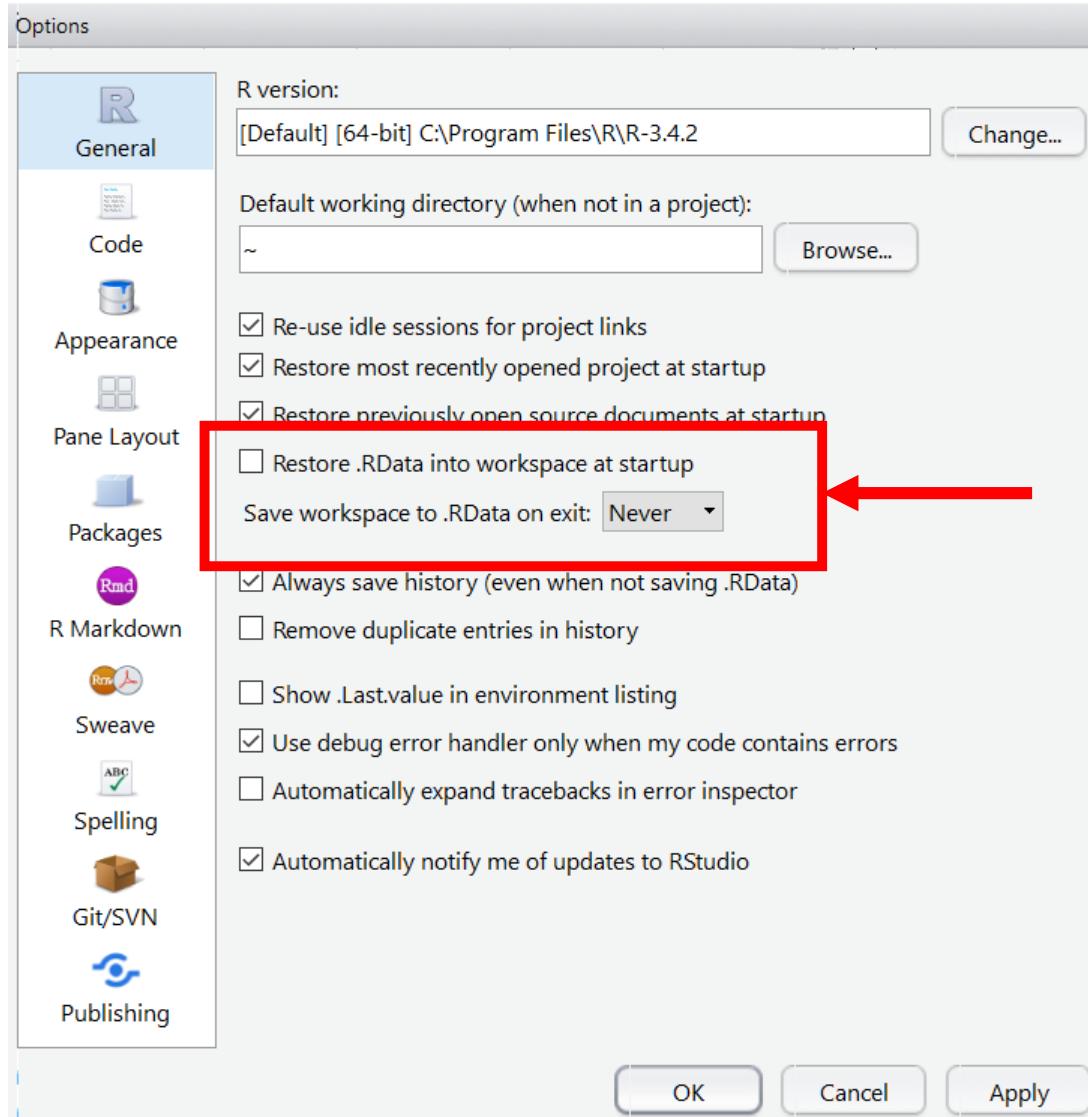
Everything in its right place

- benefits of using R projects for data analysis tasks

T_L_workshop > stats-teaching > project-structure-March-19	
Name	Type
data	File folder
docs	File folder
plots	File folder
scripts	File folder
	Text Document
all_together_now	RMD File
intro_to_RMarkdown	RMD File
LICENSE	File
project-structure-March-19	R Project
README	MD File

- Open the script 03_pm_clean_data.R
- Open the script 04_pm_plots.R
- Open the script 05_pm_analysis.R

