MONASH University

MONASH
BIOINFORMATICS
PLATFORM

Introducing R

Intro to R Workshop
Trainers –
Helpers –

@MonashBioinfo

Bioinformatics.platform@monash.edu

http://bioinformatics.erc.monash.edu

GROUP OF EIGHT AUSTRALIA

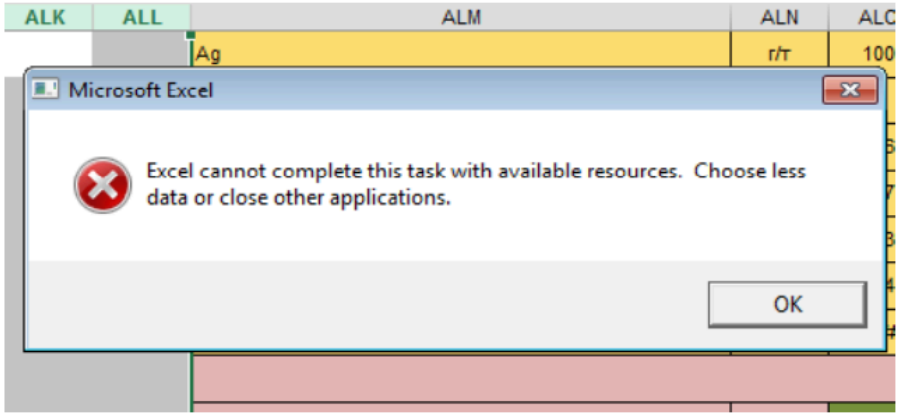
Today's Objective(s) – What we hope you will get out of today

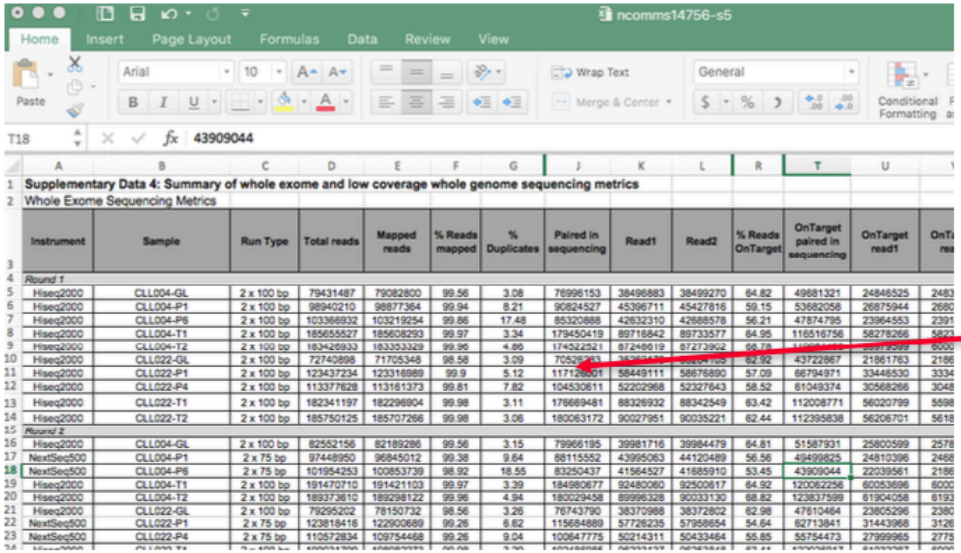
- Introduction to R
- The basics of R – data objects
- Data plotting
- How to find more information for a specific task
- The start of the journey, a network of R learners/users around you.
- Entry into more advanced and specialised uses of R for your research.

Why R? (as opposed to Excel?)

- It's free, legally free. Active development.
- Reproducible analysis.
  - Document what you have done with your data in code.
  - Come back to it in days, months, years and you should know what was done.
- Low risk of inadvertent data loss/mutation.
  - By design, R requires you to load your data in, what you do with the data is then written in code (R language).
- R can handle really large datasets.
  - Excel is limited by 1,048,576 rows and 16,384 columns.
- Collaborative.
  - Share your data and analysis.

Why R? (as opposed to Excel?)





