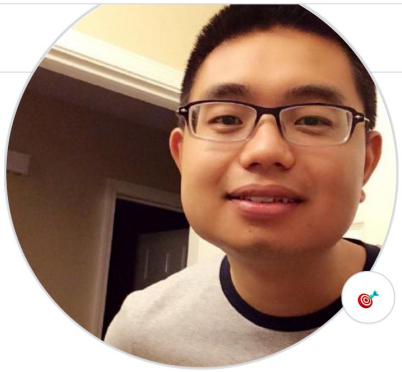
Reproducible research in genomic data science

Ming (Tommy) Tang
Senior Scientist
Dana-Farber Cancer Institute
Twitter: tangming2005

Blog: https://divingintogeneticsandgenomics.rbind.io/

Dana-Farber
Cancer Institute

Who am I ?



Ming Tang

crazyhottommy

Senior scientist at Dana-Farber Cancer Institute working on single-cell RNAseq and single-cell ATAC. Care about reproducible research and open science

Edit profile

1.4k followers · **39** following · ☆ **503**

Dana-Farber Cancer Institute
Boston, MA
tangming2005@gmail.com
http://divingintogeneticsandgenomics.r...

Achievements

Overview | Repositories 133 | Projects | Packages

crazyhottommy / README.md

Hi there 👋

- I am a computational biologist working on (single-cell) genomics, epigenomics and transcriptomics.
- I use R primary for data wrangling and visualization in the tidyverse ecosystem;
- I use python for writing Snakemake workflows and reformatting data;
- I am a unix geek learning shell tricks almost every month; I care about reproducible research and open science.

Learn more about me at my blog

Pinned

Customize your pins

ChIP-seq-analysis Public
ChIP-seq analysis notes from Ming Tang
Python ☆ 516 ⑂ 255

RNA-seq-analysis Public
RNAseq analysis notes from Ming Tang
Python ☆ 595 ⑂ 233

getting-started-with-genomics-tools-and-resources Public
Unix, R and python tools for genomics and data science
Shell ☆ 642 ⑂ 206

pyflow-ChIPseq Public
a snakemake pipeline to process ChIP-seq files from GEO or in-house
Python ☆ 82 ⑂ 36

scRNAseq-analysis-notes Public
scRNAseq analysis notes from Ming Tang
☆ 285 ⑂ 82

bioinformatics-one-liners Public
Bioinformatics one liners from Ming Tang
☆ 329 ⑂ 88

Sign in now to use ZenH

Reproducibility crisis

http://biobungalow.weebly.com/bio-bungalow-blog/everybody-knows-the-scientific-method

Most computational research is not reproducible.

I don’t know of a systematic study, but of papers that I read, approximately 95% fail to include details necessary for replication.

It’s very hard to build off of research like this.

(There’s a lot more to say about repeatability, reproducibility and replicability than I can fit in here…)

An example

- The Importance of Reproducible Research in High-Throughput Biology.
- https://www.youtube.com/watch?v=7gYIs7uYbMo
- By Dr.Keith A. Baggerly from MD Anderson Cancer Center.
- Highly recommend, Keith is very fun.

Flawed Cancer Trial at Duke Sparks Lawsuit

By Jennifer Couzin-Frankel | Sep. 9, 2011 , 3:38 PM

A dozen plaintiffs have filed a **lawsuit** against Duke University and administrators, researchers, and physicians there, alleging that they engaged in fraudulent and negligent behavior when they enrolled cancer patients in a clinical trial compromised by faulty data. The lawsuit, filed Wednesday in a North Carolina court, comes 14 months after a **scandal erupted at Duke** that finally exposed the extent of the trial's problems: in July 2010, Duke oncologist Anil Potti, whose work was central to the trial, admitted that he had embellished his resume and later **resigned**.

Method matters

RESEARCH ARTICLE

Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors

Nathaniel D. Anderson[1,2], Richard de Borja[1,*], Matthew D. Young[3,*], Fabio Fuligni[1,*], Andrej Rosic[1], Nicola D. Roberts[3], Simo…
+ See all authors and affiliations

Science 31 Aug 2018:
Vol. 361, Issue 6405, eaam8419
DOI: 10.1126/science.aam8419

Detection of gene fusions

We detected gene fusions in regions of genomic complexity using an approach that integrates multiple independent fusion algorithms, and then removed those found in normal tissue. Putative fusions were validated by de novo assembly. A total of 1277 normal (nonneoplastic) samples from 43 different tissues were obtained from the NHGRI GTEx consortium (database version 4) and used to remove artifacts. All fusions were visually inspected if one or both genes involved chromoplexy or were adjacent (up to 1 Mbp). Fusions were further filtered by quality of the realigned transcript, breakpoint coverage, and gene expression.

Why reproducibility is hard?

Why reproducibility is hard?

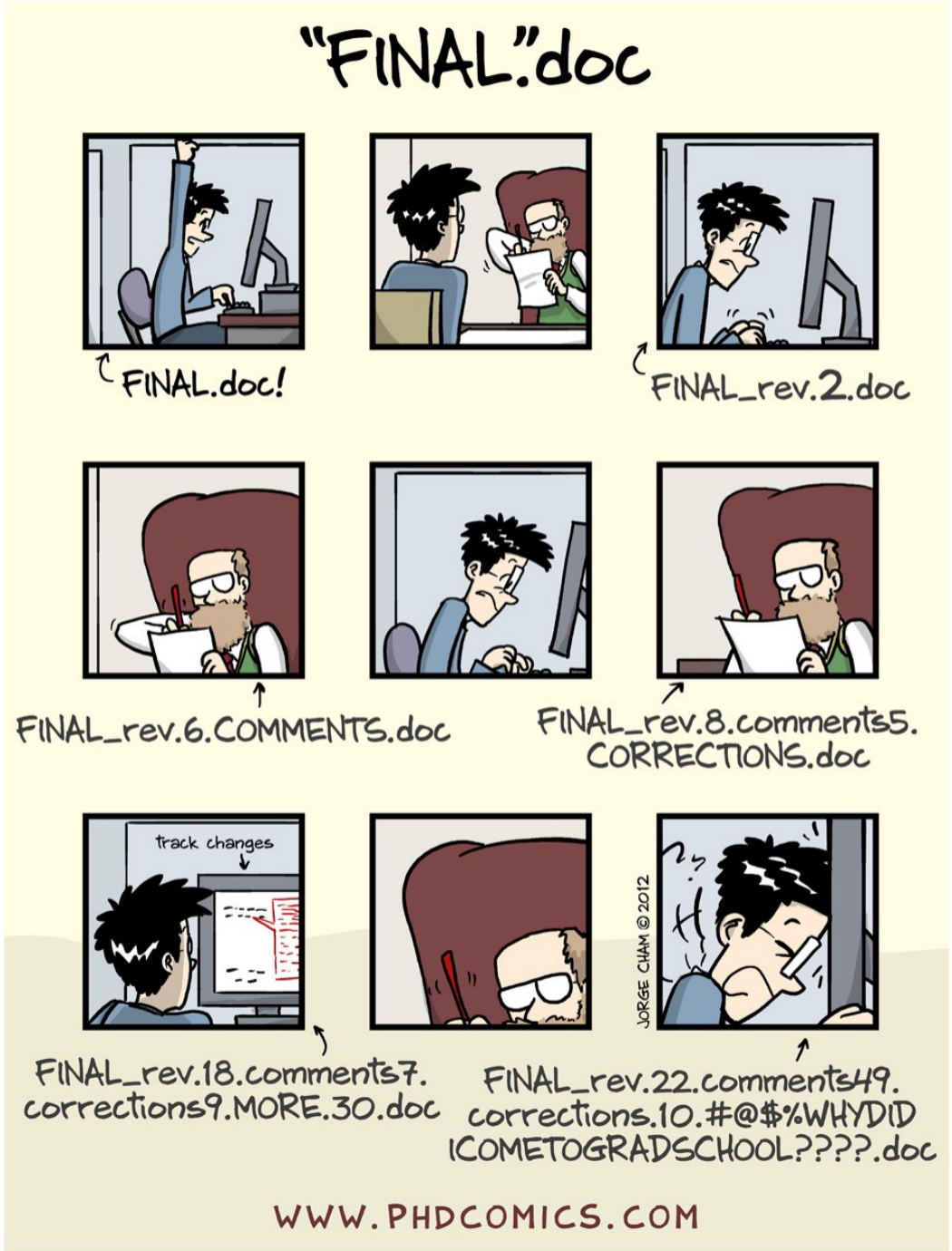
- 1. no raw data are available.
- 2. scripts available upon reasonable request ☺
- 2. lack of method description.
- 3. versions of the tools are different. (e.g. R/python/bioinformatics tools)
- 4. different machines (unix vs windows).

If it is so hard, should you care?

- Keep this in mind: You are going to do the same analysis for sure in the future yourself!
- This is for your own benefit.

How to ensure reproducibility

- Git version control
- Jupyter/R Notebook, documentation
- Containers (docker, singularity, biocontainers https://biocontainers.pro/)


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Version control

- Git
- Github
- Gitlab

Jupyter Notebook

jupyter nbviewer

JUPYTER FAQ

notebook / docs / source / examples / Notebook

Running Code

First and foremost, the Jupyter Notebook is an interactive environment for writing and running code. The notebook is capable of running code in a wide range of languages. However, each notebook is associated with a single kernel. This notebook is associated with the IPython kernel, therefor runs Python code.