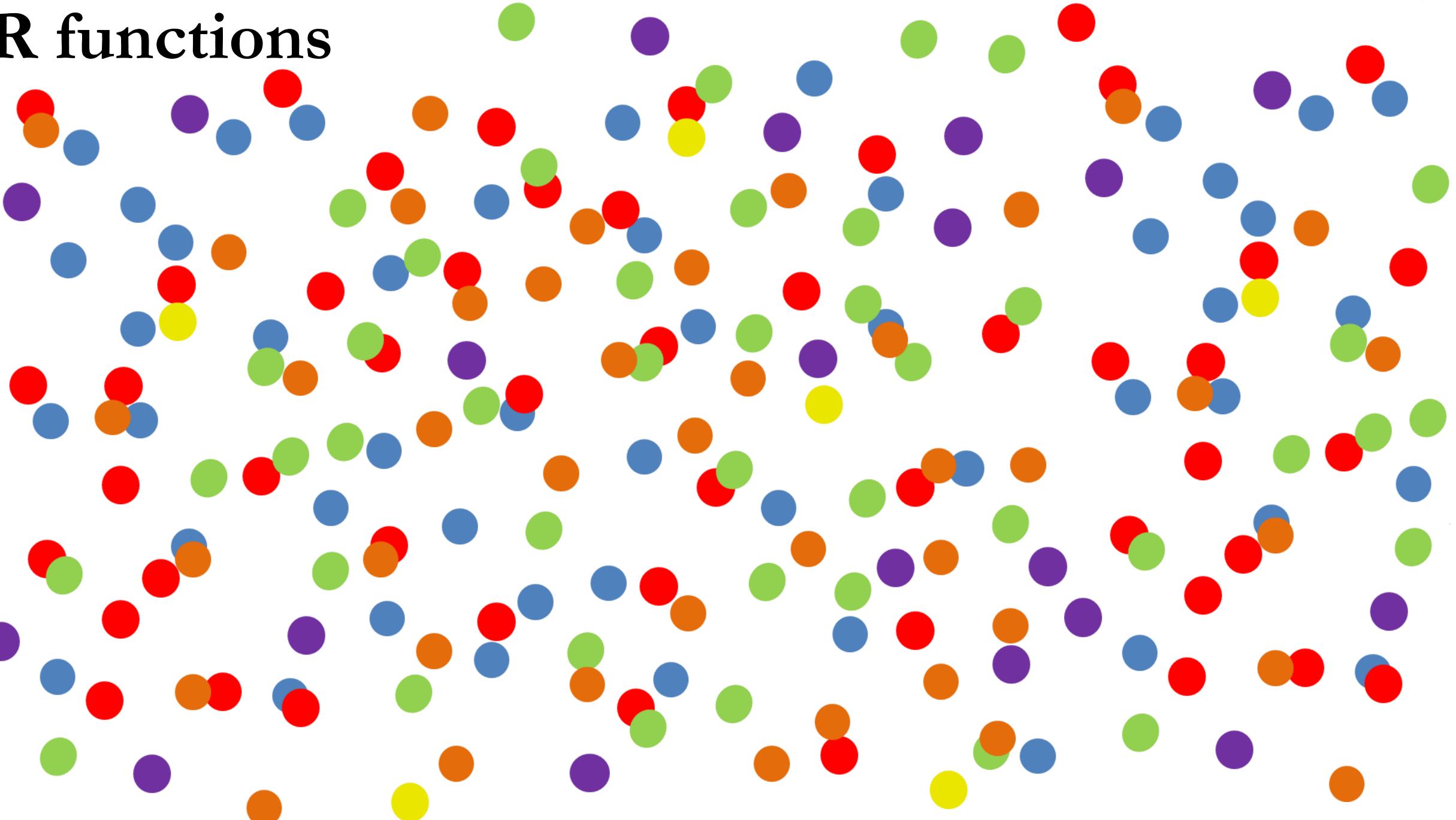
Other points to note



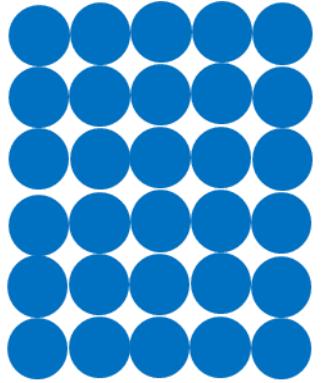
- You might consider your environment as “real”
- If you continue to use R, it is better for you to consider your R scripts as “real”, as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony

Tidyverse

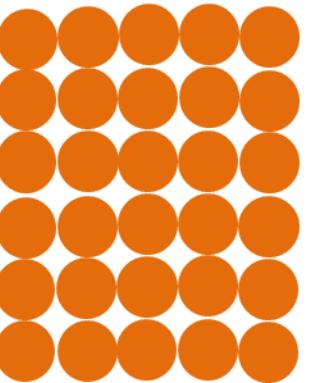
R functions



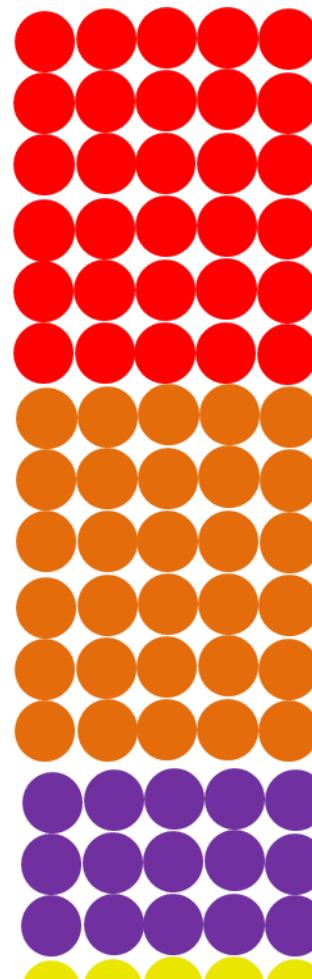
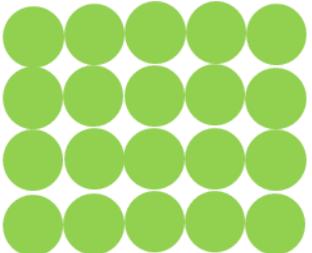
R packages



Base R:
Comes
pre-
loaded



Other packages:
Install once
Update regularly
Load each session



core
tidyverse

What is the tidyverse?

