BIOS 6660 course overview

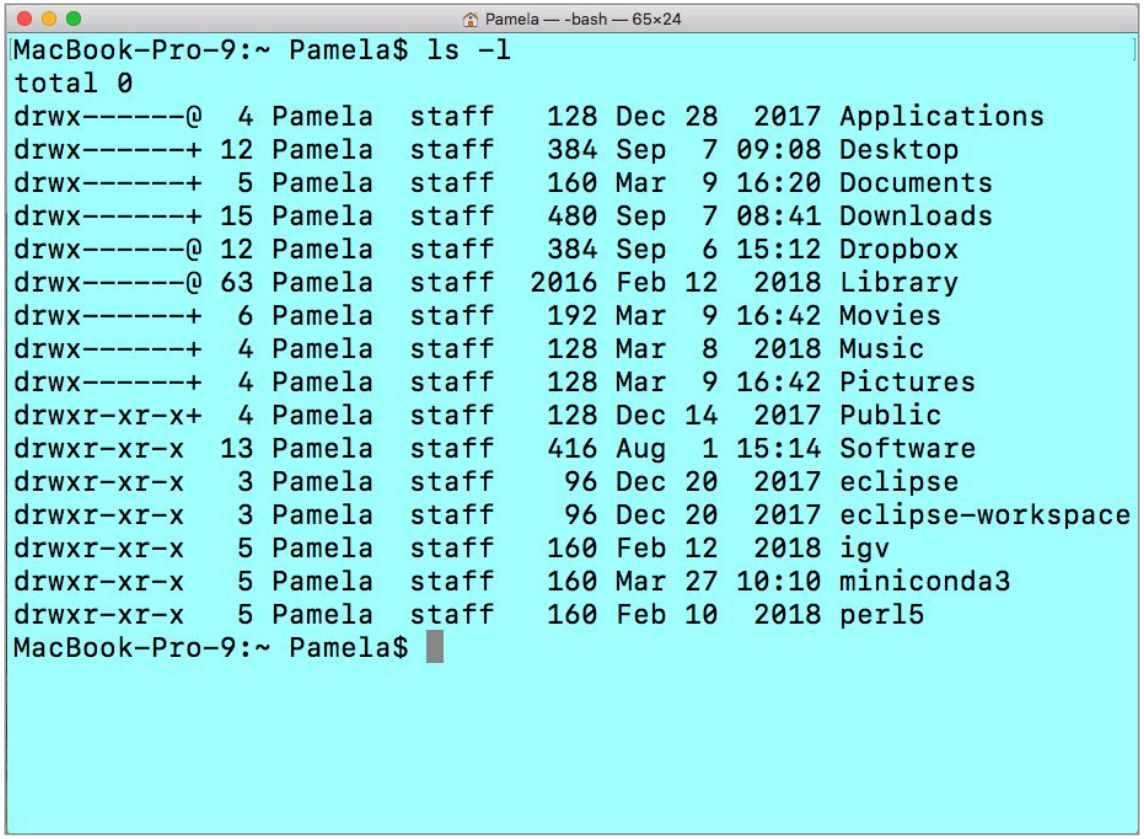
Lecture 1
BIOS 6660, Spring 2019
Instructor: Pam Russell

https://me.me/i/genome-hello-gene-morris-less-5055447

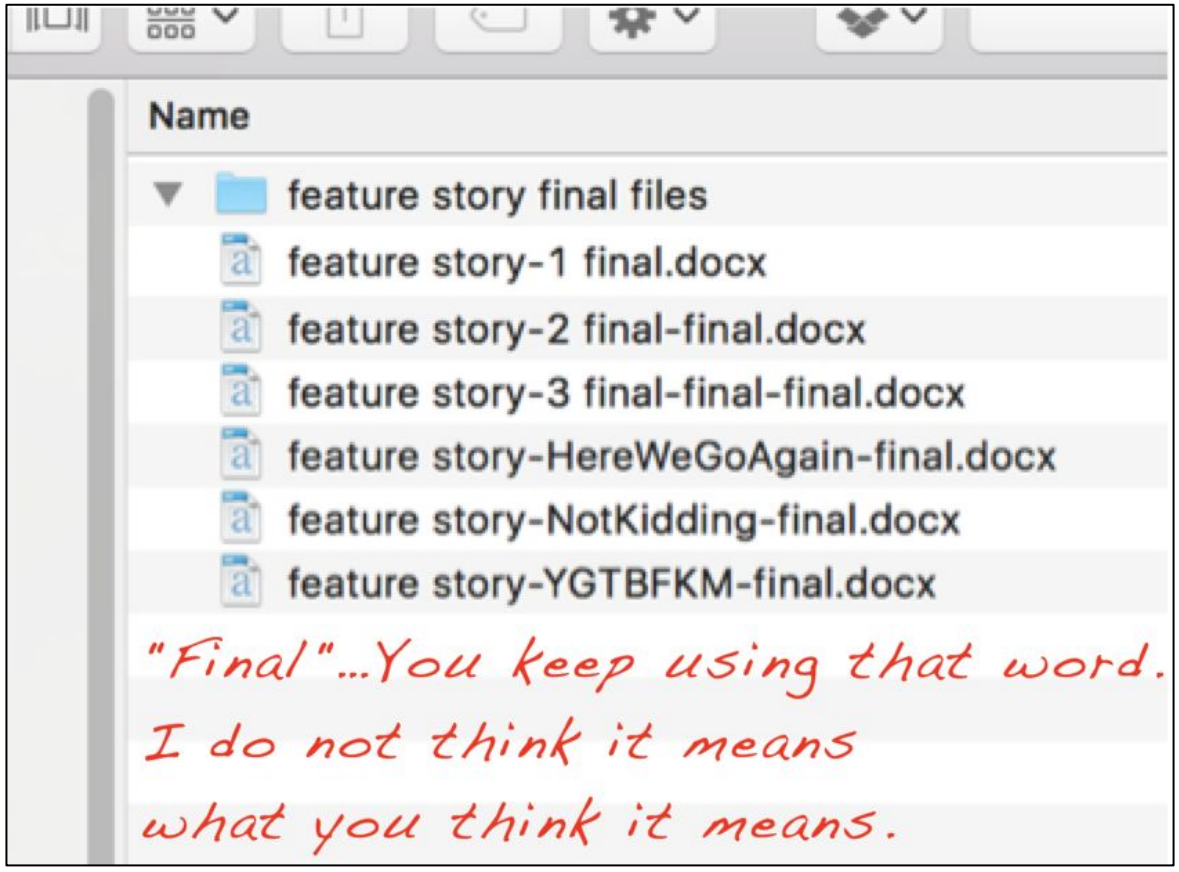
Basic overview

Weeks	Instructor	Topics
1-5	Pam Russell	Toolkit for reproducible data analysis
6-10	Lauren Vanderlinden	RNA-Seq, ChIP-Seq
11-15	Weiming Zhang	DNA methylation, genetic association

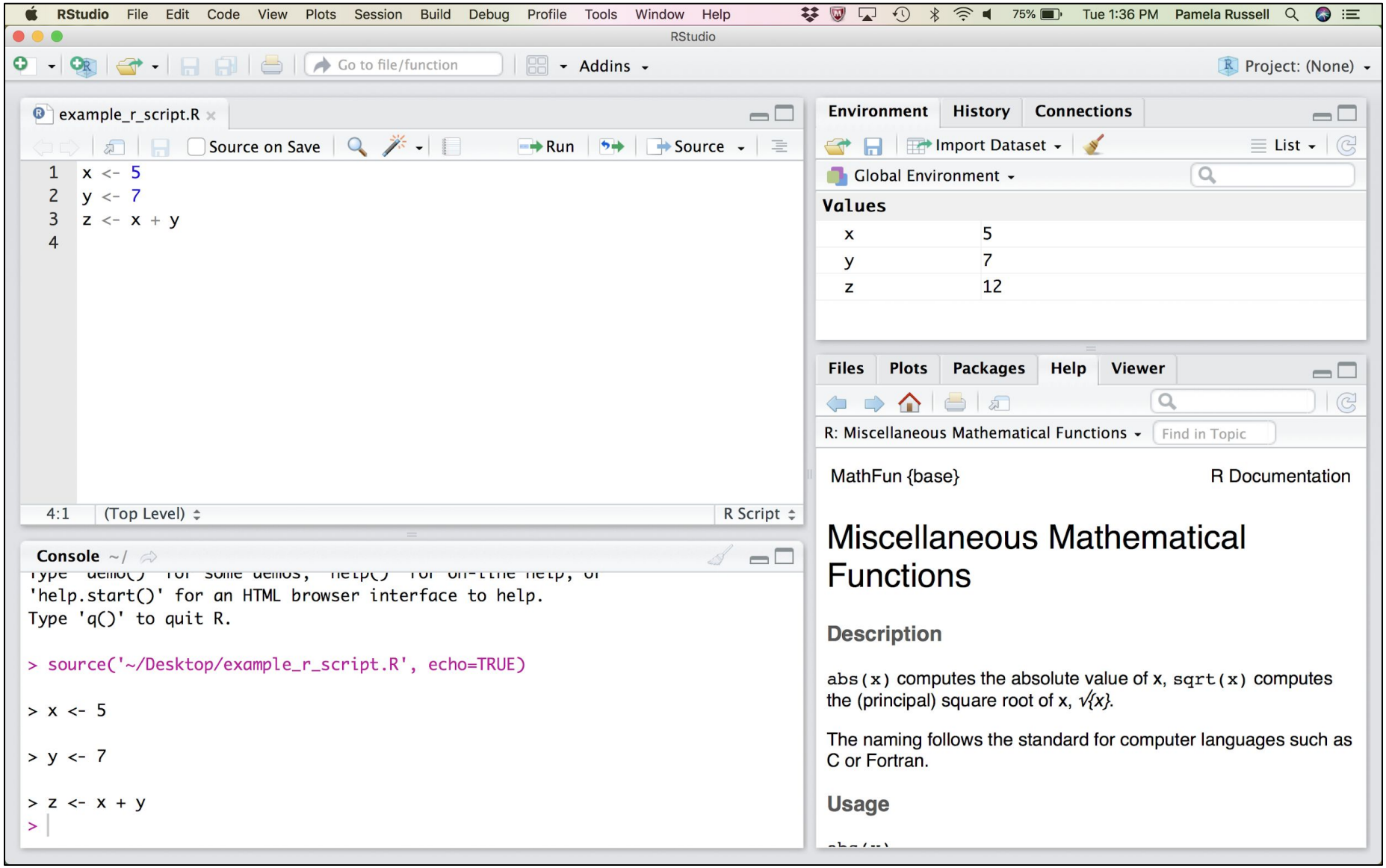
Part 1 topics: Command line



Part 1 topics: Version control



Part 1 topics: R programming



Part 1 topics: R Markdown

RStudio

First_rmd_doc.Rmd

```
---
title: "First R Markdown Document"
author: "Pam Russell"
date: "10/16/2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```
```