Code cells allow you to enter and run code

Run a code cell using `Shift-Enter` or pressing the button in the toolbar above:

In [2]:
```
a = 10
```

In [3]:
```
print(a)
```

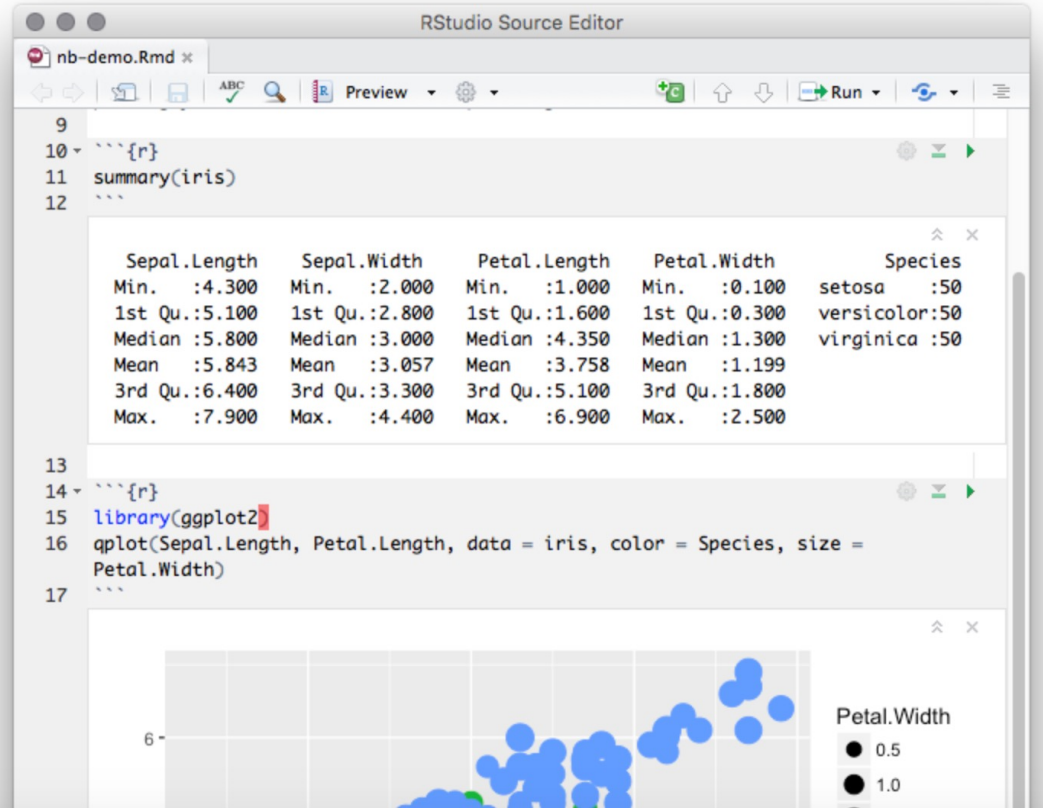
```
10
```

There are two other keyboard shortcuts for running code:

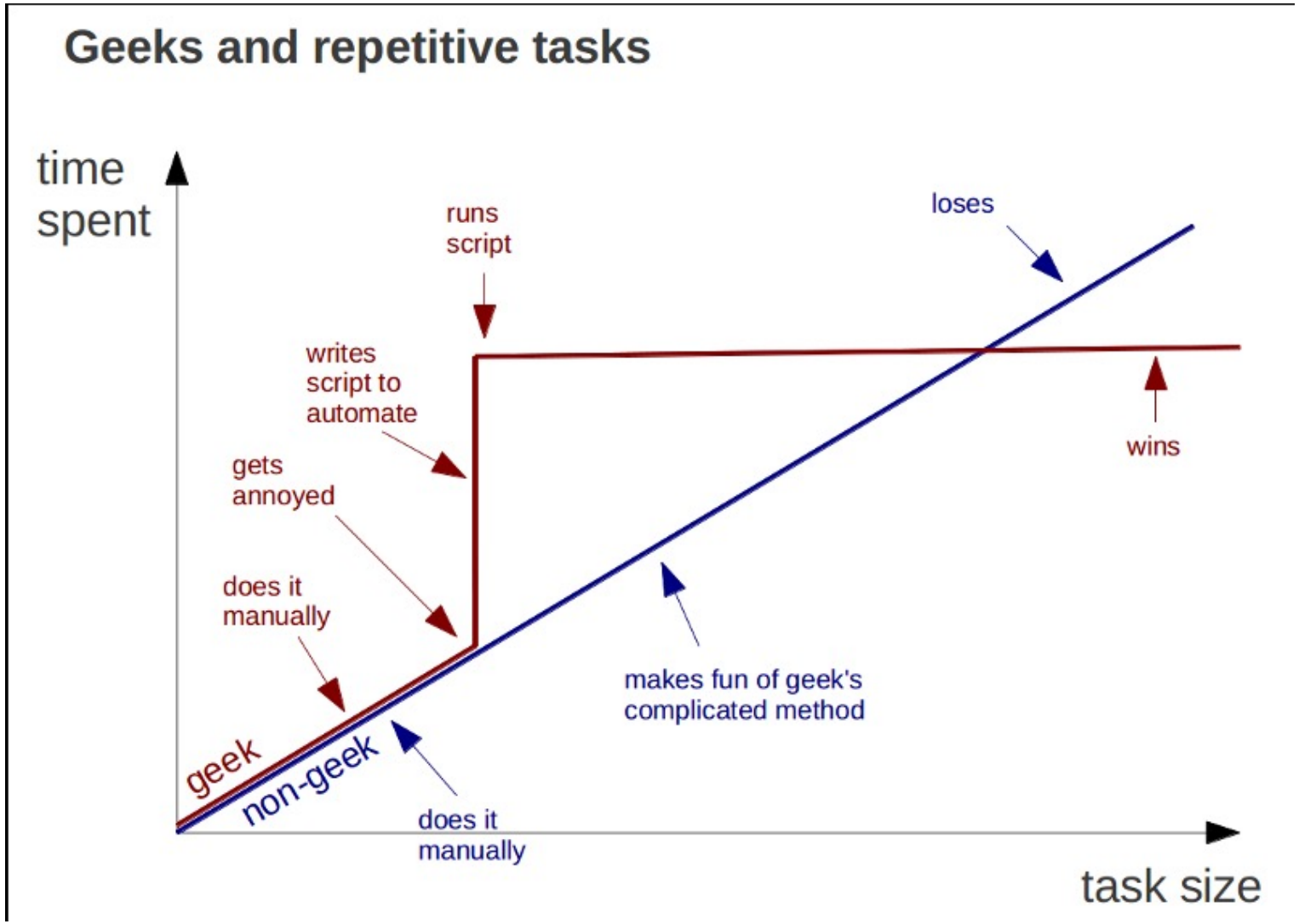
- `Alt-Enter` runs the current cell and inserts a new one below.
- `Ctrl-Enter` run the current cell and enters command mode.

R notebook/markdown

An R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input.



Automation makes your research more reproducible AND saves you time in the long run



Computers are good at repetitive work

Good Side effect of automation

- The best documentation is automation
- Write scripts for everything unless it is not possible. (manual editing, document, document, document!)
- Markdown, MKdocs https://www.mkdocs.org/

Credit to someone in the twitter-verse ☺

Tips for automation

- 1. if you have a repetitive simple task, put them in to a shell script: my_routine.sh.
- 2. good old GNU make
- 3. more recent snakemake, nextflow, WDL etc.

Awesome Pipeline

A curated list of awesome pipeline toolkits inspired by Awesome Sysadmin

Pipeline frameworks & libraries

- ActionChain - A workflow system for simple linear success/failure workflows.
- Adage - Small package to describe workflows that are not completely known at definition time.
- Airflow - Python-based workflow system created by AirBnb.
- Anduril - Component-based workflow framework for scientific data analysis.
- Antha - High-level language for biology.
- AWE - Workflow and resource management system with CWL support
- Bds - Scripting language for data pipelines.
- BioMake - GNU-Make-like utility for managing builds and complex workflows.
- BioQueue - Explicit framework with web monitoring and resource estimation.
- Bioshake - Haskell DSL built on shake with strong typing and EDAM support
- Bistro - Library to build and execute typed scientific workflows.

https://github.com/pditommaso/awesome-pipeline

519

Snakemake—a scalable bioinformatics workflow engine

Publication Article in **Bioinformatics**, published October 2012

Authors Johannes Köster, Sven Rahmann

More details

{wdl}

nextflow

docker

- Why docker?
- Imagine you are working on an analysis in R and you send your code to a friend. Your friend runs exactly this code on exactly the same data set but gets a slightly different result. This can have various reasons such as a different operating system, a different version of an R package, etc. Docker is trying to solve problems like that.
- Think it as a virtual machine!
- This just happened between me and my colleagues who used a different version of R packages!

https://cyverse-cybercarpentry-container-workshop-2018.readthedocs-hosted.com/en/latest/docker/dockerintro.html
https://ropenscilabs.github.io/r-docker-tutorial/01-what-and-why.html

conda and biocoda

Conda

CONDA

Package, dependency and environment management for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN

MENU

nature methods

Correspondence | Published: 02 July 2018

Bioconda: sustainable and comprehensive software distribution for the life sciences

Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris & Johannes Köster The Bioconda Team

Nature Methods **15**, 475–476 (2018) | Download Citation

Other important untaught skills

- Naming files
- Project organization
- Data organization, backup plans

What are your file names look like?