Tidyverse

Packages Articles Learn Help Contribute

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

- Joined up collection of packages for data analysis
 - Consistent functions
 - Uses (tidy) data
 - Supports end-to-end workflows

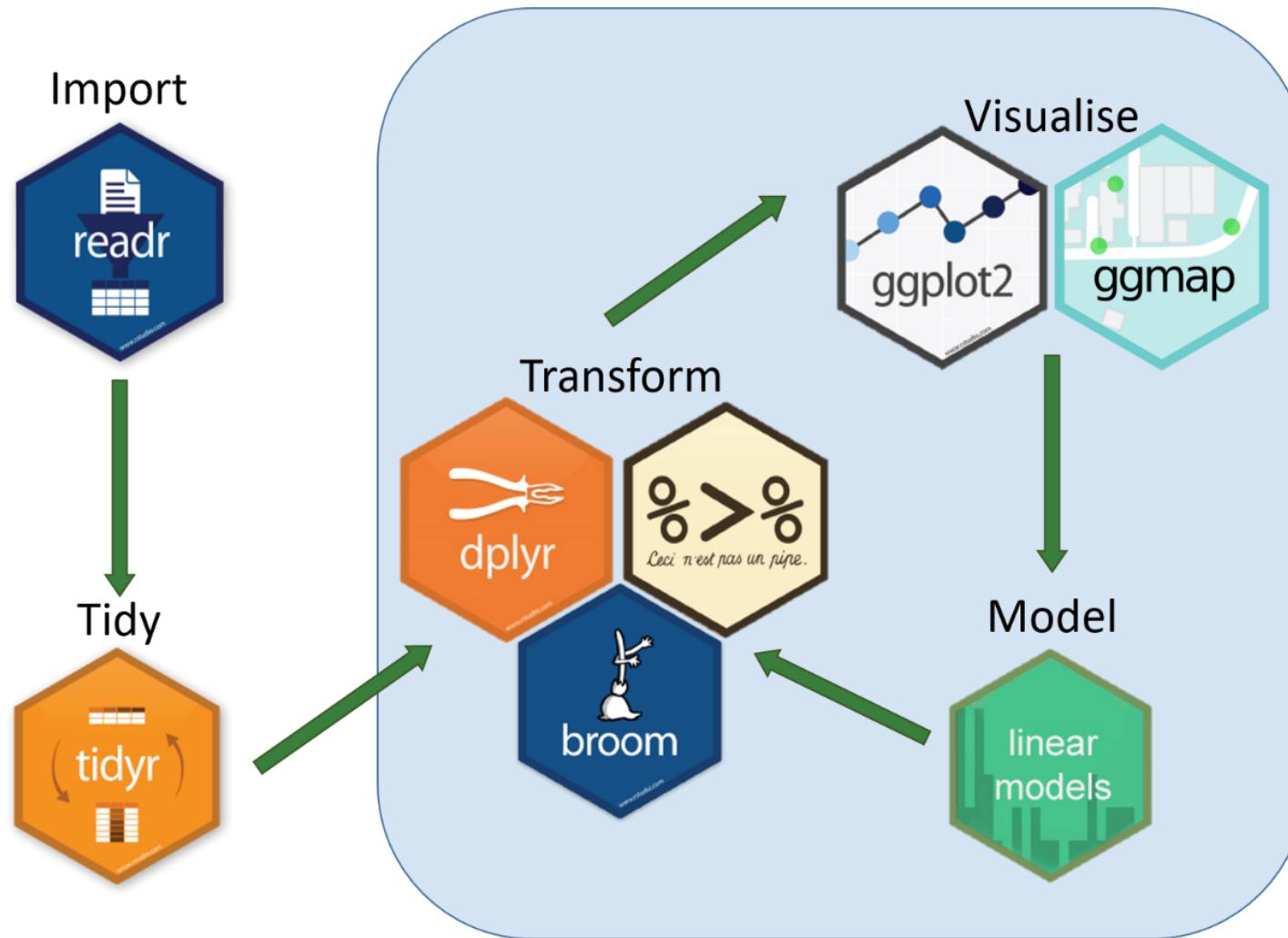
What is the tidyverse?

```
> install.packages(c("broom", "cli2", "crayon",
  "dbplyr", "dplyr", "forcats", "ggplot2", "haven",
  "hms", "httr", "jsonlite", "lubridate",
  "magrittr", "modelr", "pillar", "purrr", "readr",
  "readxl", "reprex", "rlang", "rstudioapi",
  "rvest", "stringr", "tibble", "tidyverse", "xml2"))

> install.packages("tidyverse")
```

Putting the pieces together

- Data analysis in a tidyverse nutshell



```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\|\\|\\|")
17
18 mutated_gene_df <- mutate_at(separated_gene_df,
19                               vars(name:systematic_name),
20                               funs(trimws)
21 )
22
23 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
24 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
25 nearly_there_df <- separate(gathered_gene_df, sample,
26                               c("nutrient", "rate"), sep = 1, convert = TRUE)
27 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
28                       S = "Sulfate", N = "Ammonia", U = "Uracil")
29
30 cleaned_genes_df <- mutate(nearly_there_df,
31                               nutrient = plyr::revalue(nutrient, nutrient_names)
32                               ) %>%
33 filter(!is.na(expression), systematic_name != "")
34
35
36
37
38
15:1 Section 1: Data import, tidying and transformation
  
```

Console Terminal

> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

Parsed with column specification:

```

cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
  
```

See spec(...) for full column specifications.