Typo and saved it

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

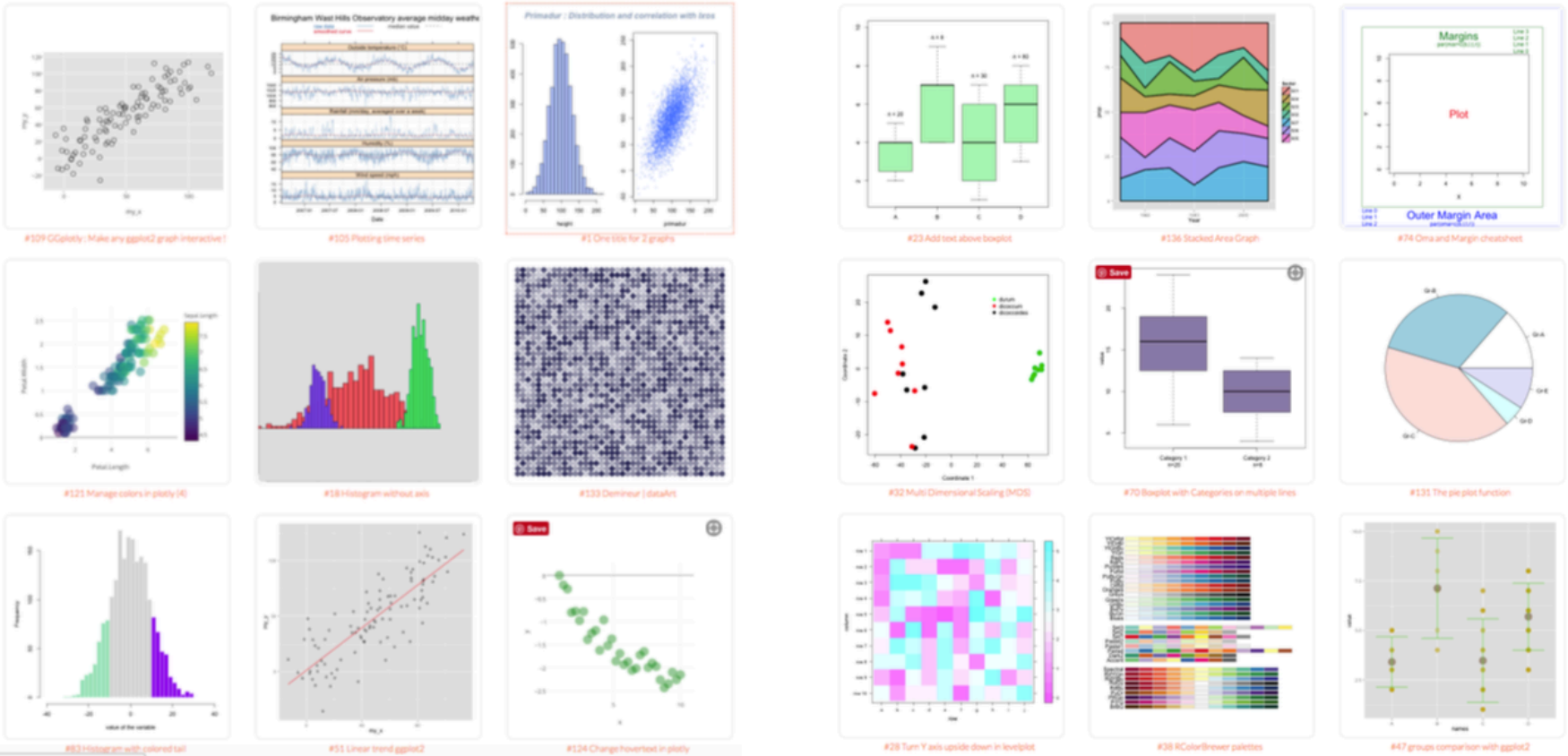
Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for