NO

myabstract.docx
Joe's Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt

http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf

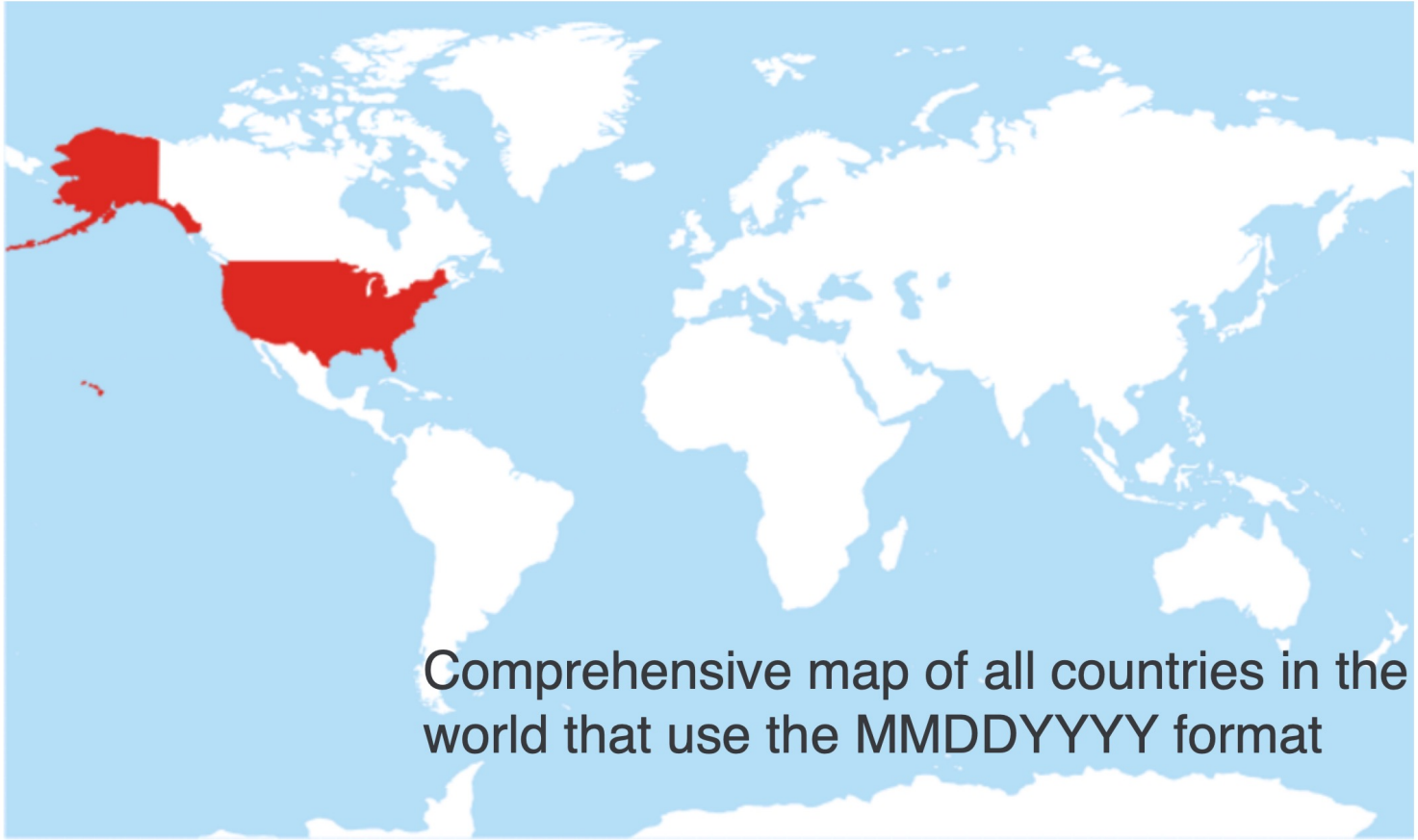
Three principles for (file) names

- 1. Machine readable (do not put special characters and space in the name)
- 2. Human readable (Easy to figure out what the heck something is, based on its name, add slug)
- 3. Plays well with default ordering:
- * Put something numeric first
- * Use the ISO 8601 standard for dates (YYYY-MM-DD)
- * Left pad other numbers with zeros

http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf

Jenny Bryan

Use the YYYY-MM-DD format for date



http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf

Punctuation

Deliberate use of `"-"` and `"_"` allows recovery of meta-data from the filenames:

- `"_"` underscore used to delimit units of meta-data I want later
- `"-"` hyphen used to delimit words so my eyes don't bleed

```
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
```

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
     [,1]         [,2]             [,3]                                  [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

date assay sample set well

This happens to be R but also possible in the shell, Python, etc.

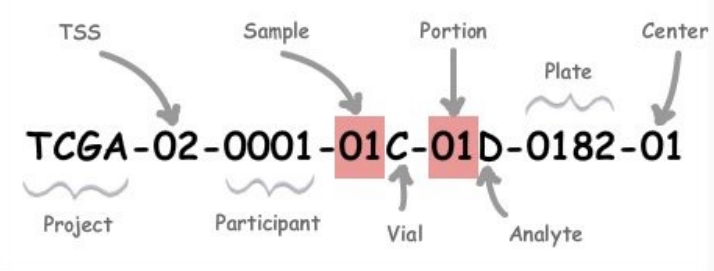
Go forth and use awesome file names :)

```
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
```

```
01_marshal-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
```

Jenny Bryan:
https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html

TCGA barcode

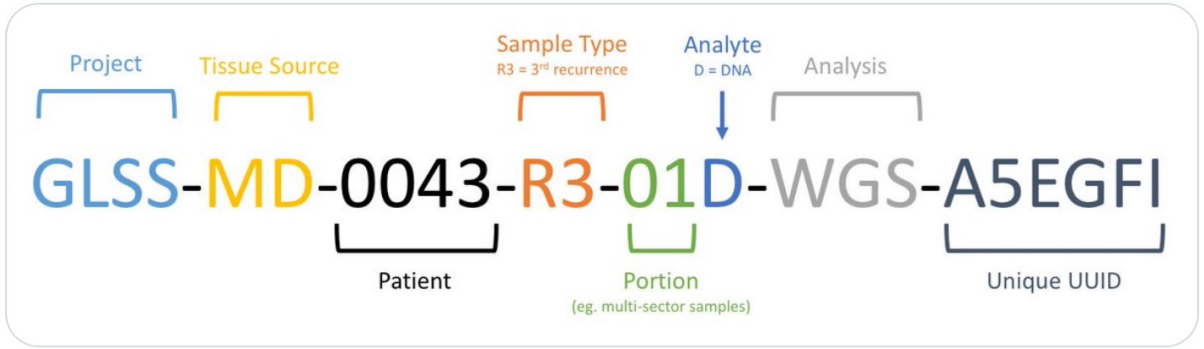


Label	Identifier for	Value	Value Description	Possible Values
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	182	The 182nd plate	4-digit alphanumeric value
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	1	The first portion of the sample	01-99
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Project	Project name	TCGA	TCGA project	TCGA
Sample	Sample type	1	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Center	Sequencing or characterization center that will receive the aliquot for analysis	1	The Broad Institute GCC	See Code Tables Report
Participant	Study participant	1	The first participant from MD Anderson for GBM study	Any alpha-numeric value
TSS	Tissue source site	2	GBM (brain tumor) sample from MD Anderson	See Code Tables Report

Good idea to encode metadata to filenames?

Ming (Tommy) Tang
@tangming2005

nice work! Also, a nice processing pipeline github.com/fpbarthel/GLAS... A general question for tweeps: is coding metadata in the file name best practice? I really love this strategy (similar to TCGA barcode). one has to think really hard designing sample ids.



Jeremy Leipzig @jermdemo · May 27
Replying to @tangming2005
Putting metadata in a filename is bad practice in the same sense as leaving your sleeping toddler in the car while you run to the ATM. What else are you going to do?
1 2

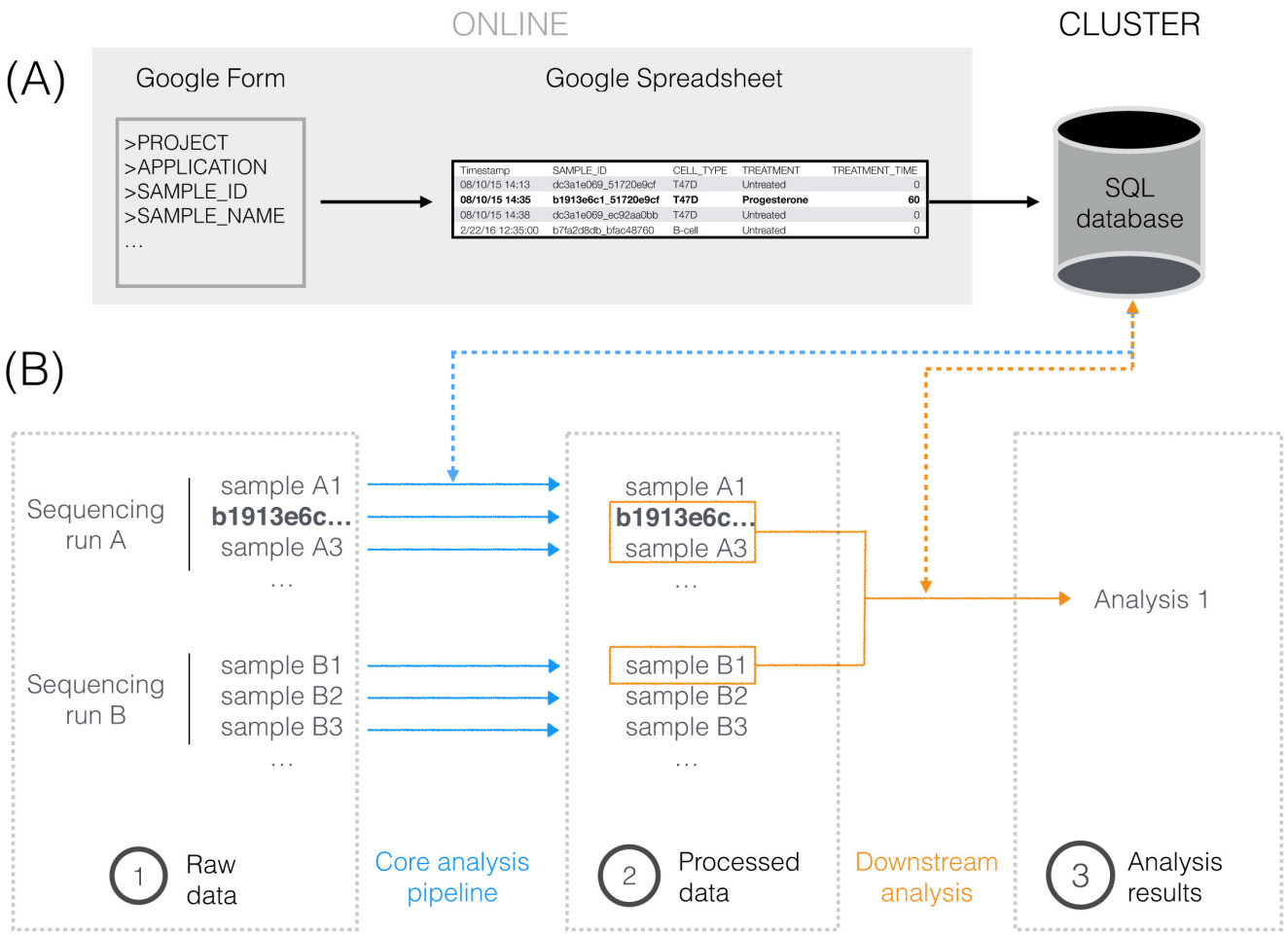
Ming (Tommy) Tang @tangming2005 · May 27
want to hear more on why? I know it might be bad to leak private information if code the metadata in the filename. on the other hand, working with a filename of uuid.txt is not fun (I know it is designed for machine not human).
2