> |

Line by line

workshop_1_project — 8_weeks_Oct-Dec_17

Environment				
Name	Type	Length	Size	Value
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables

Files	Plots	Packages	Help	Viewer
New Folder	Delete	Rename	More	
Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project				
	Name	Size	Modified	
	..			
	RData	2.5 KB	Oct 2, 2017, 1:49 PM	
	.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM	
	Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM	
	Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM	
	house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM	
	irish_population.csv	315 B	Aug 28, 2017, 4:21 PM	
	raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM	
	workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM	
	ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM	
	ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM	
	ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM	

Line by line

```
ws1_script1_stepwise_Bauer_dataset_an... * x
Source on Save | Run | Source | ...
12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\|\\|\\|")
17
18 mutated_gene_df <- mutate_at(separated_gene_df,
19                               vars(name:systematic_name),
20                               funs(trimws)
21 )
22
23 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
24
25 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
26
27 nearly_there_df <- separate(gathered_gene_df, sample,
28                               c("nutrient", "rate"), sep = 1, convert = TRUE)
29
30 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
31                       S = "Sulfate", N = "Ammonia", U = "Uracil")
32
33 cleaned_genes_df <- mutate(nearly_there_df,
34                               nutrient = plyr::revalue(nutrient, nutrient_names)
35                               ) %>%
36
37 filter(!is.na(expression), systematic_name != "")
38
20:1 Section 1: Data import, tidying and transformation R Script
Console Terminal ✎
~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/project/ ↵
> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|\\|")
```

Global Environment					
	Name	Type	Length	Size	Value
<input type="checkbox"/>	raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
<input type="checkbox"/>	separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

	Name	Size	Modified
	..		
<input type="checkbox"/>	.RData	2.5 KB	Oct 2, 2017, 1:49 PM
<input type="checkbox"/>	.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
<input type="checkbox"/>	Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
<input type="checkbox"/>	Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
<input type="checkbox"/>	house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
<input type="checkbox"/>	irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
<input type="checkbox"/>	raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
<input type="checkbox"/>	workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
<input type="checkbox"/>	ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
<input type="checkbox"/>	ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
<input type="checkbox"/>	ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\\\|")
17
18 mutated_gene_df <- mutate_at(separated_gene_df,
19                               vars(name:systematic_name),
20                               funs(trimws)
21                               )
22
23
24 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
25
26 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
27
28 nearly_there_df <- separate(gathered_gene_df, sample,
29                               c("nutrient", "rate"), sep = 1, convert = TRUE)
30
31 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
32                       S = "Sulfate", N = "Ammonia", U = "Uracil")
33
34 cleaned_genes_df <- mutate(nearly_there_df,
35                               nutrient = plyr::revalue(nutrient, nutrient_names)
36                               ) %>%
37
38 filter(!is.na(expression), systematic_name != "")

```

27:1 Section 1: Data import, tidying and transformation

Console Terminal

```

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/ ↵
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                               vars(name:systematic_name),
+                               funs(trimws)
+                               )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
>

```

Line by line

Environment History Connections

Global Environment

Name	Type	Length	Size	Value
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\|\\|\\|")
17
18 mutated_gene_df <- mutate_at(separated_gene_df,
19                               vars(name:systematic_name),
20                               funs(trimws)
21 )
22
23
24 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
25
26 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
27
28 nearly_there_df <- separate(gathered_gene_df, sample,
29                               c("nutrient", "rate"), sep = 1, convert = TRUE)
30
31 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
32                       S = "Sulfate", N = "Ammonia", U = "Uracil")
33
34 cleaned_genes_df <- mutate(nearly_there_df,
35                               nutrient = plyr::revalue(nutrient, nutrient_names)
36                               ) %>%
37
38   filter(!is.na(expression), systematic_name != "")
29:1 Section 1: Data import, tidying and transformation

```

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```

> .default = col_double(),
> GID = col_character(),
> YORF = col_character(),
> NAME = col_character(),
> GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                               vars(name:systematic_name),
+                               funs(trimws)
+ )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
>

```

Line by line

workshop_1_project — 8_weeks_Oct-Dec_17

Environment History Connections

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\\t")
13
14 separated_gene_df <- separate(raw_gene_df, NAME,
15   c("name", "BP", "MF", "systematic_name",
16     "number"),
17   sep = "\\\\|\\\")
18
19 mutated_gene_df <- mutate_at(separated_gene_df,
20   vars(name:systematic_name),
21   funs(trimws)
22 )
23
24
25 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29 nearly_there_df <- separate(gathered_gene_df, sample,
30   c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33   S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35 cleaned_genes_df <- mutate(nearly_there_df,
36   nutrient = plyr::revalue(nutrient, nutrient_names)
37   ) %>
38   filter(!is.na(expression), systematic_name != "")
32:1 Section 1: Data import, tidying and transformation