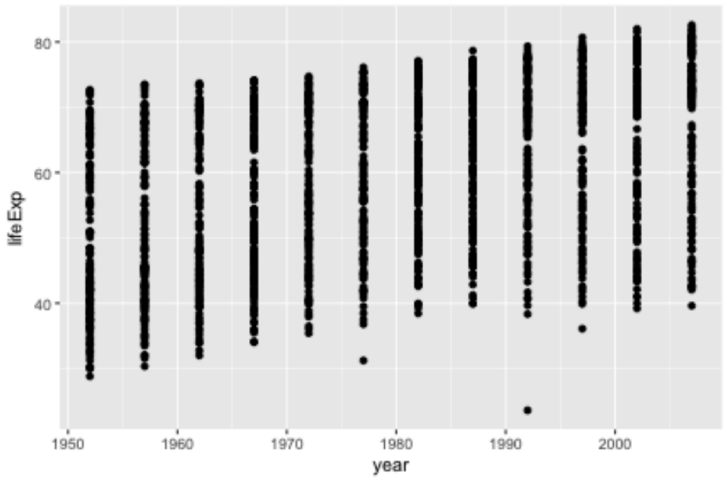
Why R? (as opposed to Excel?)

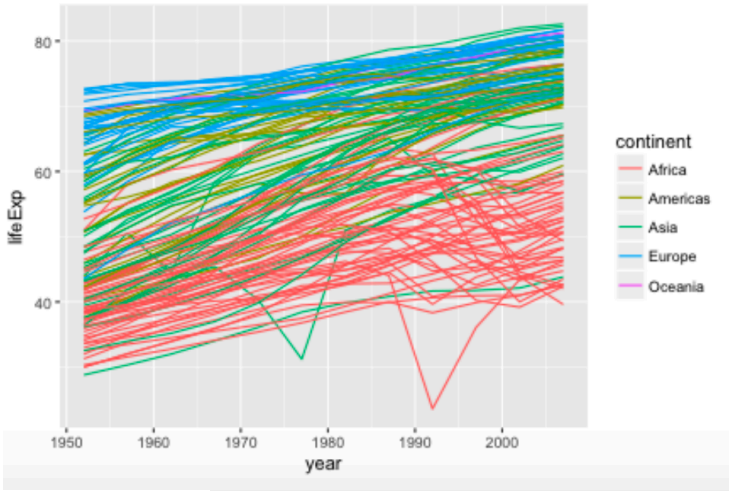
- Graphs and plotting.
- Can you plot a box and whisker plot in excel?

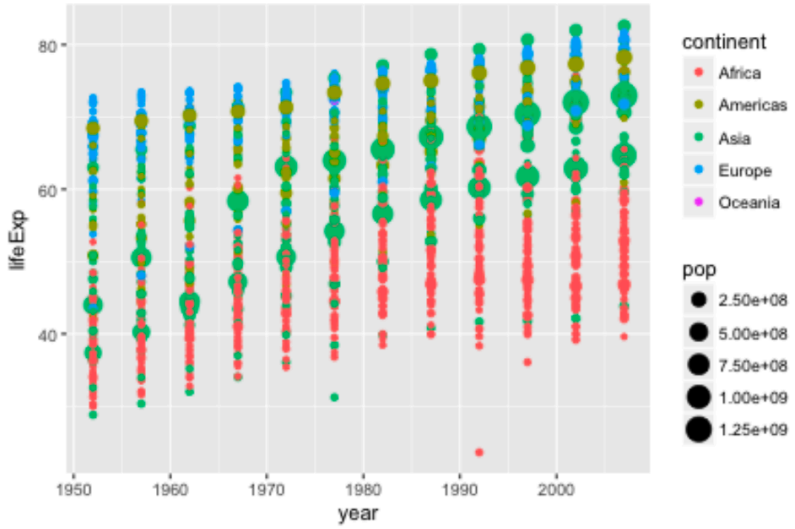


@MonashBioinfo
Bioinformatics.Platform@monash.edu
http://bioinformatics.erc.monash.edu

A better way to explore, present and interpret your data….







Learn to encode data
In a graph

How to get R?