Jeremy Leipzig @jermdemo · May 27
If the metadata is wrong you need to change the filename and change it everywhere it might have been referenced. Also some pipeline frameworks don't respect filenames to the same extent you might.
2 1

Jeff Gentry @geoffjentry · May 27
💯- seems like a good idea until it's not and then you're hosed
1

https://twitter.com/tangming2005/status/1398094371594149893

Make large sequencing project successful



Listen

Article

Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

Pages 2-10 | Received 01 Jun 2017, Accepted author version posted online: 29 Sep 2017, Published online:24 Apr 2018

Download citation https://doi.org/10.1080/00031305.2017.1375989 Check for updates

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

Common mistakes

- https://datacarpentry.org/spreadsheet-ecology-lesson/02-common-mistakes/

  "There are a few potential errors to be on the lookout for in your own data as well as data from collaborators or the Internet. If you are aware of the errors and the possible negative effect on downstream data analysis and result interpretation, it might motivate yourself and your project members to try and avoid them. **Making small changes to the way you format your data in spreadsheets can have a great impact on efficiency and reliability when it comes to data cleaning and analysis**"

No multiple tables in the same sheet

Using multiple tables

A common strategy is creating multiple data tables within one spreadsheet. This confuses the computer, so don't do this! When you create multiple tables within one spreadsheet, you're drawing false associations between things for the computer, which sees each row as an observation. You're also potentially using the same field name in multiple places, which will make it harder to clean your data up into a usable form. The example below depicts the problem:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1																																
2	lake site May 29 2012						29-May		lake site Jun 12. 2012						12-Jun		lake site Jun 19. 2012						19-Jun		Lake site Jun 26. 2012						26-Jun	
3			Bug1	bug2			avr	SEM		plot	bug1	bug2			avr	SEM		plot	bug1	bug2	general					plot	bug1	bug2	general			
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	1	0	1				4	T1	0	6	6				4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	0	3	3				5	T1	3	20	23				5	T1	10	70	80				5	T1	4	96	100			
9	6	T2	1	0	1				6	T2	0	0	0				6	T2	1	7	8				6	T2	0	1	1			
10	7	T2	0	0	0				7	T2	0	0	0				7	T2	0	1	1				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	1	0	1				8	T2	0	0	0				8	T2	0	0	0			
12	9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0			
14	11	contro	0	0	0				11	contro	0	0	0				11	control	0	0	0				11	control	0	0	0			
15	12	contro	0	0	0				12	contro	0	0	0				12	control	0	0	0				12	control	0	0	0			
16	13	contro	0	0	0				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
17	14	contro	0	0	0				14	contro	0	0	0				14	control	0	1	1				14	control	0	0	0			
18	15	contro	1	0	1				15	contro	3	0	3				15	control	0	1	1				15	control	0	0	0			
19																																
20																																
21	Barn site May 29. 2012								Barn site Jun 12. 2012								Barn site Jun 19 . 2012								Barn Site Jun 26 . 2012							
22		plot	bug1	bug2	general		29-May			plot	bug1	bug2	general		12-Jun			plot	bug1	bug2	general		19-Jun			plot	bug1	bug2	general		26-Jun	
23	1	T1	3	3	6				1	T1	21	0	21				1	T1	5	0	5				1	T1	0	0	0		avr	SEM
24	2	T1	1	4	5		avr	SEM	2	T1	36	74	110		avr	SEM	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10			
28	6	T2	0	0	0				6	T2	1	0	1				6	T2	0	8	8				6	T2	0	0	0			
29	7	T2	0	0	0				7	T2	0	4	4				7	T2	0	12	12				7	T2	0	0	0			
30	8	T2	0	1	1				8	T2	0	0	0				8	T2	0	0	0				8	T2	0	0	0			
31	9	T2	0	1	1				9	T2	0	0	0				9	T2	3	0	3				9	T2	0	0	0			
32	10	T2	0	0	0				10	T2	0	0	0				10	T2	2	0	2				10	T2	0	2	2			
33	11	contro	0	0	0				11	contro	1	0	1				11	control	0	5	5				11	control	0	2	2			
34	12	contro	0	1	1				12	contro	0	0	0				12	control	1	1	2				12	control	1	0	1			
35	13	contro	0	1	1				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
36	14	contro	0	1	1				14	contro	8	1	9				14	control	0	5	5				14	control	0	3	3			
37	15	contro	0	2	2				15	contro	0	1	1				15	control	0	2	2				15	control	1	0	0			
38																																

Using problematic null values

Example: using -999 or other numerical values (or zero) to represent missing data.

Solutions:

There are a few reasons why null values get represented differently within a dataset. Sometimes confusing null values are automatically recorded from the measuring device. If that's the case, there's not much you can do, but it can be addressed in data cleaning with a tool like OpenRefine before analysis. Other times different null values are used to convey different reasons why the data isn't there. This is important information to capture, but is in effect using one column to capture two pieces of information. Like for using formatting to convey information it would be good here to create a new column like 'data_missing' and use that column to capture the different reasons.

Whatever the reason, it's a problem if unknown or missing data is recorded as -999, 999, or 0. Many statistical programs will not recognize that these are intended to represent missing (null) values. How these values are interpreted will depend on the software you use to analyze your data. It is essential to use a clearly defined and consistent null indicator. Blanks (most applications) and NA (for R) are good choices. White et al, 2013, explain good choices for indicating null values for different software applications in their article: Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution.

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Using formatting to convey information

Example: highlighting cells, rows or columns that should be excluded from an analysis, leaving blank rows to indicate separations in data.

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Solution: create a new field to encode which data should be excluded.

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y

Using problematic field names

Choose descriptive field names, but be careful not to include spaces, numbers, or special characters of any kind. Spaces can be misinterpreted by parsers that use whitespace as delimiters and some programs don't like field names that are text strings that start with numbers.

Underscores (`_`) are a good alternative to spaces. Consider writing names in camel case (like this: ExampleFileName) to improve readability. Remember that abbreviations that make sense at the moment may not be so obvious in 6 months, but don't overdo it with names that are excessively long. Including the units in the field names avoids confusion and enables others to readily interpret your fields.

Examples

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Be cautious with excel

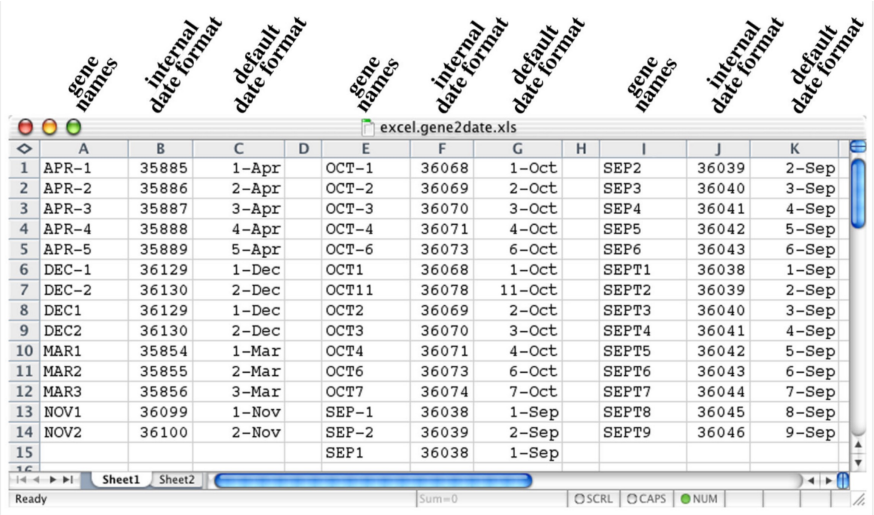
Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

115k Accesses | **38** Citations | **2375** Altmetric | Metrics



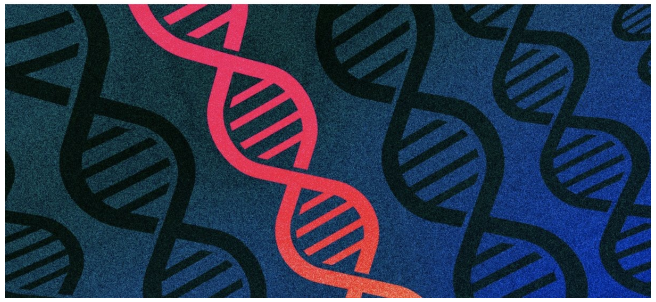
	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT



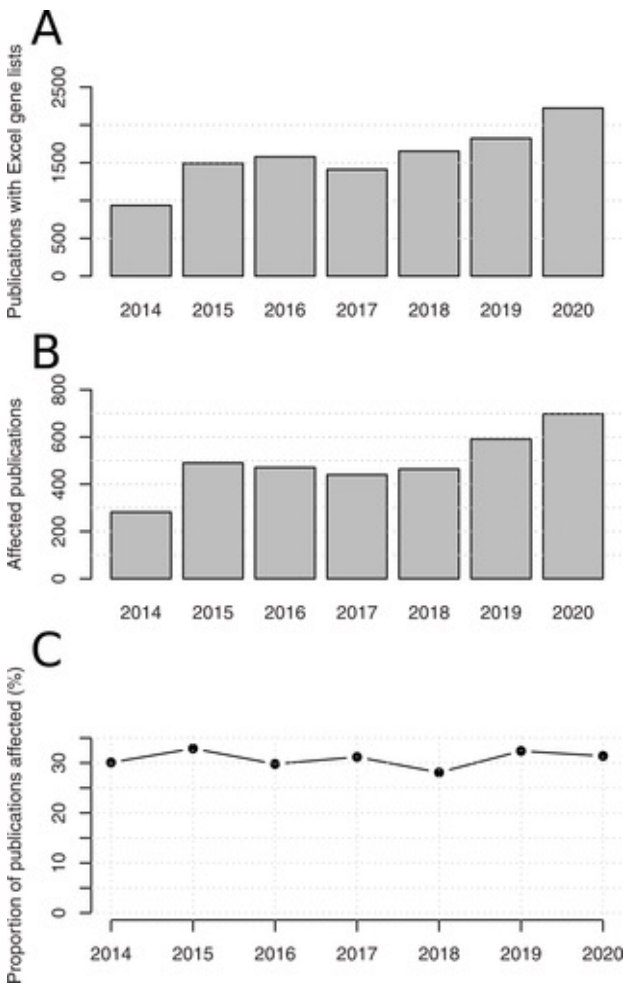
https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Gene name errors: Lessons not learned

Retraction Watch
@RetractionWatch

An Excel screw-up leads to a retraction. "This technological issue caused rows to shift and the data from the different groups got mixed up."
sciencedirect.com/science/articl...