Console

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

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

>
```

First R Markdown Document

Pam Russell

10/16/2018

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

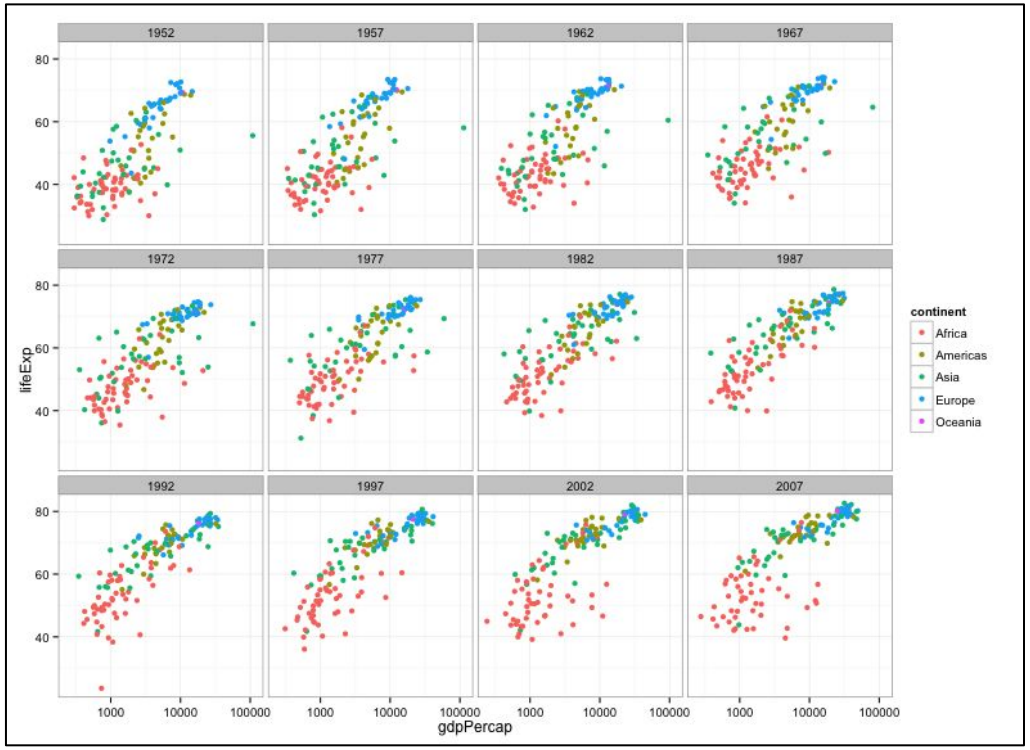
```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

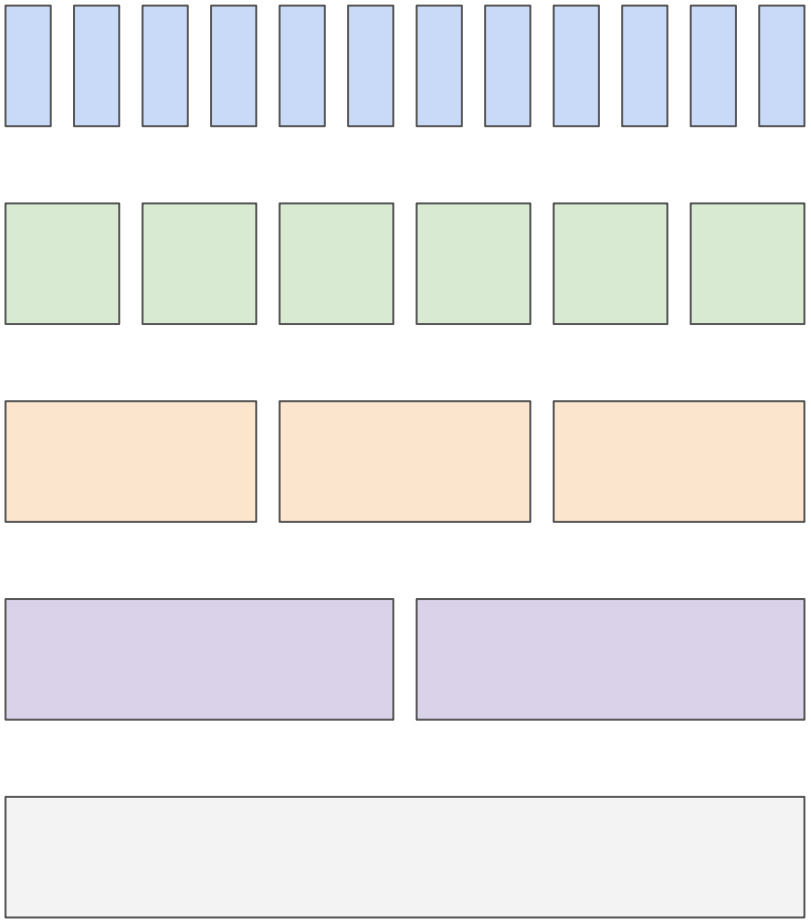
You can also embed plots, for example:

Part 1 topics: R Tidyverse



https://karawoo.com/2015-04-27-wsu/syllabus/ggplot2.html

Part 1 topics: Code organization



Individual statement

Block of code

Whole file

Multiple files

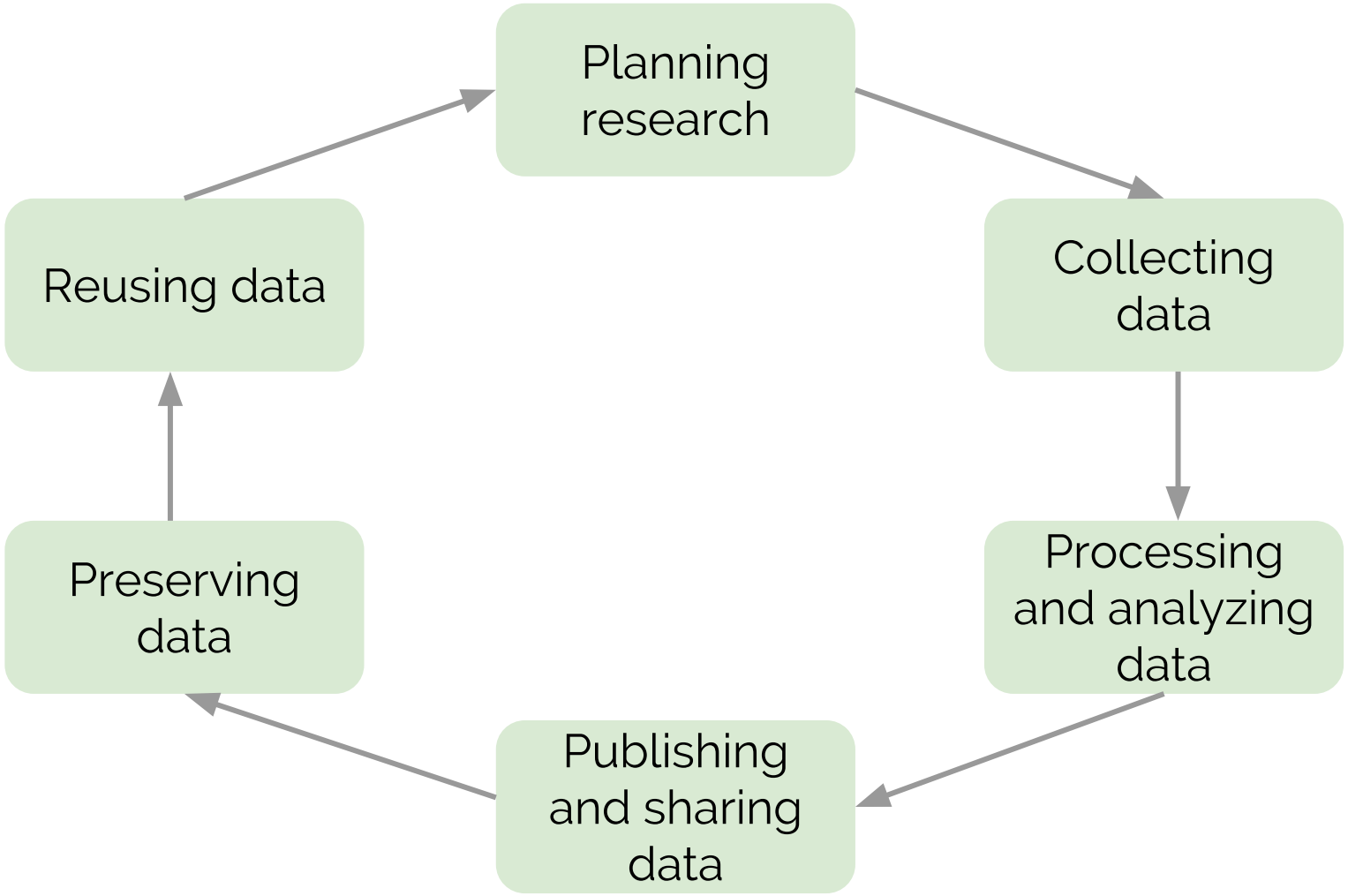
Package

Part 1 topics: Code quality

Correctness

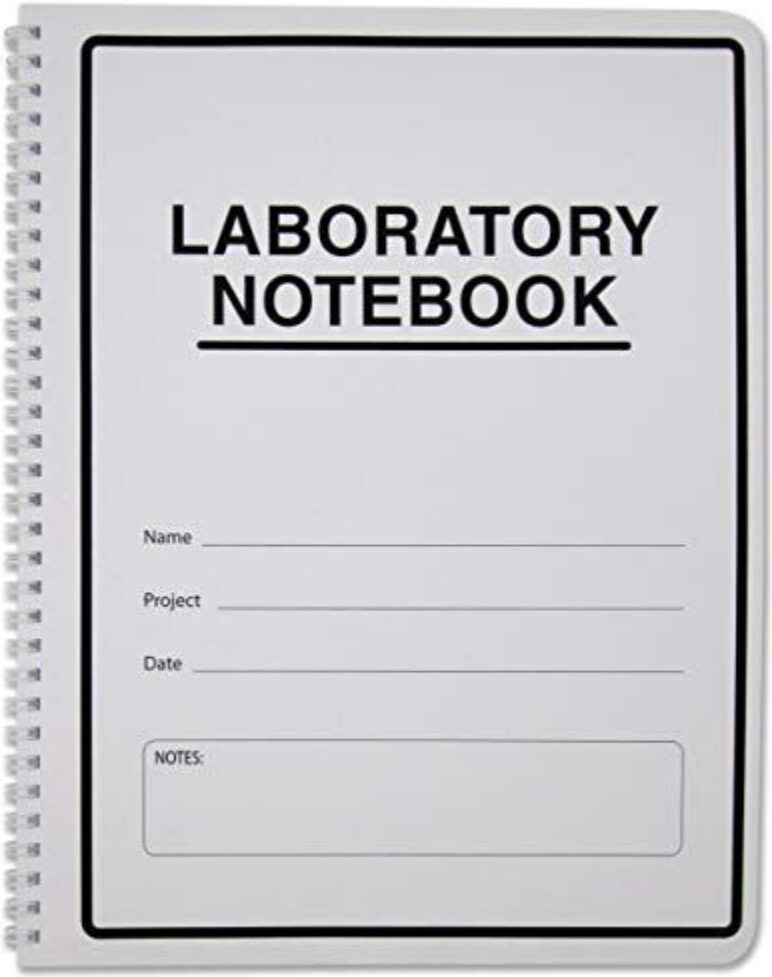
Effective communication

Part 1 topics: Data management

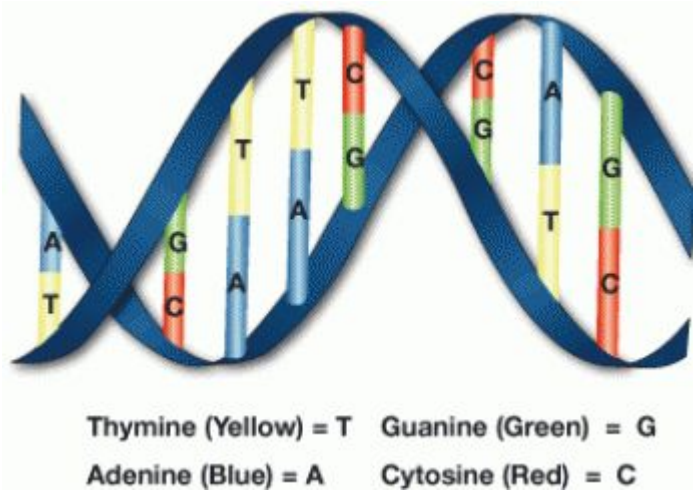


https://www.ukdataservice.ac.uk/manage-data/lifecycle

Part 1 topics: Reproducibility

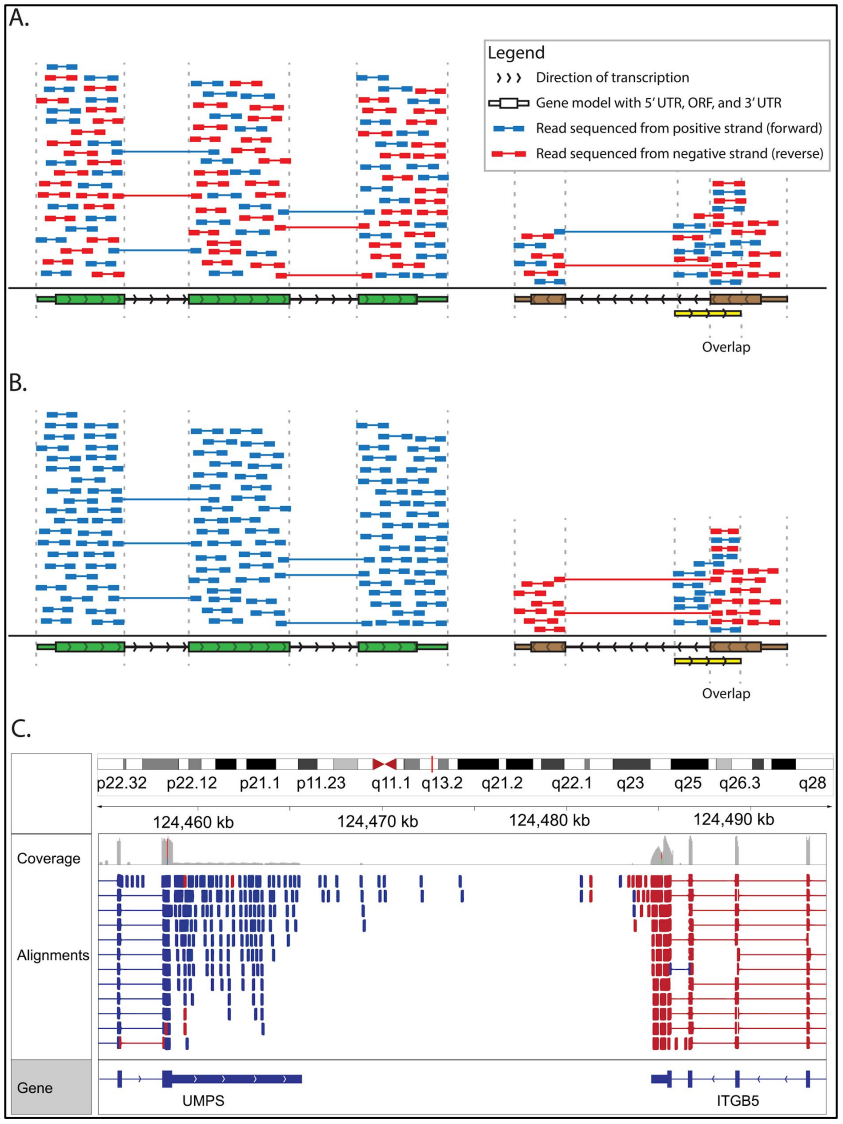


Parts 2 and 3: Genomic data analysis



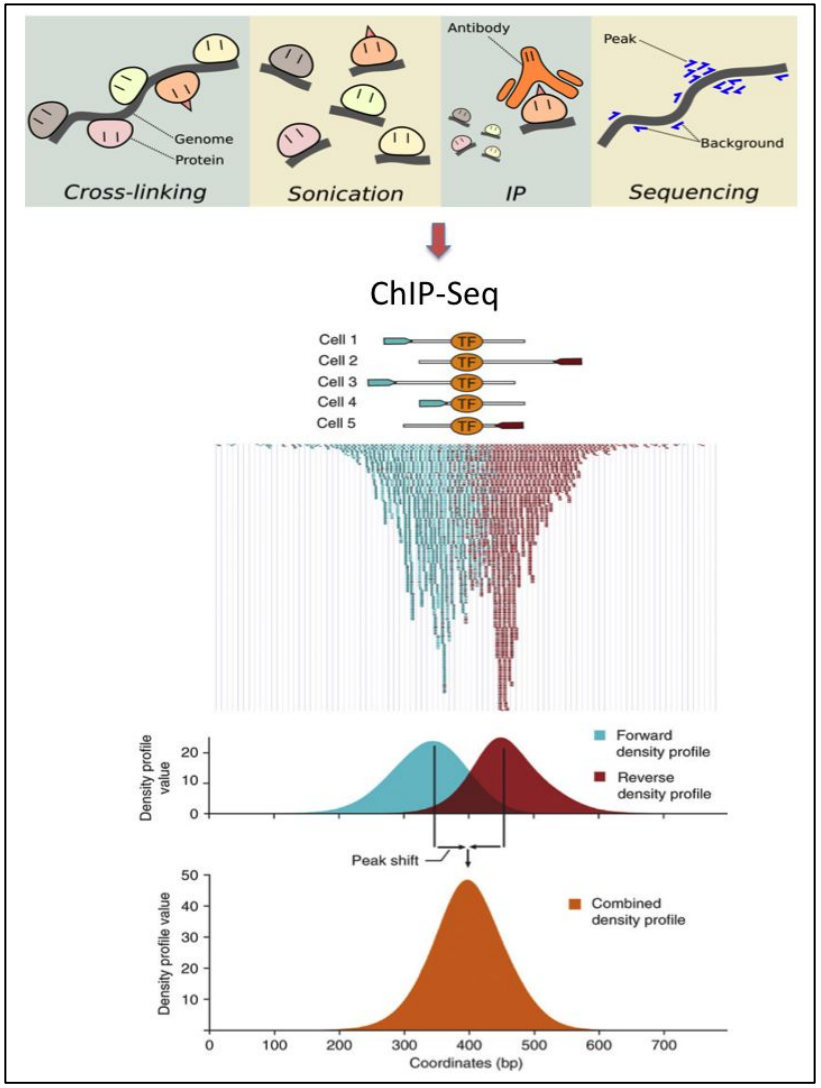
http://reynoldsgen677s10.weebly.com/dna-sequence.html
https://www.illumina.com/systems/sequencing-platforms/hiseq-2500.html