- Many specialised tools/libraries for specific purposes
  - Collection of functions

https://cran.r-project.org)

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

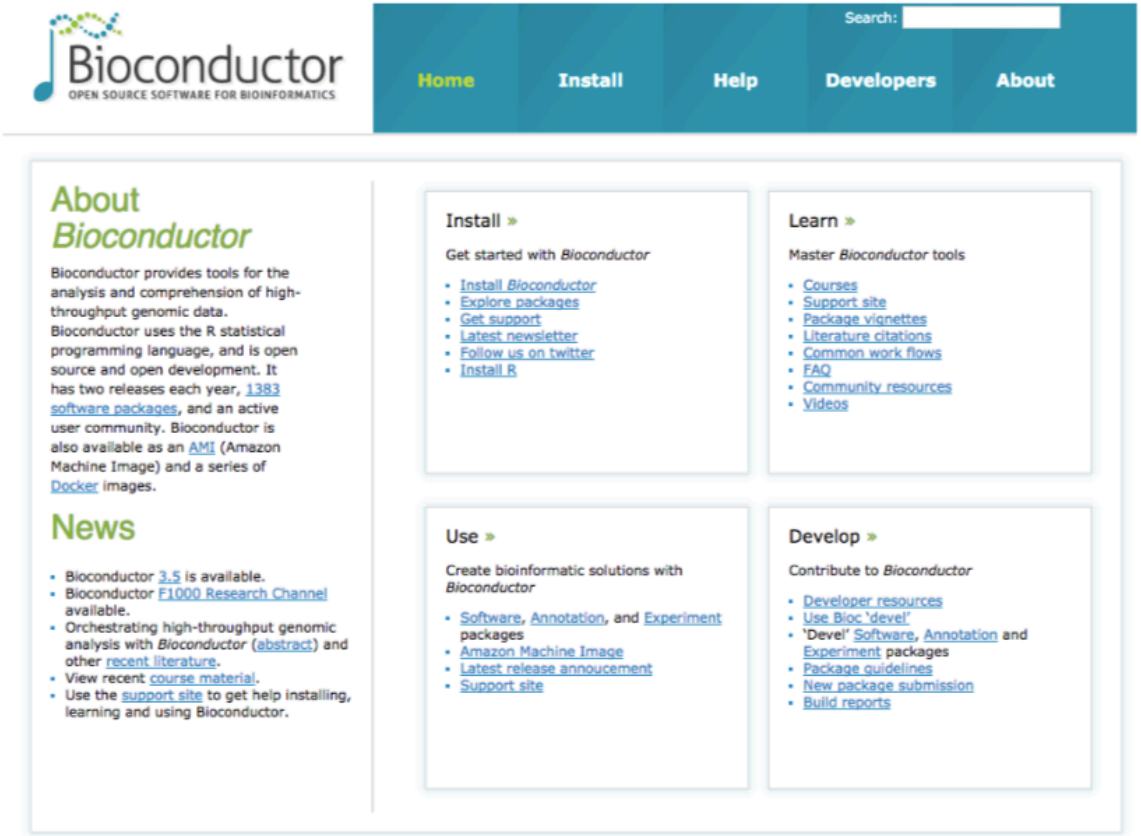
- The latest release (Thursday 2017-09-28, Short Summer) R-3.4.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Tools and libraries used in R?