3:27 PM · Aug 6, 2018 · Twitter Web Client



https://www.sciencedirect.com/science/article/pii/S0018506X18302599?via%3Dihub

https://github.com/jennybc/scary-excel-stories by Jenny Bryan

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008984

Why not excel?

BBC | Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | US Election | Coronavirus | Video | World | US & Canada | UK | Business | Tech | Science | Stories

Tech

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October

Coronavirus pandemic

The problem is that PHE's own developers picked an old file format to do this - known as XLS.

As a consequence, each template could handle only about 65,000 rows of data rather than the one million-plus rows that Excel is actually capable of.

And since each test result created several rows of data, in practice it meant that each template was limited to about 1,400 cases.

When that total was reached, further cases were simply left off.

https://www.bbc.com/news/technology-54423988

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

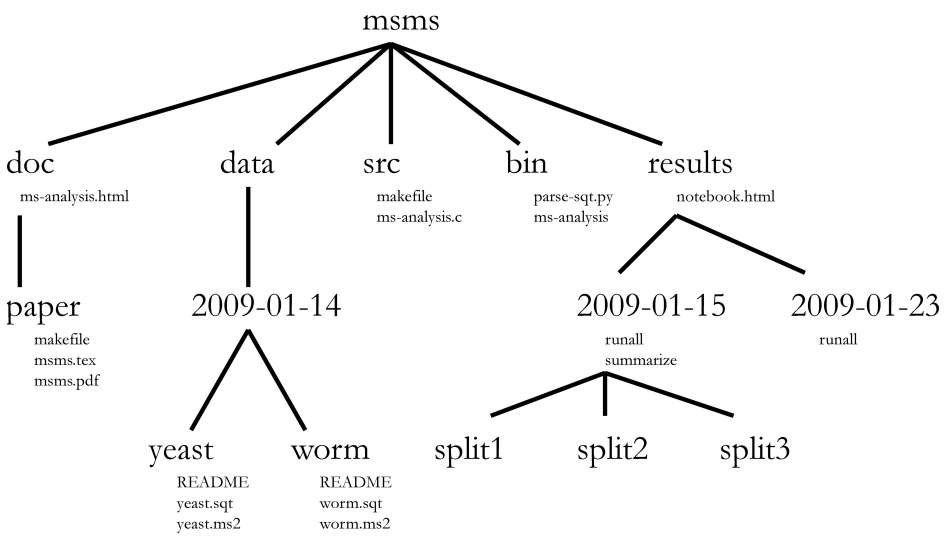
OPEN ACCESS

EDUCATION

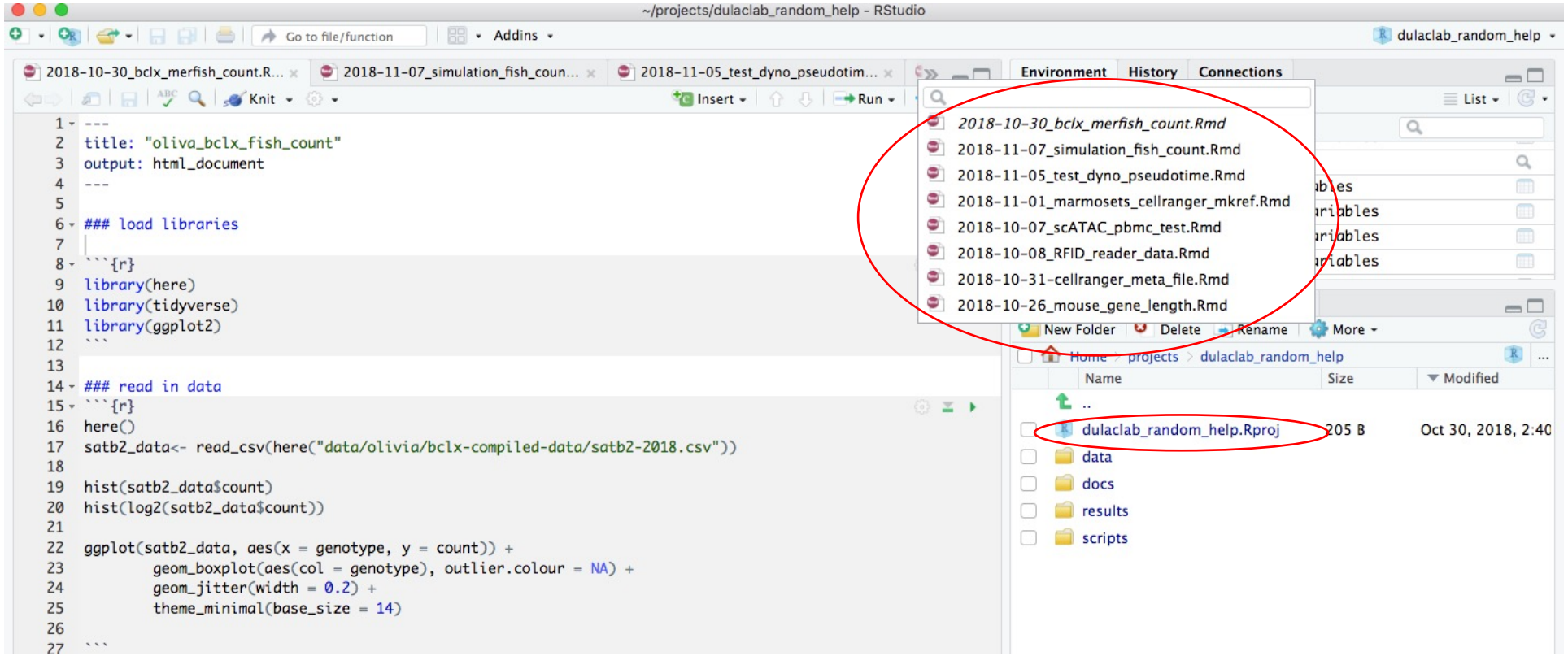
A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

Published: July 31, 2009 • https://doi.org/10.1371/journal.pcbi.1000424



Rstudio R project



Also check workflowr: https://github.com/jdblischak/workflowr

An example from me: https://crazyhottommy.github.io/scRNA-seq-workshop-Fall-2019/index.html

Always use here() to construct relative path.

To continue our example, start R in the `foofy` directory, wherever that may be. Now the code looks like so:

```
library(ggplot2)
library(here)

df <- read.delim(here("data", "raw_foofy_data.csv"))
p <- ggplot(df, aes(x, y)) + geom_point()
ggsave(here("figs", "foofy_scatterplot.png"))
```

Tidyverse Packages

2017/12/12

Jenny Bryan

I was honored to speak this week at the IASC-ARS/NZSA Conference, hosted by the Stats Department at The University of Auckland. One of the conference themes is to celebrate the accomplishments of Ross Ihaka, who got R started back in 1992, along with Robert Gentleman. My talk included advice on setting up your R life to maximize effectiveness and reduce frustration.

Two specific slides generated much discussion and consternation in #rstats Twitter:

> *If the first line of your R script is*
>
> `setwd("C:\Users\jenny\path\that\only\I\have")`
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.*

> *If the first line of your R script is*
>
> `rm(list = ls())`
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.*

https://www.tidyverse.org/blog/2017/12/workflow-vs-script/

Remember, always keep the data in the data folder untouched, I usually do
$ chmod u-w –R data/
To revoke the user's write right so you can not edit or delete the files in the data folder.

Always generate the output/intermediate files/figures in the results folder using the scripts in the scripts folder