Part 2 topics: RNA-Seq analysis



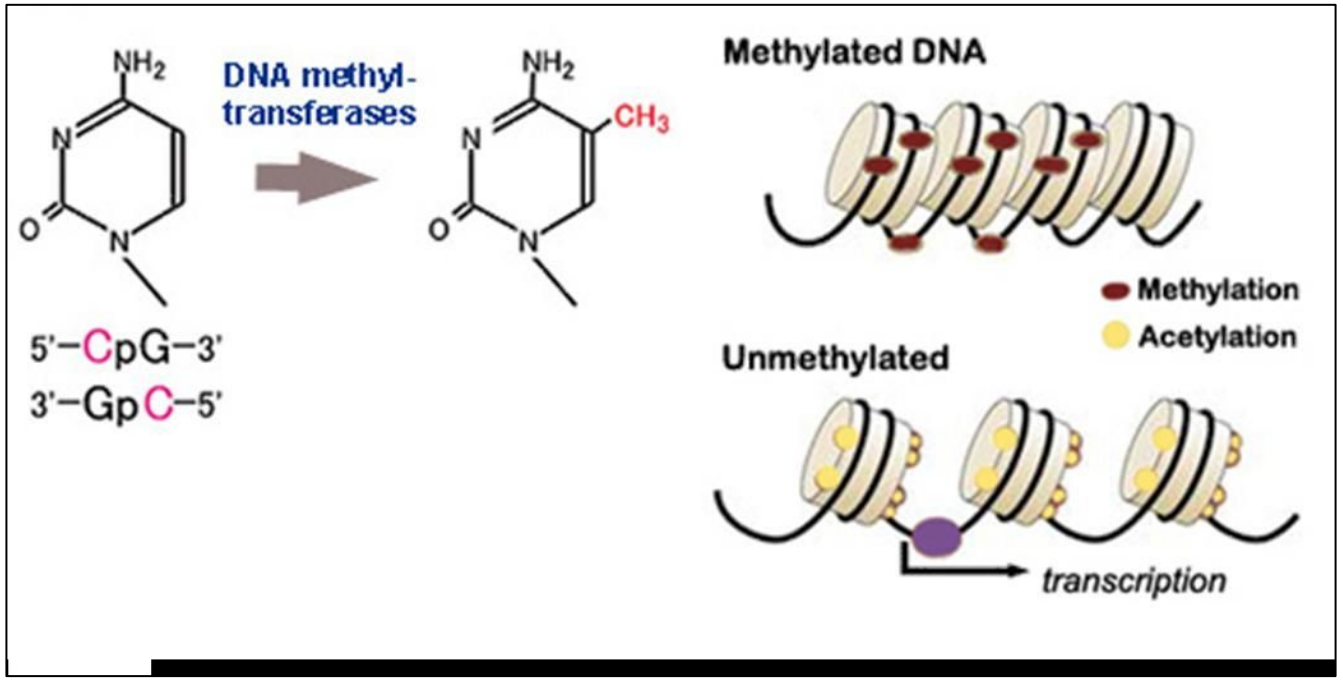
https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1004393.g006

Part 2 topics: ChIP-Seq analysis



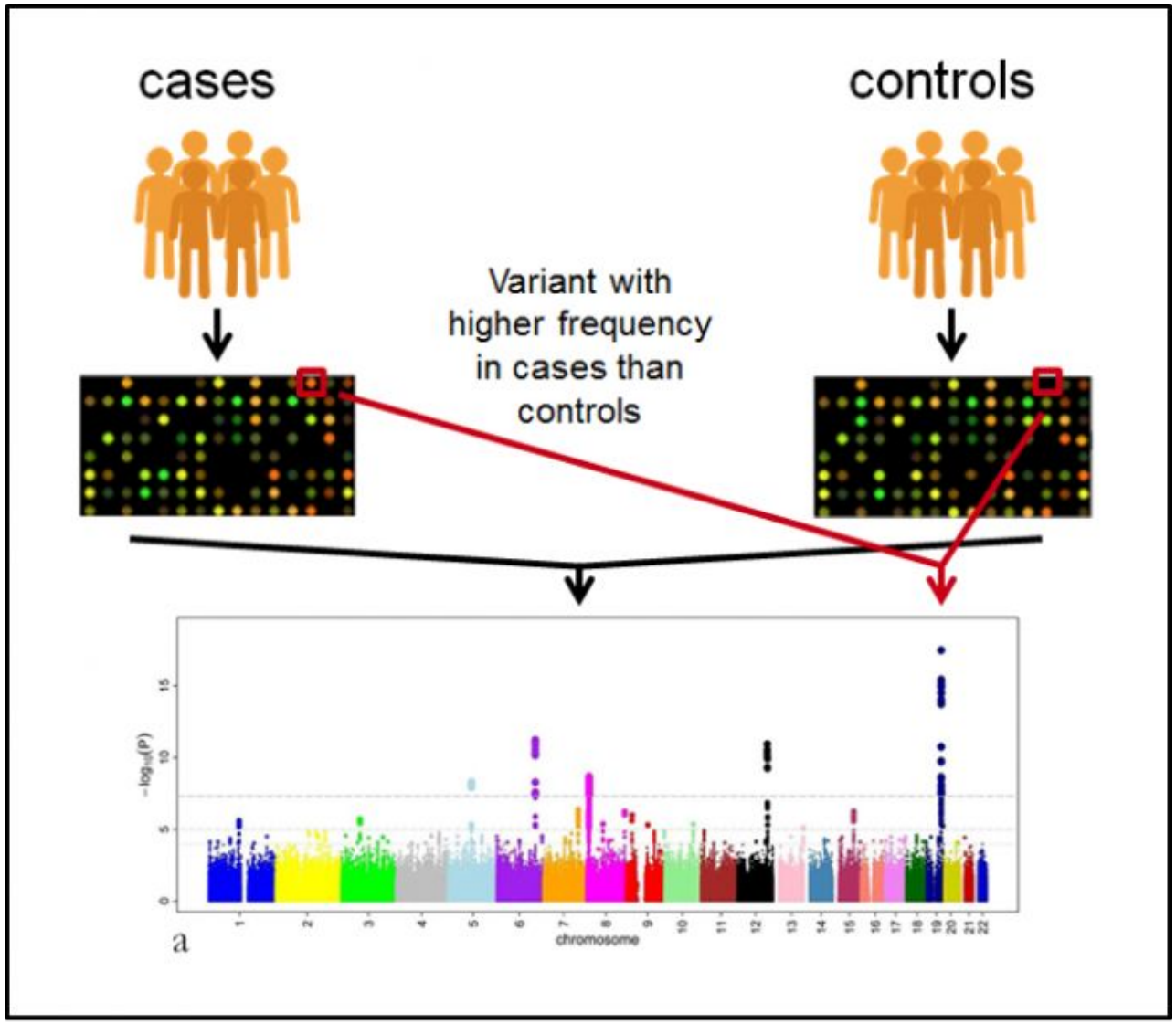
http://biocluster.ucr.edu/~rkaundal/workshops/R_feb2016/ChIPseq/ChIPseq.html

Part 3 topics: DNA methylation analysis



http://helicase.pbworks.com/w/page/17605615/DNA%20Methylation

Part 3 topics: Genetic association



https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome

Everything is on Canvas

- Syllabus
- Schedule
- Lecture slides
- Homework assignments
- Turn in homework

Evaluation

12 homework assignments

No exams

Tools

- ~~Book~~
- R
- RStudio
- R packages
- Command line
- Version control
- Yampa server
- Publically available data

Yampa server details

Some of our work will be done on a remote server called Yampa.

You should have received an email from Pam with login credentials for Yampa. If not, let Pam know ASAP.

Homework 1 first requires you to set up a Linux shell on your computer. From the shell, you can log into Yampa with the command `ssh <username>@yampa.ucdenver.pvt`. You must be on the university network or VPN to access Yampa.

The first time you log into Yampa, immediately type `passwd+[enter]` to change your password.