```

```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\|\\|\\|")
17 mutated_gene_df <- mutate_at(separated_gene_df,
18                               vars(name:systematic_name),
19                               funs(trimws)
20 )
21
22 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
23
24 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
25
26 nearly_there_df <- separate(gathered_gene_df, sample,
27                               c("nutrient", "rate"), sep = 1, convert = TRUE)
28
29 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
30                       S = "Sulfate", N = "Ammonia", U = "Uracil")
31
32 cleaned_genes_df <- mutate(nearly_there_df,
33                               nutrient = pply::revalue(nutrient, nutrient_names)
34                               ) %>%
35     filter(!is.na(expression), systematic_name != "")
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600

```

Line by line

Environment History Connections

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" ...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

ws1_script1_stepwise_Bauer_dataset_analysis.R

```

15 separated_gene_df <- separate(raw_gene_df, NAME,
16                               c("name", "BP", "MF", "systematic_name",
17                                 "number"),
18                               sep = "\\\\|")
19
20 mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23 )
24
25 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29 nearly_there_df <- separate(gathered_gene_df, sample,
30                               c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                       S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35 cleaned_genes_df <- mutate(nearly_there_df,
36                               nutrient = plyr::revalue(nutrient, nutrient_names)
37                               ) %>%
38 filter(!is.na(expression), systematic_name != "")
39
40
41 < Section 1: Data import, tidying and transformation
42
43
44:1

```

Console Terminal

```

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/ 
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                               vars(name:systematic_name),
+                               funs(trimws)
+ )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                               c("nutrient", "rate"), sep = 1, convert = TRUE)
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                       S = "Sulfate", N = "Ammonia", U = "Uracil")
> cleaned_genes_df <- mutate(nearly_there_df,
+                               nutrient = plyr::revalue(nutrient, nutrient_names)
+                               ) %>%
+ filter(!is.na(expression), systematic_name != "")
>

```

Environment History Connections

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" ...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

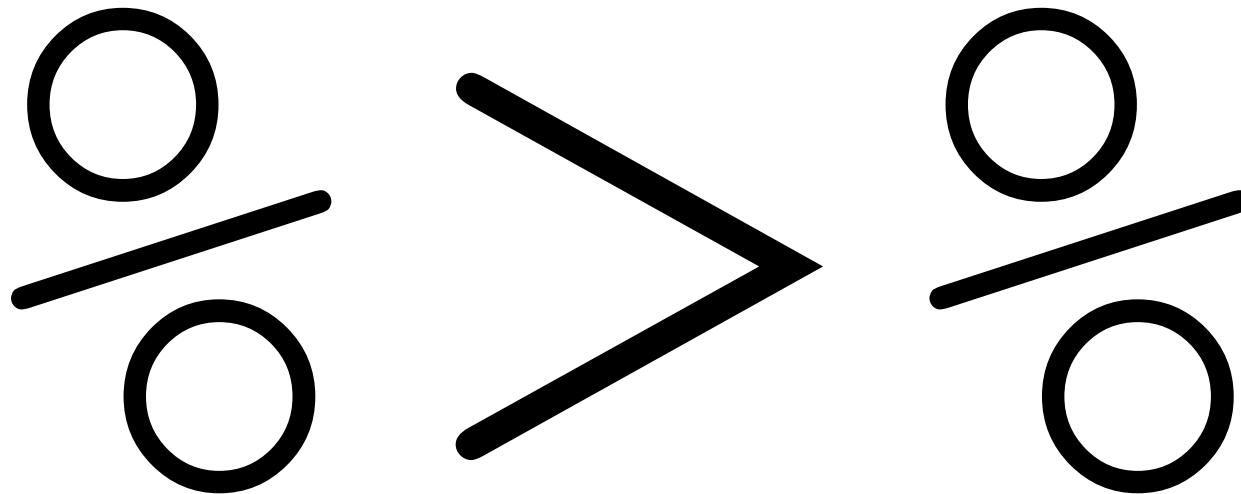
Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Putting the pieces together



Tidyverse code structure

```
new_object <- input_data %>%
```



The input data is outside
the function

```
function( data_to_be_modified, arguments_to_the_function )
```

- | | |
|------------|--------------------------------------|
| new_object | - assign the output to a new object |
| <- | - the assign operator |
| input_data | - data to be manipulated |
| %>% | - the magrittr/pipe operator |
| function | - the function you are calling on |
| data_ | - elements of the input data to use |
| arguments_ | - how you want to apply the function |

```

1 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
2                     S = "Sulfate", N = "Ammonia", U = "Uracil")
3
4 cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t"
5                                 ) %>%
6
7   separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|\\|")
8
9   mutate_at(vars(name:systematic_name), funs(trimws))
10
11 select(-number, -GID, -YORF, -GWEIGHT)
12
13 gather(sample, expression, G0.05:U0.3
14
15 ) %>%
16
17 separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
18
19 ) %>%
20
21 mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
22
23 ) %>%
24
25 filter(!is.na(expression), systematic_name != ""
26
27 )

```

9:18 (Top Level) ▾

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/project/ ↵

```

+   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
+   ) %>%
+
+   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
+   ) %>%
+
+   filter(!is.na(expression), systematic_name != ""
+   )

```

Parsed with column specification:

```

cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)

```

See spec(...) for full column specifications.

> |

Piped

workshop_1_project — 8_weeks_Oct-Dec_17

Environment History Connections

Import Dataset

Global Environment

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "Le...