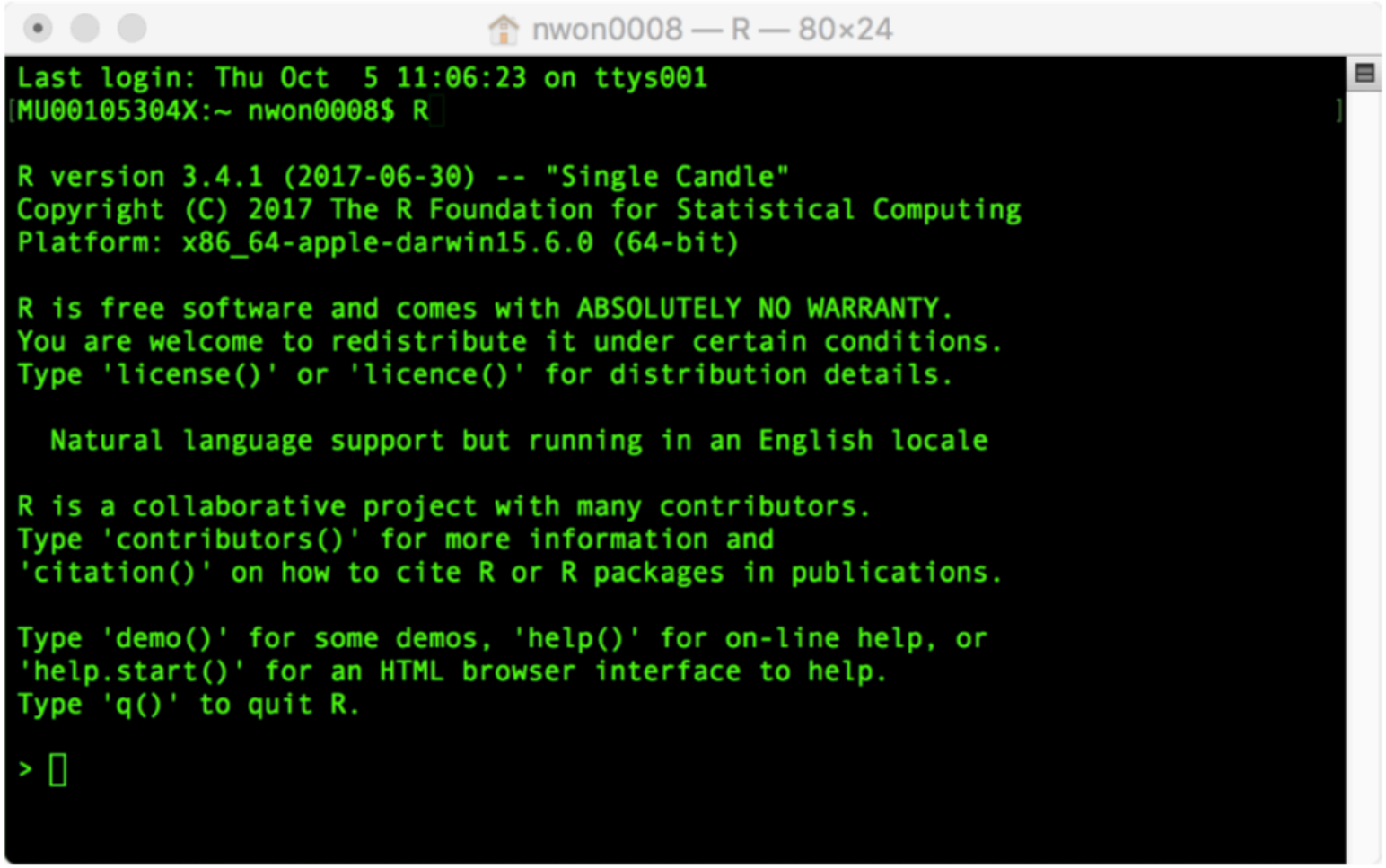
- Bioconductor, source for most bioinformatics libraries and tools for R. Home of LIMMA.



- Defined rules.
- Good documentation.
- Worked examples.