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

OPEN ACCESS

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

Browse | Publish | About

OPEN ACCESS

COMMUNITY PAGE

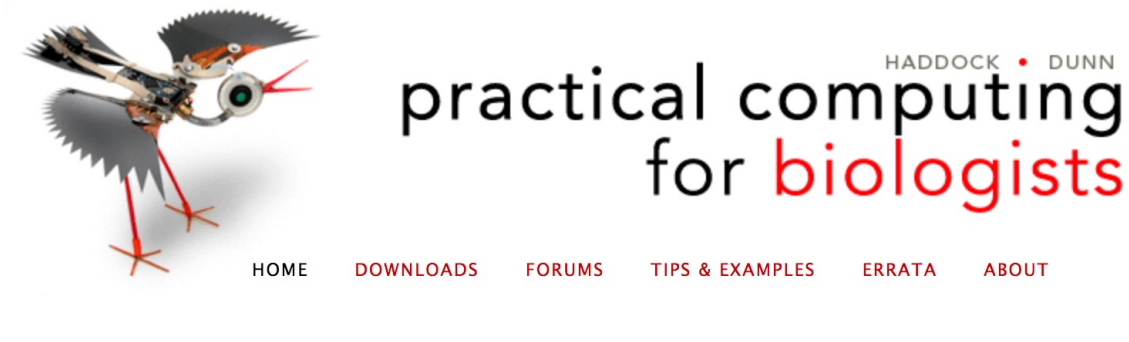
Best Practices for Scientific Computing

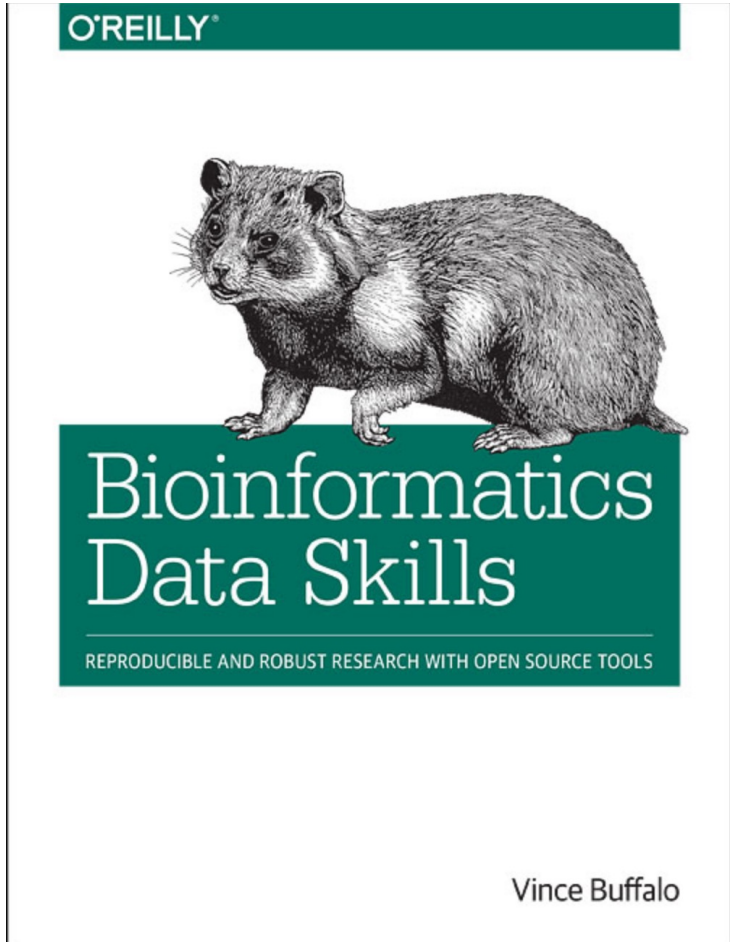
Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

More readings

- **What They Forgot to Teach You About R** https://rstats.wtf/
- The renv package is a new effort to bring project-local R dependency management to your projects. https://rstudio.github.io/renv/articles/renv.html
- A Reproducible Data Analysis Workflow with R Markdown, Git, Make, and Docker: https://psyarxiv.com/8xzqy/
- https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources#automate-your-workflow-open-science-and-reproducible-research

Learn by doing, enjoy!





https://divingintogeneticsandgenomics.rbind.io/post/my-opinionated-selection-of-books-for-bioinformatics-data-science-curriculum/

What questions do you have?

Acknowledgments

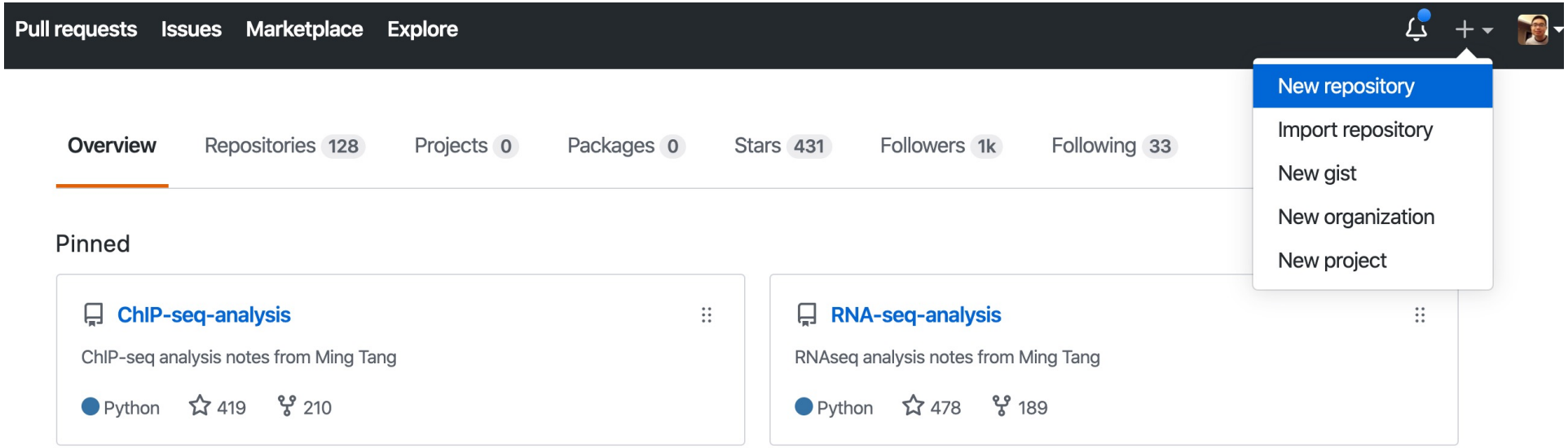
Liu Lab
Shirley Liu

Jenny Bryan
Titus Brown
Data Carpentry https://datacarpentry.org/
All the people who share their wisdom on the web
Thanks!