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

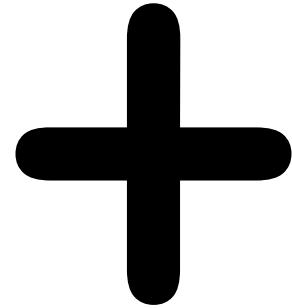
Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Piped

```
1 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
2                      S = "Sulfate", N = "Ammonia", U = "Uracil"
3                      )
4
5 cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t"
6                                  ) %>%
7
8   separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|\\|\\|"
9           ) %>%
10
11  mutate_at(vars(name:systematic_name), funs(trimws))
12
13  select(-number, -GID, -YORF, -GWEIGHT
14         ) %>%
15
16  gather(sample, expression, G0.05:U0.3
17         ) %>%
18
19  separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
20         ) %>%
21
22  mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
23         ) %>%
24
25  filter(!is.na(expression), systematic_name != ""
26         )
```

The moral of the story.....

You can go from this

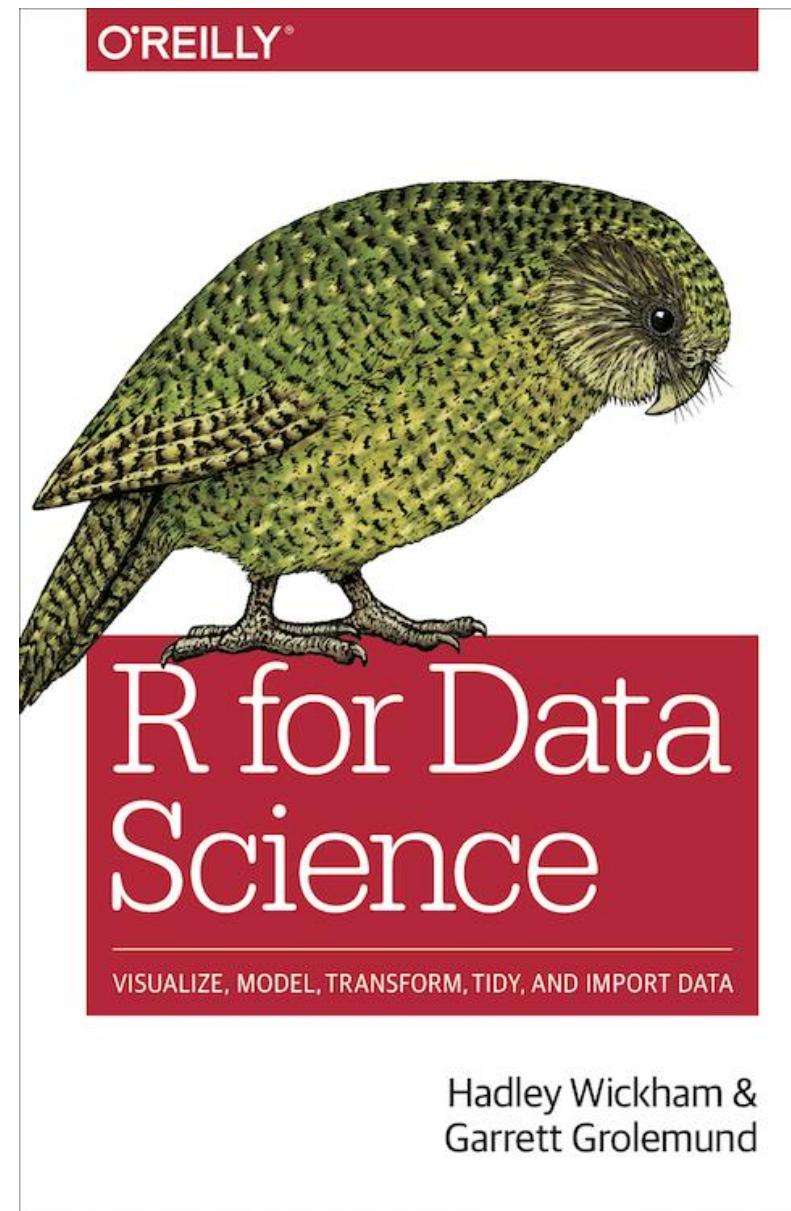


Completely ordinary

To this!!