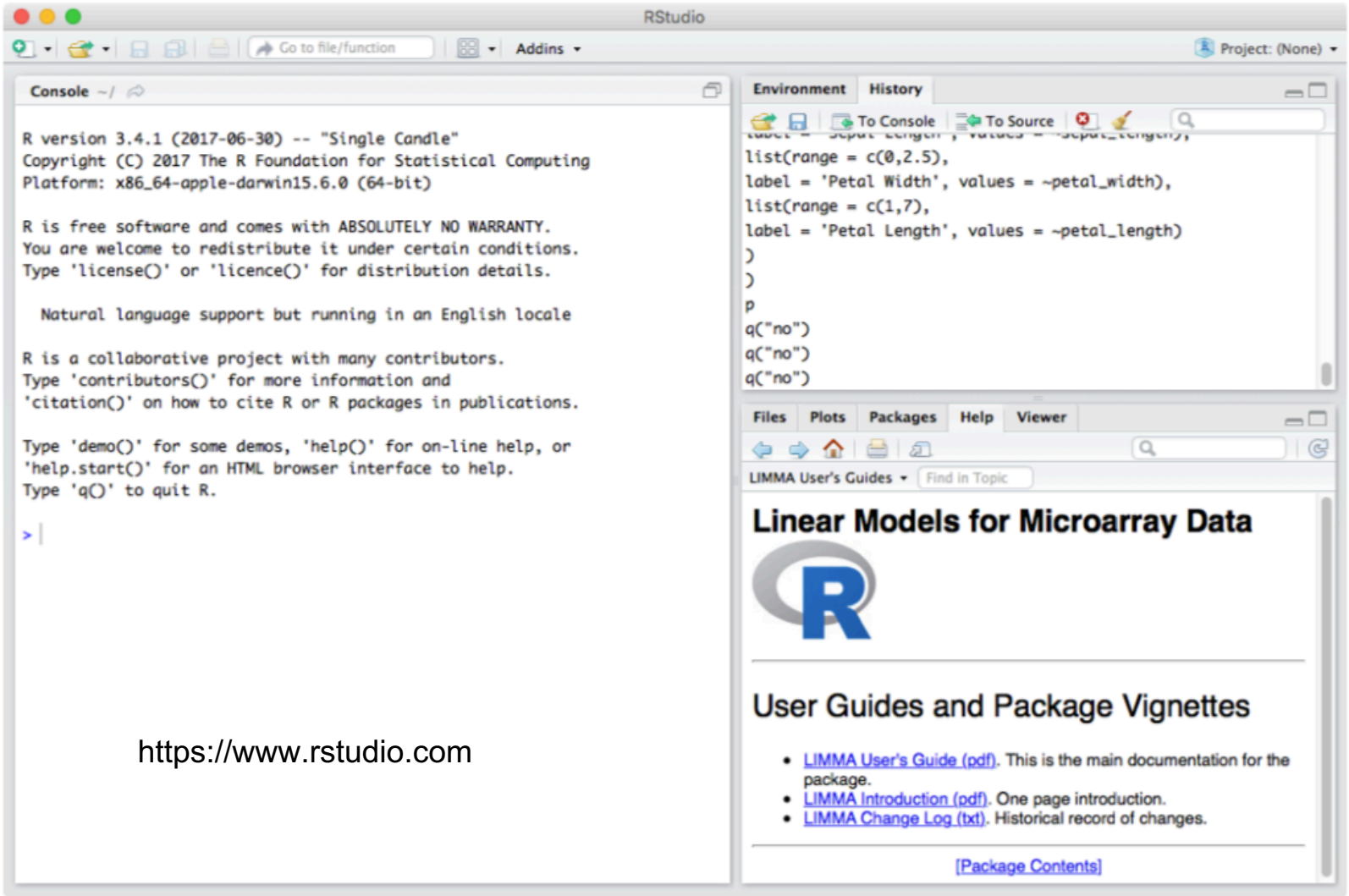
R and the interfaces to R

- R on the command line



R and the interfaces to R

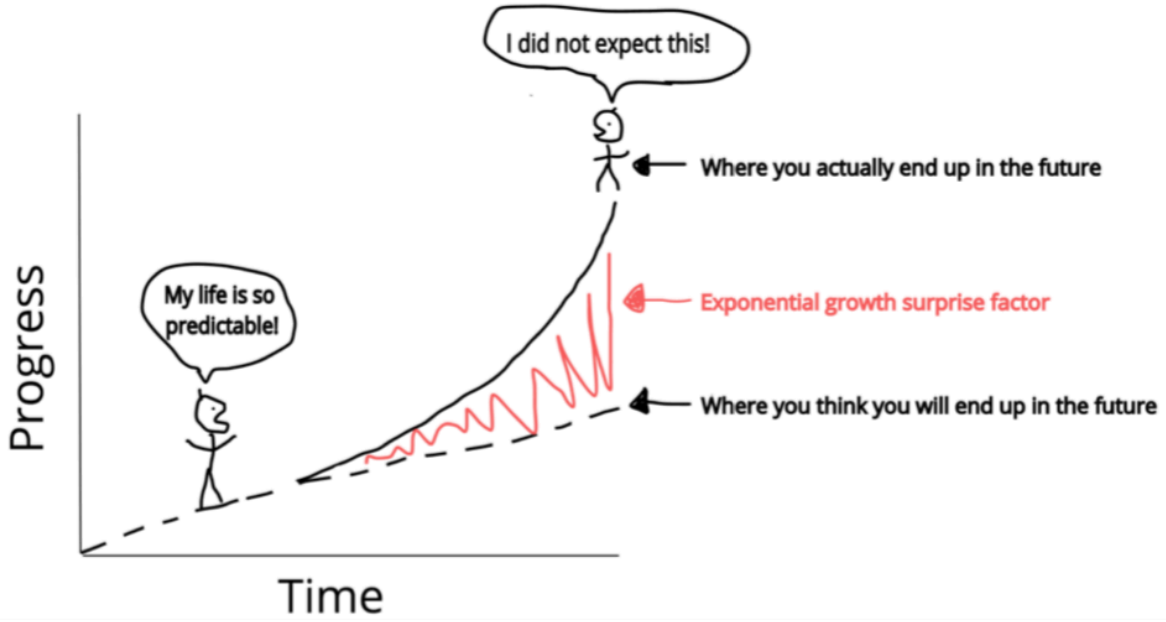
- RStudio – RStudio Server, a graphical interface to R