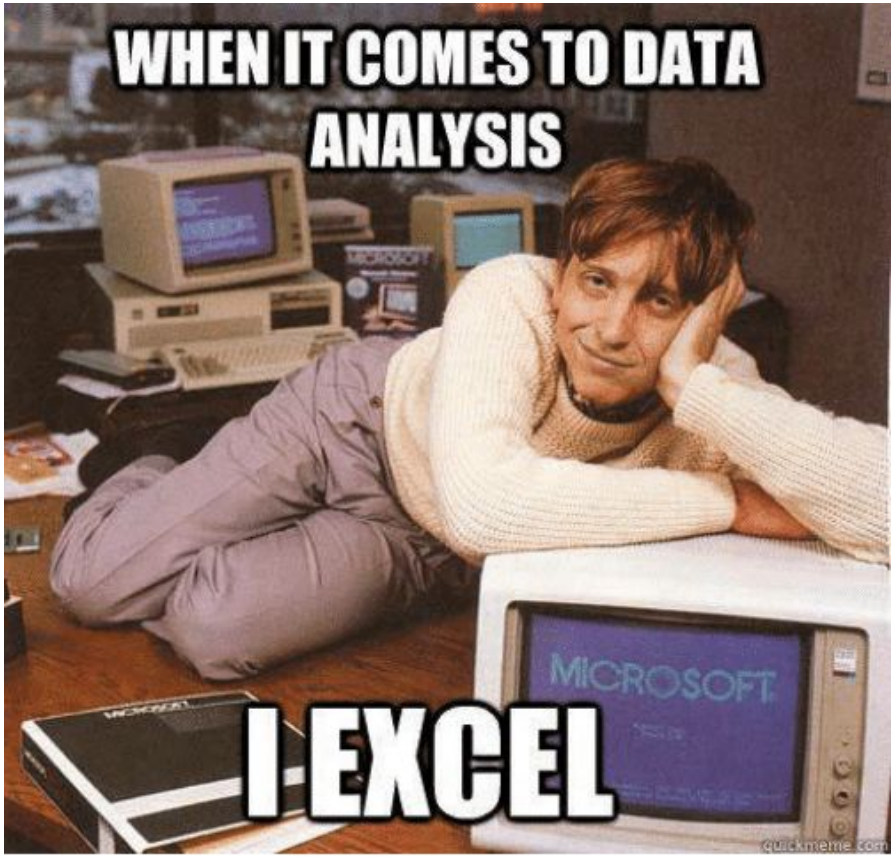
Pam's contact info

Office hours Fridays 10:00-12:00
Or by appointment
Building 406 room 107

pamela.russell@ucdenver.edu

Analysis horror stories

Lecture 1
BIOS 6660, Spring 2019
Instructor: Pam Russell



https://me.me/i/when-it-comes-to-data-analysis-microsoft-10047450

https://tccl.arcc.albany.edu/knilt/index.php?title=File:Keep-calm-and-go-practice-2.png

Data management

Data management

Acquiring, organizing, maintaining, and protecting data

“Then we saw sequences start to vanish and we were like, ‘Oh my god’”

rm -r -f *

https://thenextweb.com/media/2012/05/21/how-pixars-toy-story-2-was-deleted-twice-once-by-technology-and-again-for-its-own-good/

Paper retracted when data can't be found

Retraction: CPAP for the Metabolic Syndrome in Patients with Obstructive Sleep Apnea. N Engl J Med 2011;365:2277-86

TO THE EDITOR:

As the authors of the article entitled "CPAP for the Metabolic Syndrome in Patients with Obstructive Sleep Apnea"[1] published in the *Journal* on December 15, 2011, we regret to report that transcription errors occurred in the Supplementary Appendix, available with the full text of the article at NEJM.org. There were multiple errors in the table on pages 18 and 19 of the Supplementary Appendix concerning data on the accumulation of abdominal fat as assessed with the use of computed tomography and on carotid intima–media thickness as assessed with the use of ultrasonography. These errors, in turn, changed some values in Table 4 of the article. Although these changes do not alter the conclusions of the article, the primary data could not be located to verify corrections made from secondary tables. Accordingly, we have no way of confirming the correct data and, with regret, wish to retract the article.

https://www.ncbi.nlm.nih.gov/pubmed/22168642
https://retractionwatch.com/2013/10/30/nejm-paper-on-sleep-apnea-retracted-when-original-data-cant-be-found/

R

R

Transparent, automated, reproducible analysis

Excel error leads to purchase of toxic assets

THE TI

Max 18C, min 5C

Tuesday September 16 2008 timesonline.co.uk No 69430

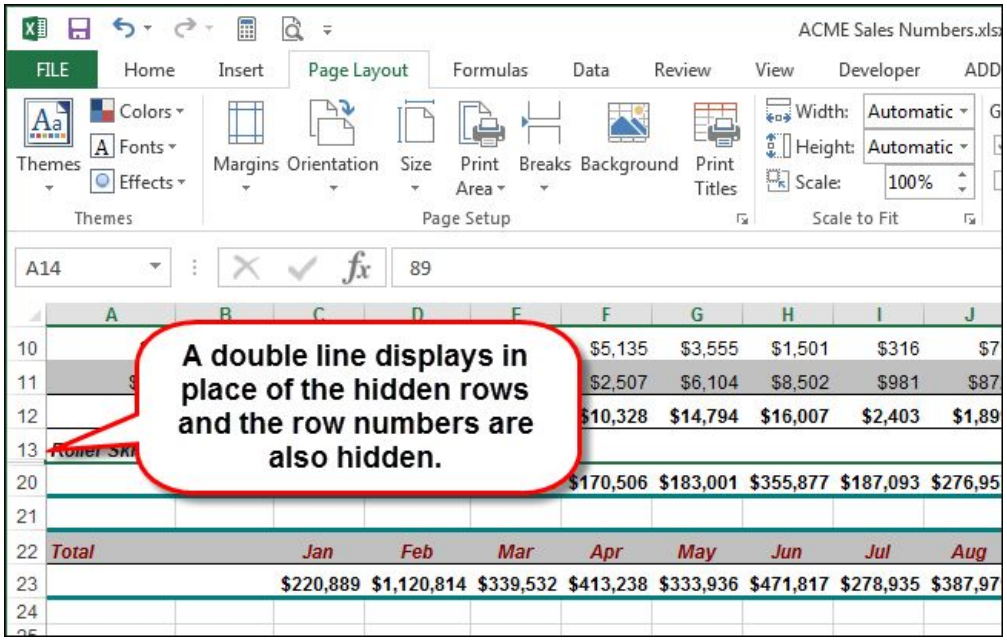
Lehman collapse sends shockwave round world

hares and oil prices plunge, thousands lose jobs

ry Duncan Economics Editor

rs of a global financial meltdown w yesterday as the world's biggest kruptcy plunged markets into noil.

Dow Jones industrial average was down 300 points, or 2.6 per cent. Sentiment was also bolstered by steep falls in oil prices, which dropped by more than $5 a barrel to $96, closing



https://www.computerworld.com/article/2533631/desktop-apps/excel-error-leaves-barclays-with-more-lehman-assets-than-it-bargained-for.html

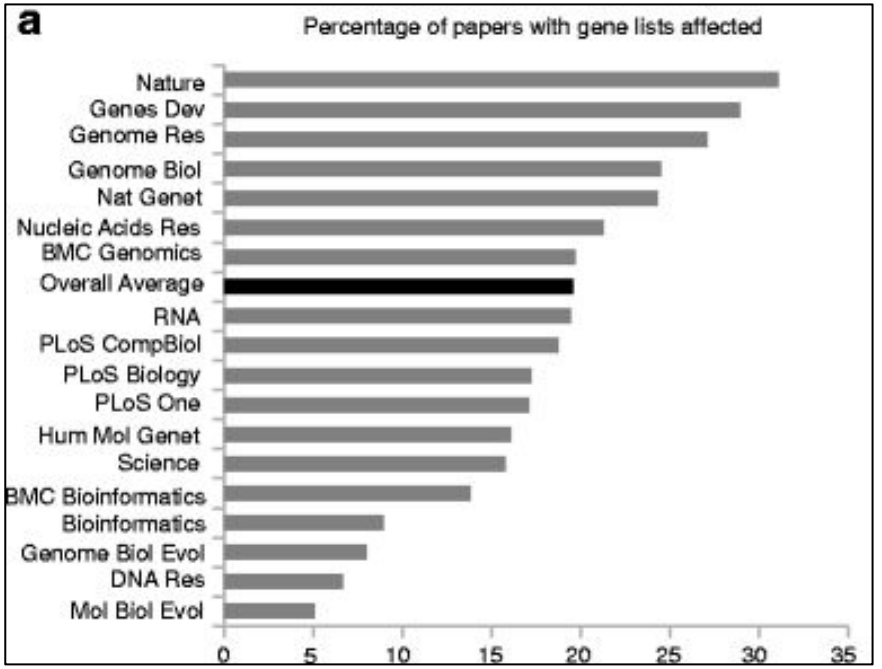
https://www.howtogeek.com/218582/how-to-hide-cells-rows-and-columns-in-excel/

Excel is just trying to help

"MARCH1" ⟶ "1-Mar"