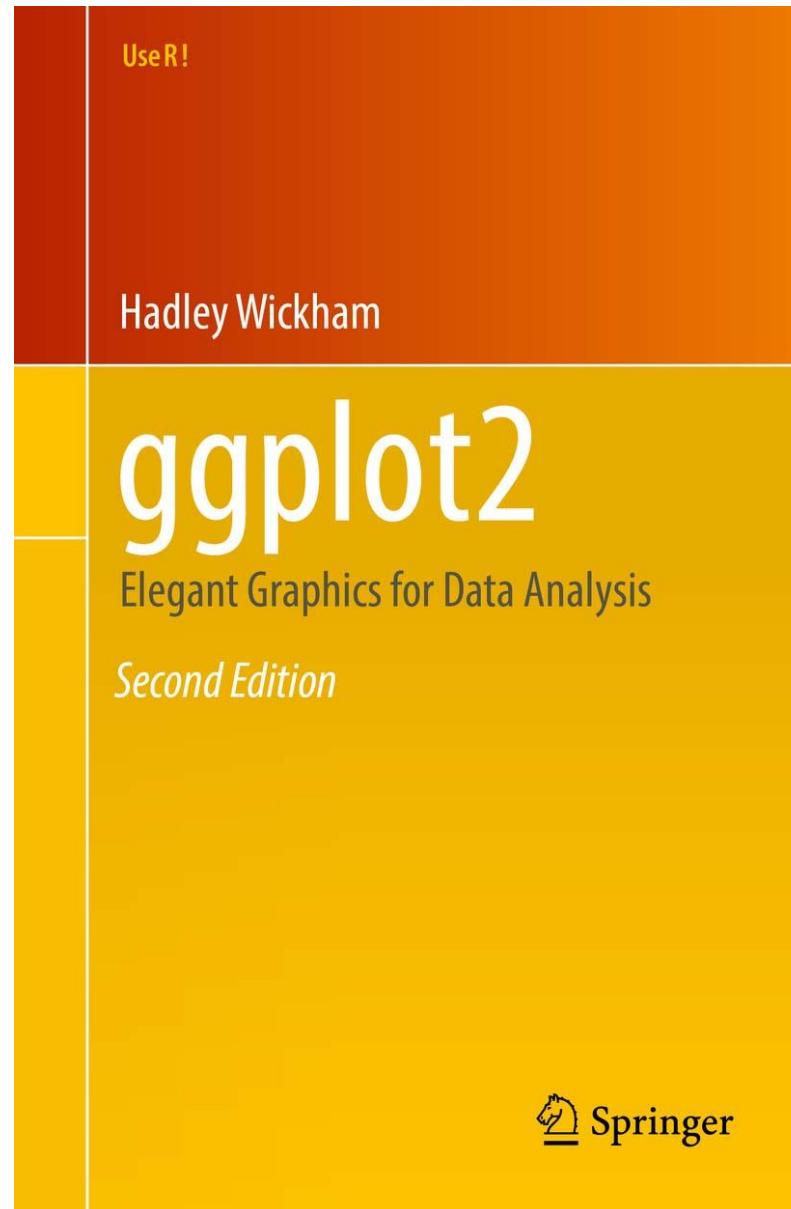
Master Builder!

You could write a book on that!!



[R for Data Science webpage](#)

And on this!!



Explore the tidyverse example scripts

- Open the scripts in the tidyverse folder
- Make sure you are using the correct R-project file

R Markdown

R Markdown

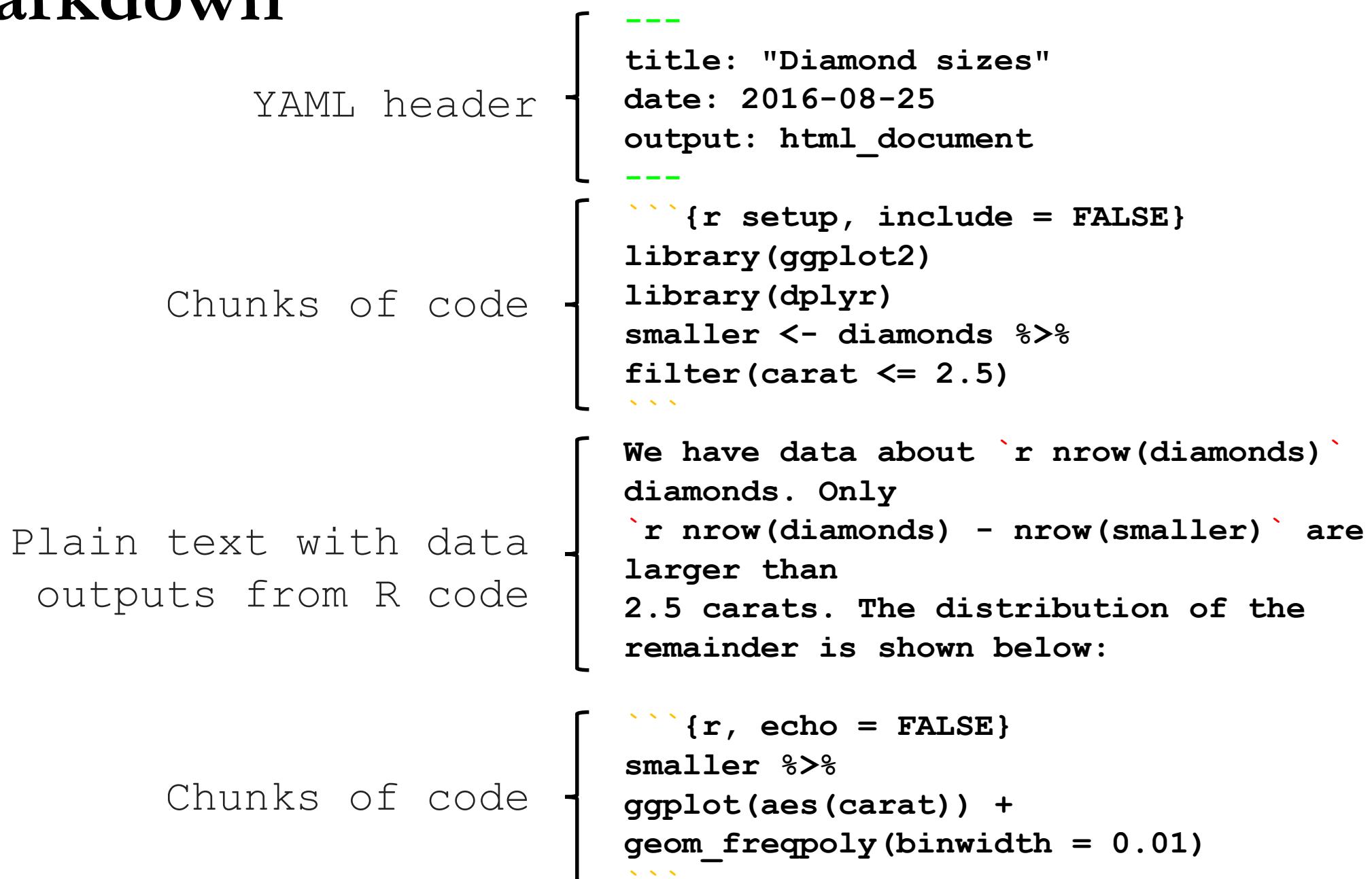
- R Markdown combines the code you wrote, the output produced and your own comments
- You can view it as a digital lab notebook, where you are both recording what you're doing, and what you were thinking while you were doing it!
- R Markdown outputs can take many forms
 - Word documents, PDFs, slideshows etc.
- Once created the .Rmd file gets sent to knitr, which executes the chunks of code and creates a new markdown document
 - this is then processed by pandoc which creates the finished file
 - knitr and pandoc are external websites