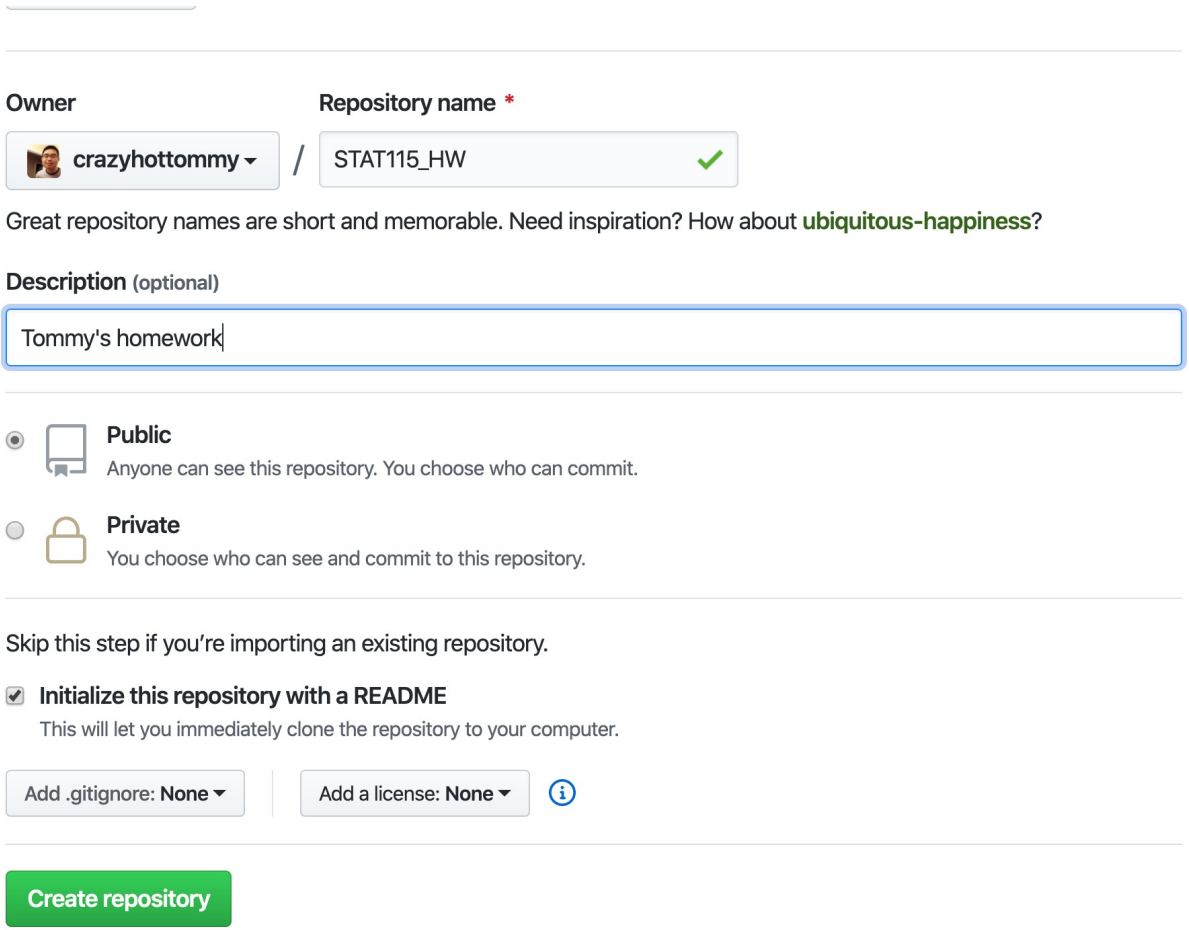
Dana-Farber
Cancer Institute

Reproducible computing using Rstudio: A walk through

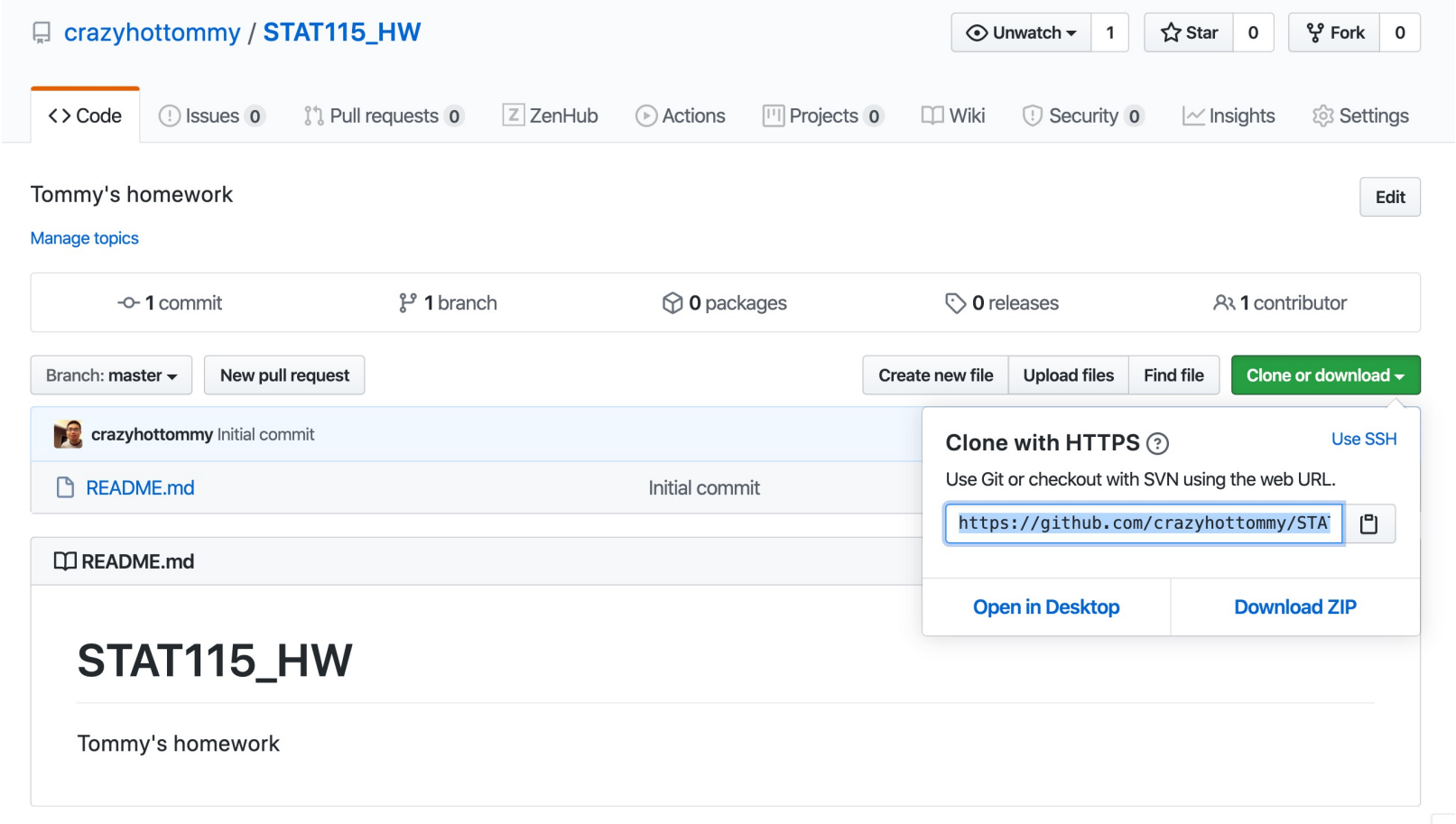
- Go to https://github.com/username
- Create a new repository



Create the new repository
Check [] Initialize this repository with a README



Copy the link from “Clone with HTTPS”



Go back to your local computer, open terminal

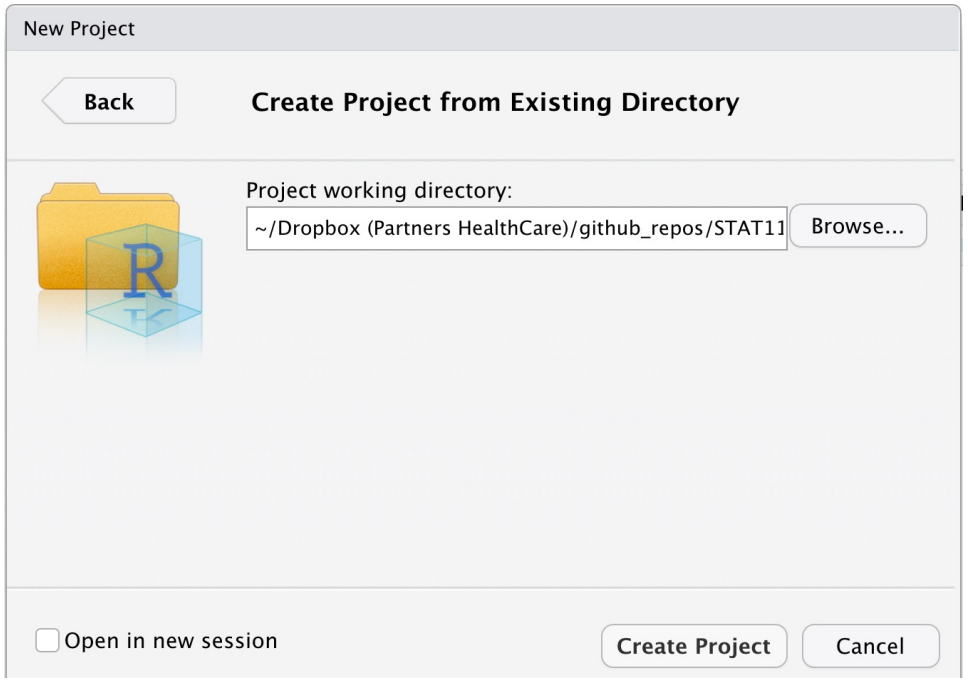
- $ cd /Users/mtang/Dropbox (Partners HealthCare)
- $ mkdir github_repos; cd github_repos
- $ git clone https://github.com/crazyhottommy/STAT115_HW.git
- You should see STAT115_HW folder in the github_repos folder.

e.g.,

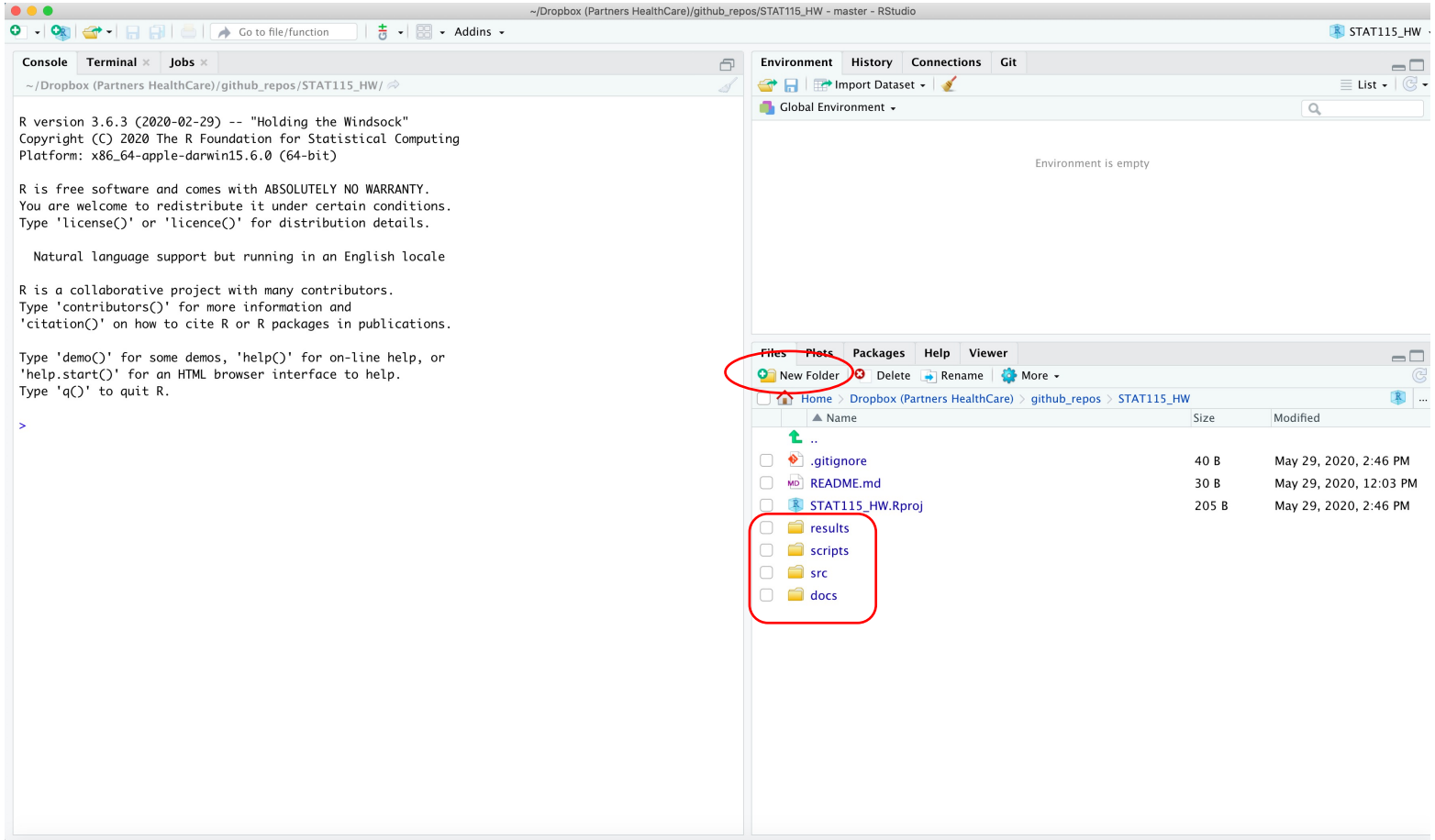
```
(base) ➜  github_repos ls
CIDC_single_cell                chips_automator
DivingIntoGeneticsAndGenomics   cidc_chips
MAESTRO                         computation_wiki
STAT115_HW                      pyflow-ChIPseq
```

I put it in the Dropbox folder since we have unlimited space with Partner's email, so it get backed up in dropbox as well!