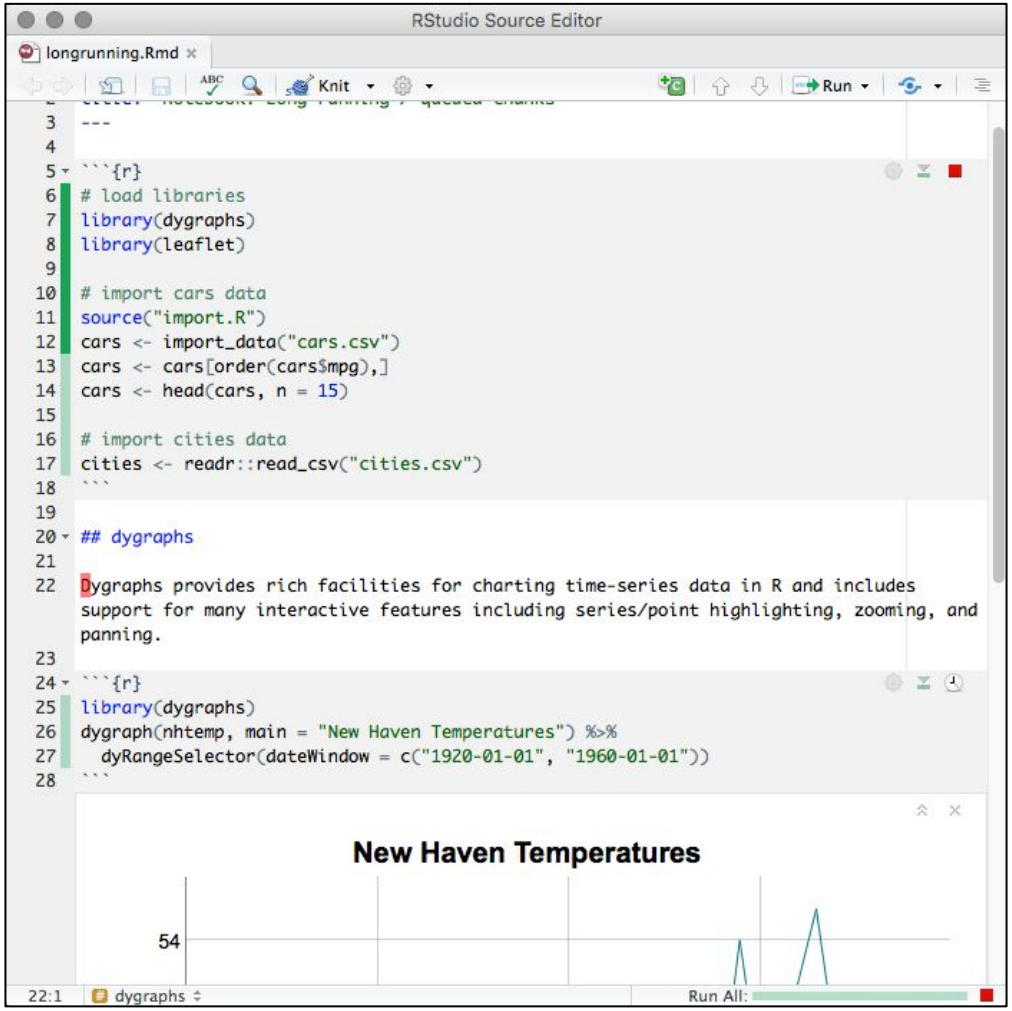
"SEPT2" ⟶ "2006/09/02"

"2310009E13" ⟶ "2.31E+13"



https://doi.org/10.1186/s13059-016-1044-7

R for transparent, end-to-end workflows



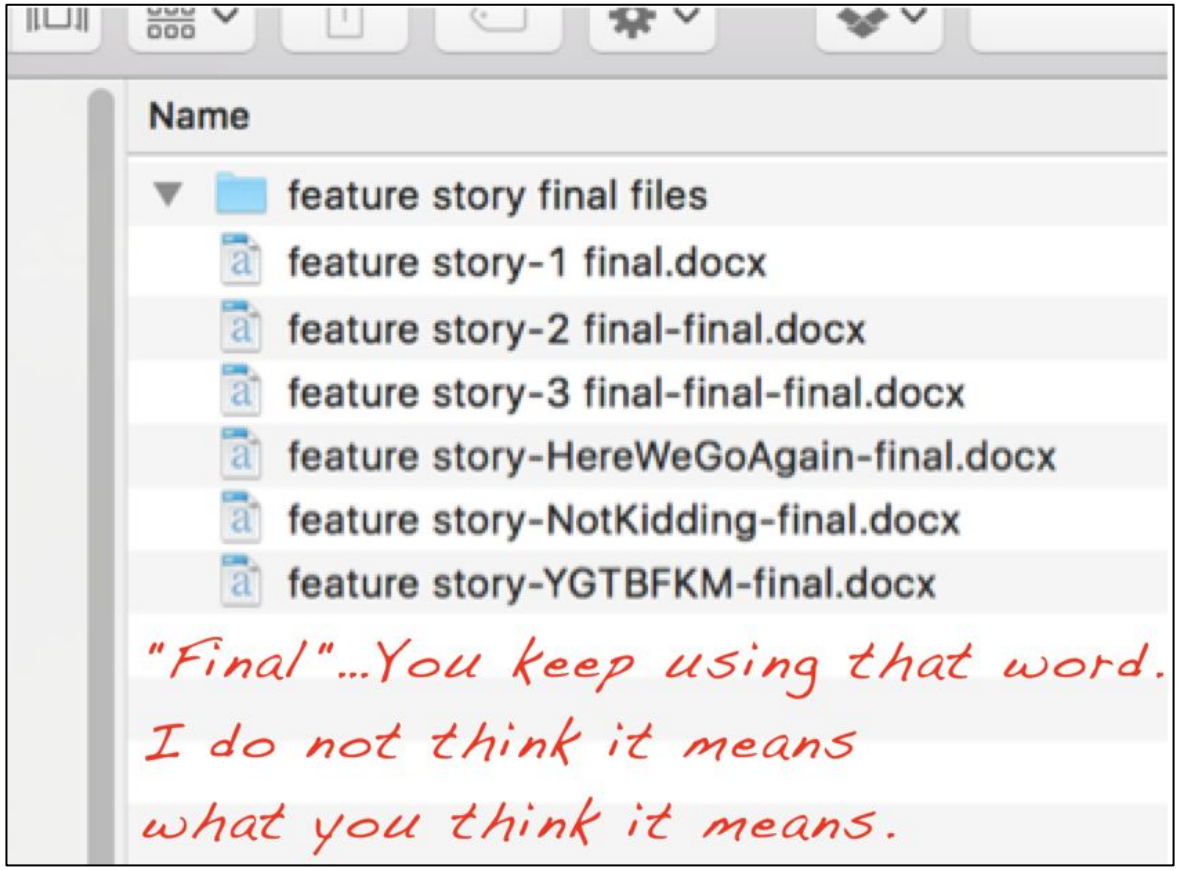
https://bookdown.org/yihui/rmarkdown/notebook.html

Version control

Version control

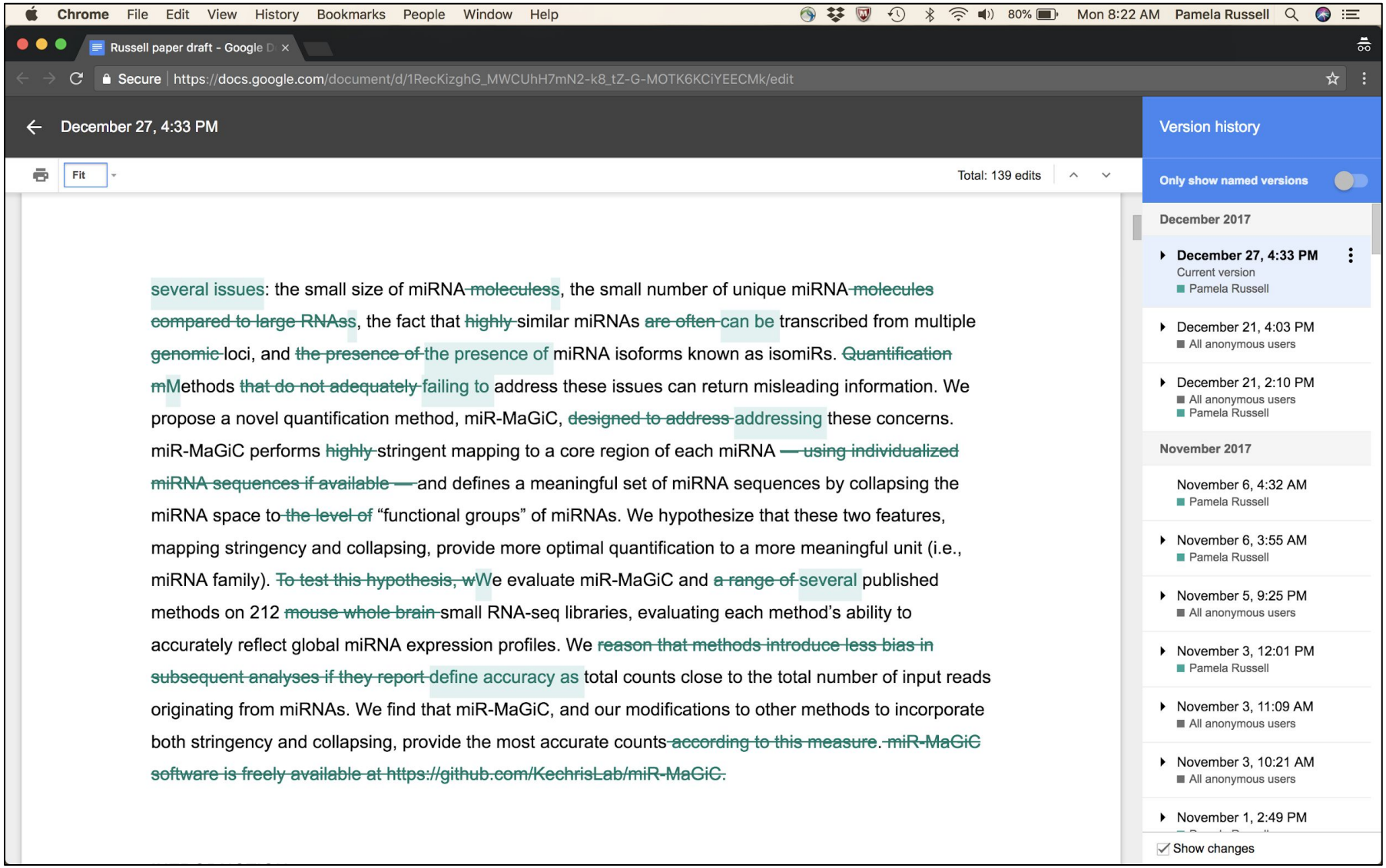
Maintaining an archive of snapshots throughout project life cycle