What has R Markdown ever done for us?

The image displays a grid of 14 screenshots illustrating the versatility of R Markdown across different output formats. Each screenshot shows a slide or document from the 'Topographic Data in R' package, demonstrating how the same content can look in various styles.

- html**: A standard HTML page with a light blue header and footer.
- ioslides**: A presentation slide with a dark background and a large title.
- reveal.js**: A presentation slide with a dark background and a large title.
- rtf**: A Microsoft Word document with a white background and a small header.
- tufte handout**: A document with a white background and a small header.
- book**: A book cover for 'R for Data Science' by Garrett Grolemund & Hadley Wickham.
- pdf**: A standard PDF document with a white background and a small header.
- dashboard**: A dashboard interface with multiple panels showing data visualizations.
- slidy**: A presentation slide with a white background and a small header.
- markdown**: A plain text file with code and a small image.
- package vignette**: A document with a white background and a small header.
- website**: A screenshot of a website with a light blue header and footer.
- Word**: A Microsoft Word document with a white background and a small header.
- notebook**: An R notebook interface with a white background and a small header.
- beamer**: A presentation slide with a white background and a small header.
- latex**: A plain text file with code and a small image.
- custom template**: A document with a white background and a small header.
- shiny app**: A shiny application interface with a light blue header and footer.

R Markdown



Introduction to R Markdown

- We're now going to look at a R Markdown file that provides some of the tips and tricks you'll need yourselves
 - Code chunks
 - Formatting
 - Tables
 - Figures etc.
- Open the R Markdown file `intro_to_RMarkdown.rmd`

Where to next?

- Understanding basic statistical concepts

www.khanacademy.org

- Collection of YouTube videos describing statistics through R

<http://rafalab.github.io/pages/harvardx.html>

- You know what you want to do, but don't know how to do it

<https://stats.stackexchange.com/>

Structured training.....

Course Languages

- English 557
- Spanish 12
- Chinese 7
(Simplified)

[Show More](#)

Subtitle Languages

- English 586
- Chinese 62
(Simplified)
- Spanish 60

[Show More](#)

All Topics

- Data Science 212
- Business 207
- Computer Science 162

[Show More](#)You searched for **statistics with r**. 586 matchesActive filters: **English**

Courses and Specializations



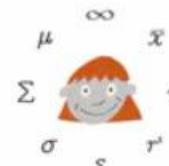
Statistics with R

5-course Specialization · Duke University



The R Programming Environment

Johns Hopkins University



Basic Statistics

University of Amsterdam



Advanced Linear Models for Data Science 2: Statistical Linear Models

Johns Hopkins University

Viewing 42 results matching

Search:

"statistics r" ×[CLEAR ALL](#)

Refine your search



Availability

Current

20

Starting Soon

7

Upcoming

5

Self-Paced

24

Archived

8

Subjects

Biology & Life Sciences

9

Business & Management

3

Computer Science

12

Data Analysis & Statistics

Statistics and R	HarvardX	Course	7/12/2017
Explore Statistics with R	Klx	Course	7/7/2015
Introduction to R for Data Science	Microsoft	Course	10/1/2017
Programming with R for Data Science	Microsoft	Course	10/1/2017
Analyzing Big Data with Microsoft R	Microsoft	Course	10/1/2017
Statistical Analysis in Bioinformatics	USMx	Course	10/23/2017

PAID COURSE

Introduction to the Tidyverse

[Start Course For Free](#)[▶ Play Intro Video](#)

⌚ 4 hours | ▶ 16 Videos | </> 50 Exercises | 🌍 10,553 Participants | 💼 4,150 XP

Course Description

This is an introduction to the programming language R, focused on a powerful set of tools known as the "tidyverse". In the course you'll learn the intertwined processes of data manipulation and visualization through the tools dplyr and ggplot2. You'll learn to manipulate data by filtering, sorting and summarizing a real dataset of historical country data in order to answer exploratory questions. You'll then learn to turn this processed data into informative line plots, bar plots, histograms, and more with the ggplot2 package. This gives a taste both of the value of exploratory data analysis and the power of tidyverse tools. This is a suitable introduction for people who have no previous experience in R and are interested in learning to perform data analysis.

**David Robinson**

Chief Data Scientist, DataCamp



Cork (Ireland) R-Users Group