Open Rstudio -- > File -- > New Project --> Existing Directory -- > Browse and select the STAT115_HW folder --> Create Project

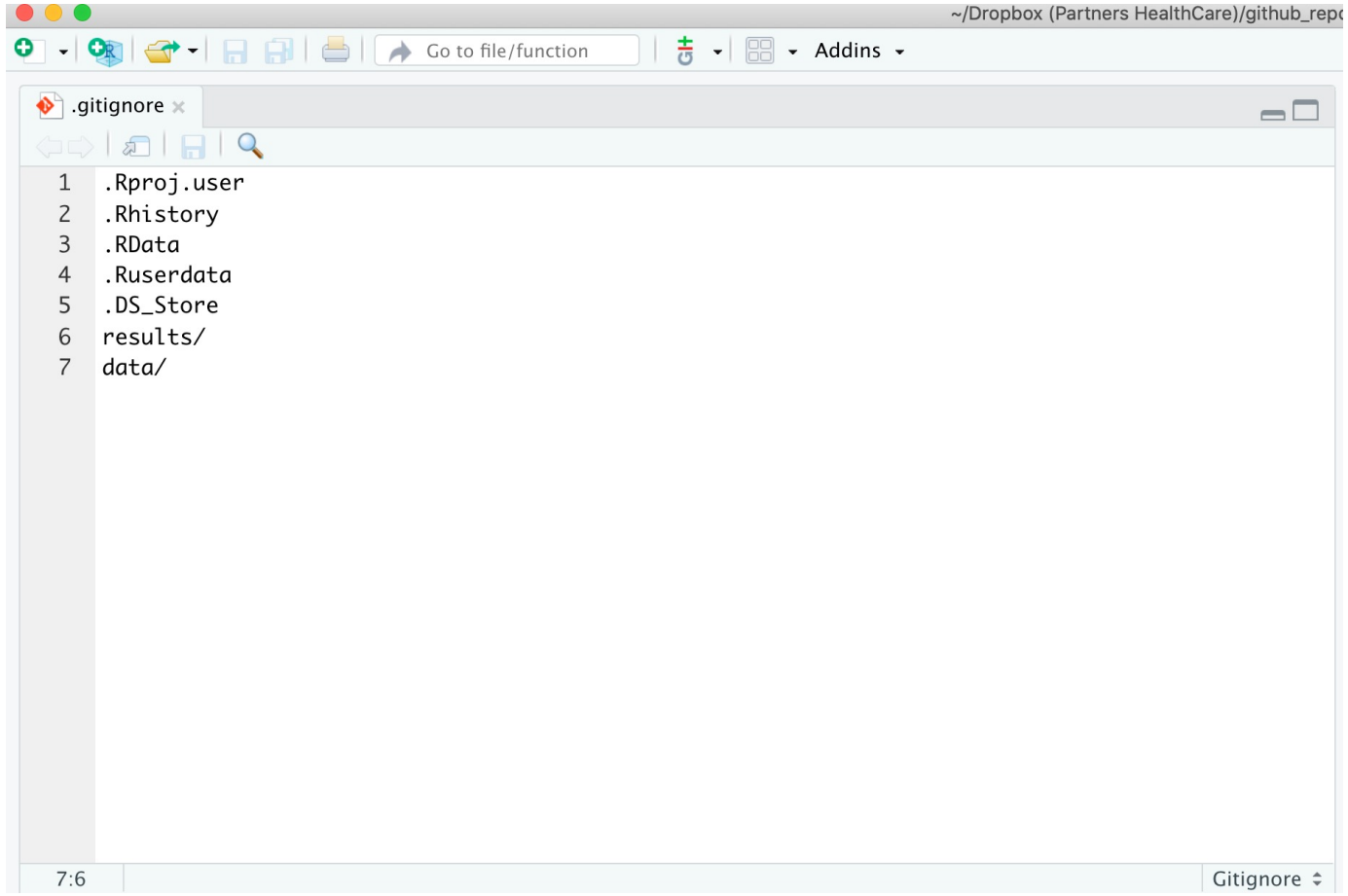


In the Files tab, click New Folder and create data, results, scripts, src and docs folder



The results folder will contain all the results obtained from the script in the scripts folder.
src folder contains R function that you can source from the script in the scripts folder.
Docs folder contains any documentations/manuscripts.

Edit the .gitignore file by clicking it



Ignore
.DS_Store file on mac

I also ask git not to track
Files in the results/ and
data/ folder since they usually
contain big files and intermediate
Files.

This how I do it, you do not have to follow.

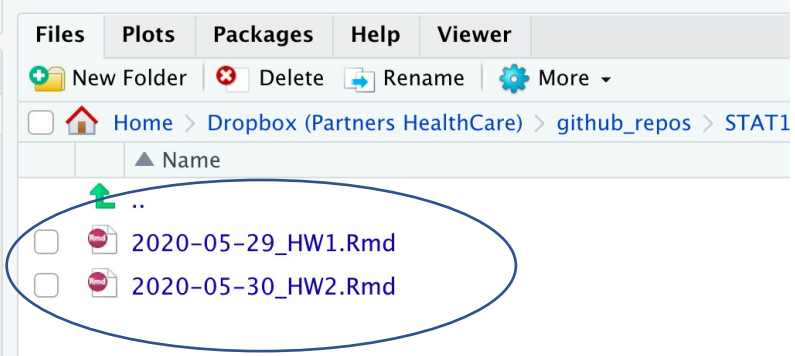
Remember I have them backed
up in dropbox if I want them.

If you want to version control
Large files, check
Git lfs https://git-lfs.github.com/

Now, you can either go to
File -- > New File -- > Rmarkdown

or download the homework Rmd file to the scripts folder.
Click Terminal tab, and use curl to download the Rmd file

Note, I renamed them by prefixing date so they are nicely sorted.

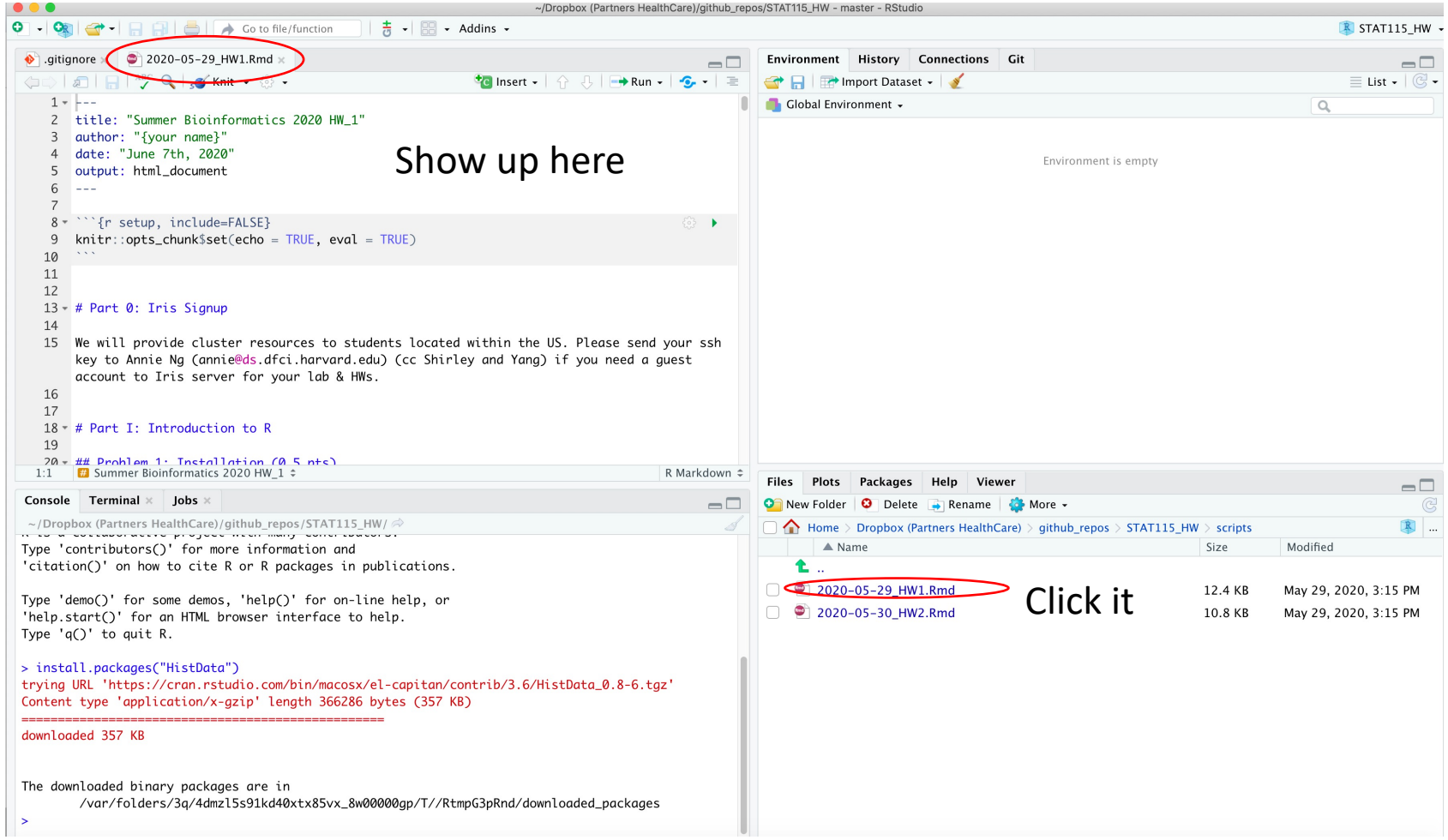


They will show up in the scripts folder

If you name Them:
HW1.Rmd
HW2.Rmd
HW3.Rmd.

These are sorted as well, but I personally like to add date so I have an idea when did I wrote the script.
Or better to use 0 to pad the file name if you have more than 10 files so they are sorted nicely.
01_HW.Rmd
02_HW.Rmd … 10_HW.Rmd

Now, click 2020-05-29_HW1.Rmd and start to work on it.



Git version control

After you worked on the Rmd file and knitted to html, you want to push it to the github. You can either use the Rstudio built-in Git tab or use the Terminal:

- In Rstudio, click the Terminal tab:
- $ git add scripts/2020-05-29_HW1.Rmd
- $ git commit -m “homework 1 done”
- $ git push

- More reading:
- Happy Git with R https://happygitwithr.com/

- 1. we created the github repo first → clone to local → set up Rstudio project.
- 2. if you have already created and worked on a local Rstudio project, you have to do something else:
- $ cd STAT115_HW
- $ git init
- $ git add .
- $ git commit –m “first commit”
- $ git remote add origin https://github.com/crazyhottommy/STAT115_HW.git
- $ git push -u origin master
- Reference:
- https://help.github.com/en/github/importing-your-projects-to-github/adding-an-existing-project-to-github-using-the-command-line