DIY version control



https://doctorfreelance.com/file-naming-conventions/

Google Docs


Chrome File Edit View History Bookmarks People Window Help
80% Mon 8:22 AM Pamela Russell
Russell paper draft - Google D
Secure https://docs.google.com/document/d/1RecKizghG_MWCUhH7mN2-k8_tZ-G-MOTK6KCiYEECMk/edit
December 27, 4:33 PM
Fit
Total: 139 edits
several issues: the small size of miRNA moleculess, the small number of unique miRNA molecules compared to large RNAss, the fact that highly similar miRNAs are often can be transcribed from multiple genomic loci, and the presence of the presence of miRNA isoforms known as isomiRs. Quantification mMethods that do not adequately failing to address these issues can return misleading information. We propose a novel quantification method, miR-MaGiC, designed to address addressing these concerns. miR-MaGiC performs highly stringent mapping to a core region of each miRNA — using individualized miRNA sequences if available — and defines a meaningful set of miRNA sequences by collapsing the miRNA space to the level of "functional groups" of miRNAs. We hypothesize that these two features, mapping stringency and collapsing, provide more optimal quantification to a more meaningful unit (i.e., miRNA family). To test this hypothesis, wWe evaluate miR-MaGiC and a range of several published methods on 212 mouse whole brain small RNA-seq libraries, evaluating each method's ability to accurately reflect global miRNA expression profiles. We reason that methods introduce less bias in subsequent analyses if they report define accuracy as total counts close to the total number of input reads originating from miRNAs. We find that miR-MaGiC, and our modifications to other methods to incorporate both stringency and collapsing, provide the most accurate counts according to this measure. miR-MaGiC software is freely available at https://github.com/KechrisLab/miR-MaGiC.
Version history
Only show named versions
December 2017
December 27, 4:33 PM
Current version
Pamela Russell
December 21, 4:03 PM
All anonymous users
December 21, 2:10 PM
All anonymous users
Pamela Russell
November 2017
November 6, 4:32 AM
Pamela Russell
November 6, 3:55 AM
Pamela Russell
November 5, 9:25 PM
All anonymous users
November 3, 12:01 PM
Pamela Russell
November 3, 11:09 AM
All anonymous users
November 3, 10:21 AM
All anonymous users
November 1, 2:49 PM
Show changes

Find and replace

10 –> ten

A clearer head would have foreseen the carnage. 110 became **1ten**, 5.10 became **5.ten** and 101010 became, well, you get the idea. One of the things I learnt that day is that banking websites have a lot of numbers!

ten –> 10

This was clearly pure genius as it immediately fixed everything I'd just broken. It also brought another problem to my **at10tion**.

Code quality

Code quality

Effective code design => fewer errors

NASA loses Mars Climate Orbiter

Confusion leads to Mars failure

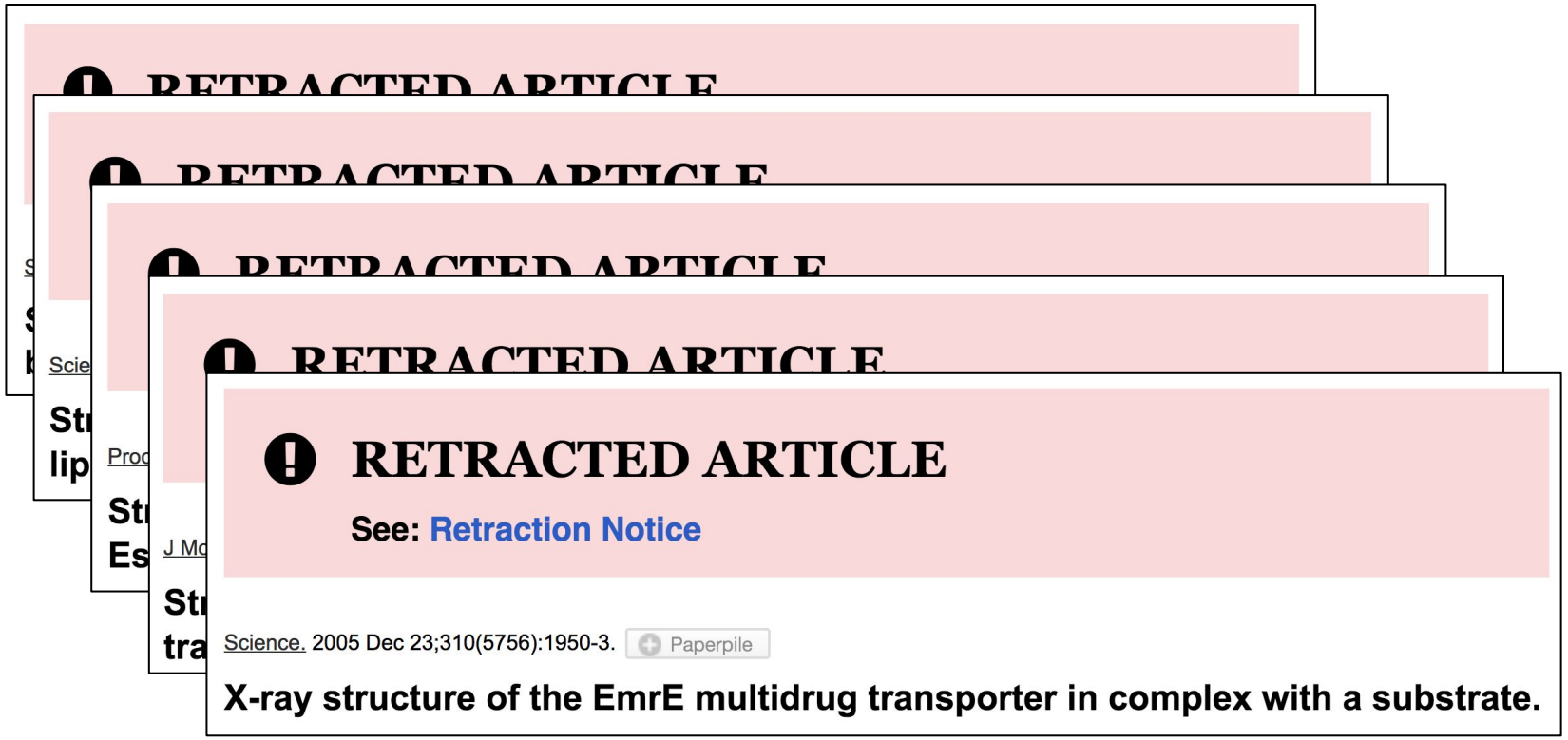
The Mars Climate Orbiter: Now in pieces on the planet's surface

The Mars Climate Orbiter Spacecraft was lost because one Nasa team used imperial units while another used metric units for a key spacecraft operation.

"**People sometimes make errors**," said Dr. Edward Weiler, NASA's Associate Administrator for Space Science. "**The problem here was not the error**, it was the failure of NASA's systems engineering, and the checks and balances in our **processes to detect the error**. That's why we lost the spacecraft."

http://news.bbc.co.uk/2/hi/science/nature/462264.stm
https://www.jpl.nasa.gov/news/releases/99/mcoloss1.html

"I hope people will understand that it was a mistake, and I'm very sorry for it."



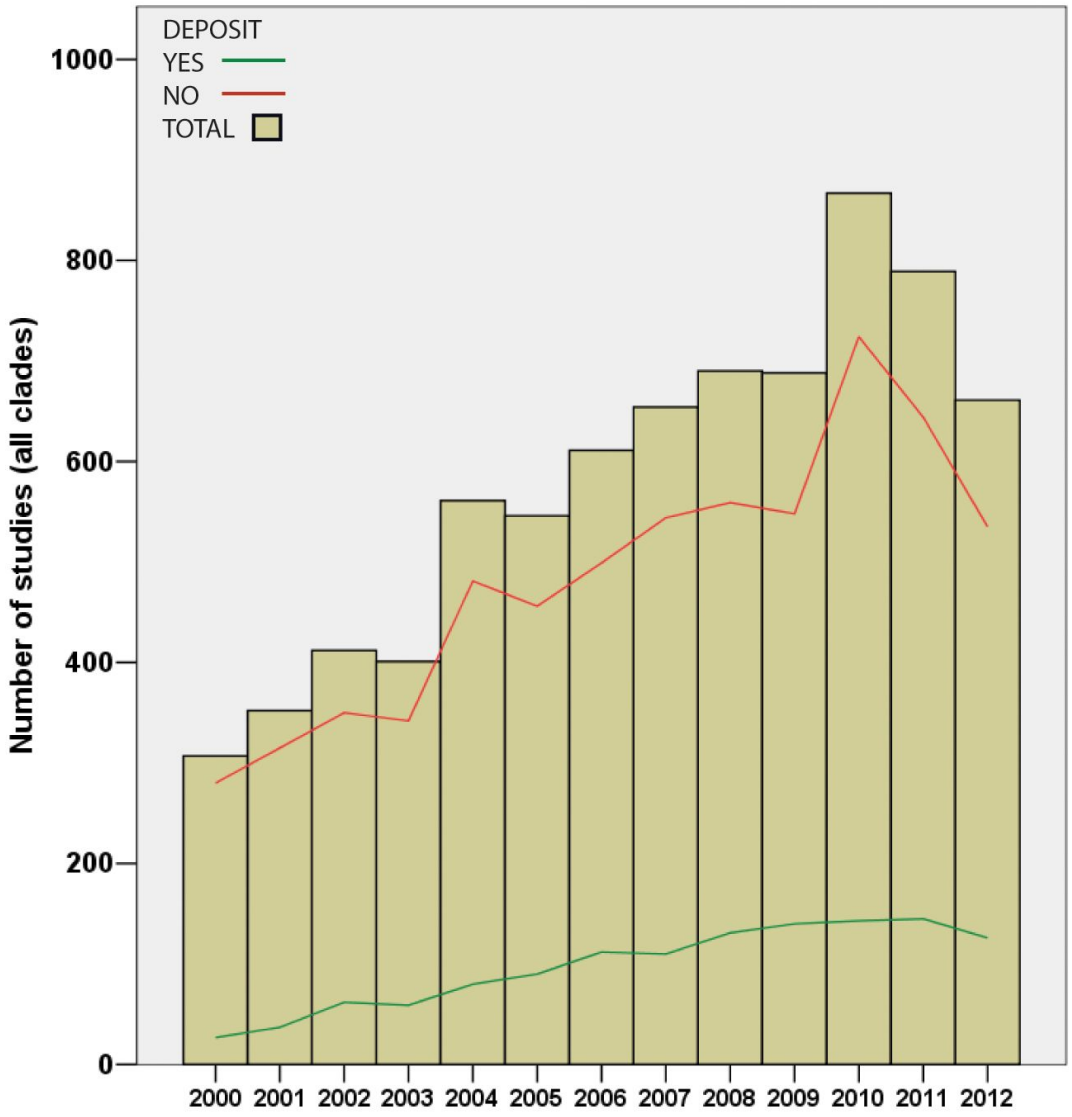
https://doi.org/10.1126/science.314.5807.1856

Reproducibility

Reproducibility

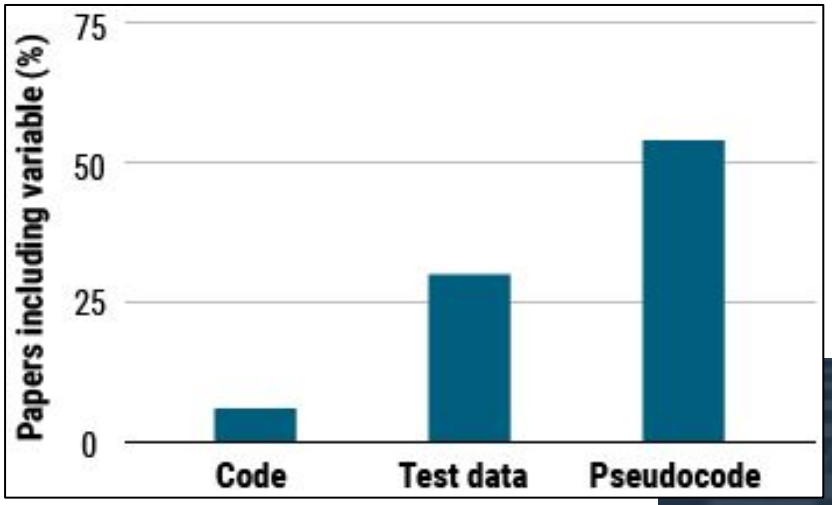
The ability to recapitulate the results in a paper from publicly available data, code, and workflows

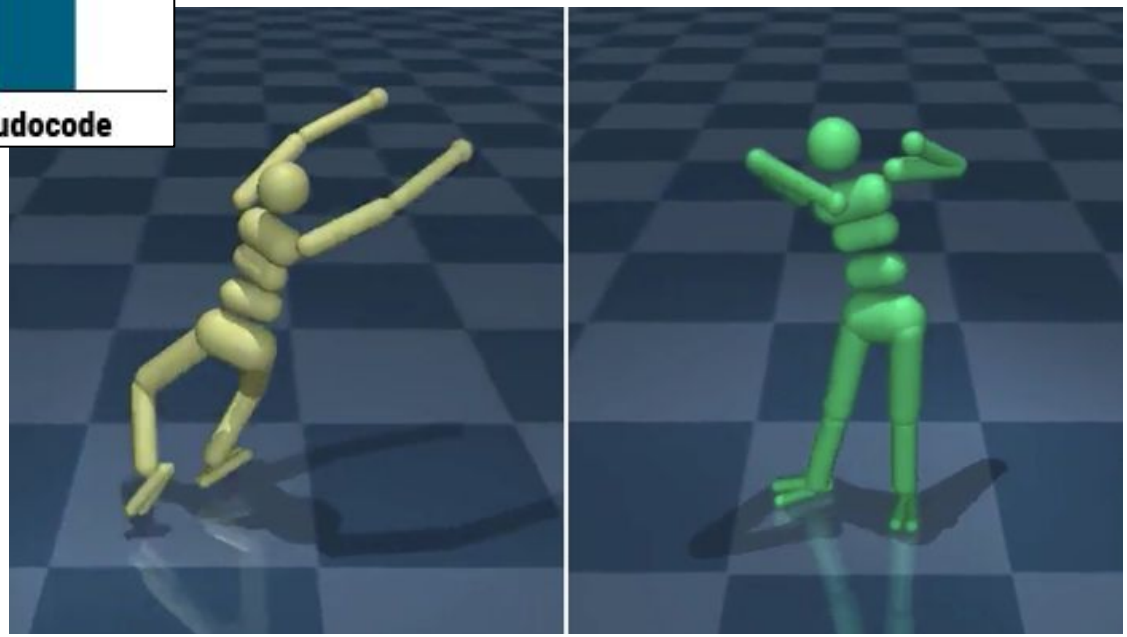
Lost branches on the tree of life



https://doi.org/10.1371/journal.pbio.1001636

“We tried for two months and we couldn’t get anywhere close.”





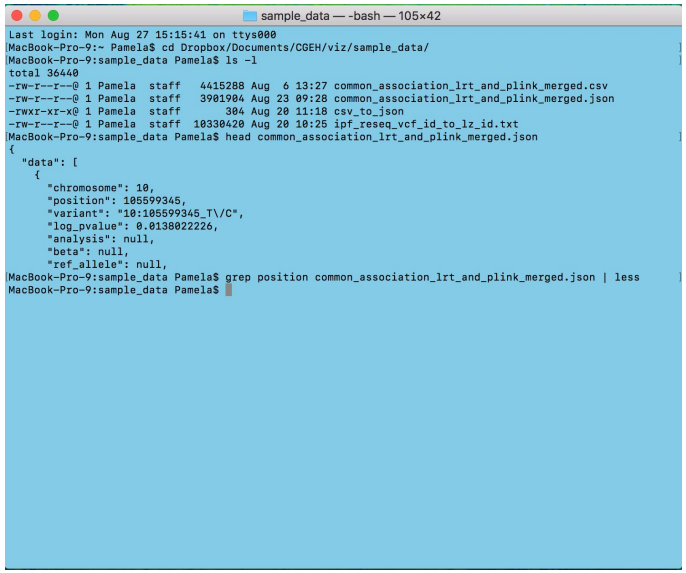
https://doi.org/10.1126/science.aat3298

Command line

Command line

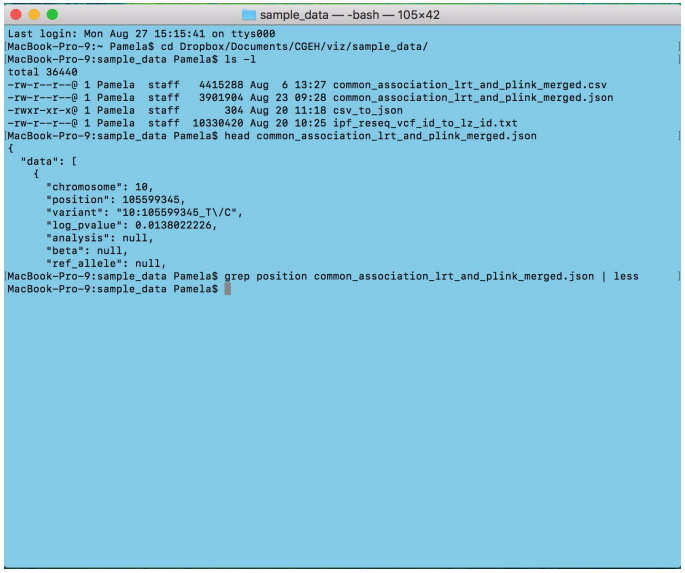
Interact with computer through commands

Text processing one liners



- Print first 10 lines of file
- Find and replace string inside file
- Count lines or characters in file
- Sort file in decreasing order by column 5
- Make a new file with columns 7, 3, 6 of original file
- Find differences between two files
- Delete all blank lines from file
- Remove duplicate lines from file
- View file with long lines chopped

Running analyses



- Most genomics tools work from command line
- Keep sets of commands in scripts for later reference and reuse
- Keep scripts under version control
- Manage version control from command line

Preview of what's next

Lecture 2

Command line and version control

Homework 1

- Currently posted; due next Tuesday
- Some problems are just tool setup that you can begin working on now
- Some problems use material from lecture 2

Getting help with homework

Office hours on Fridays are timed with assignments due Tuesdays

I won't answer last-minute emails on Monday night! :)