https://www.rstudio.com

Today's agenda

- Starting out in R
- Working with data in a matrix
- Working with data frames
- Plotting with ggplot2



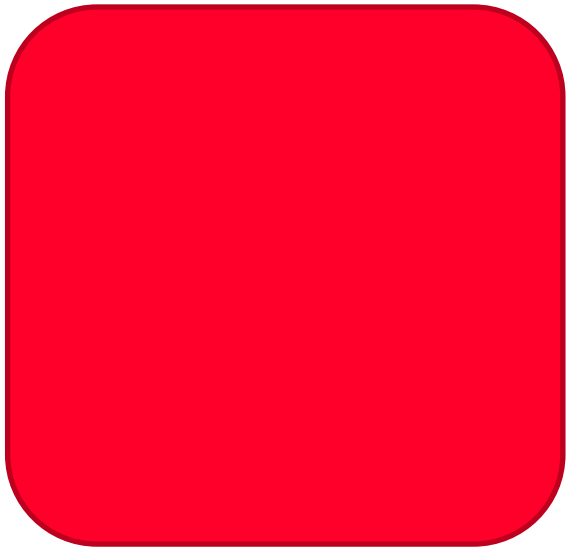
What we won't be covering but are important in the interest of time.

- Things we won't be covering but are of note:
  - –Statistical modelling (lm, glm……).
  - –Packages specific to bioinformatics. (RNA Seq)
  - –Tidyverse and Tidyr (except ggplots)

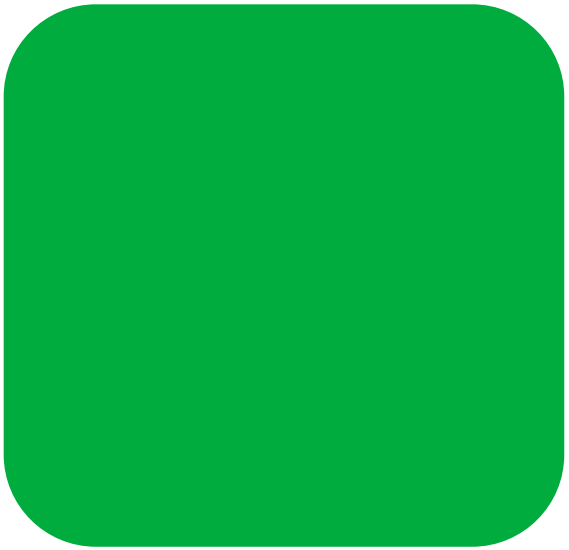
- Advanced R workshops

How today's workshop works – signalling for help

- Workshop notes
- Sticky notes



I need help please



I am all good

Working with data in a matrix

Explanatory text

Loading data

Our example data is quality measurements (particle size) on PVC plastic production, using eight different resin batches, and three different machine operators.

The data set is stored in comma-separated value (CSV) format. Each row is a resin batch, and each column is an operator. In RStudio, open `pvc.csv` and have a look at what it contains.

```
read.csv("r-intro-files/pvc.csv", row.names=1)
```

The code

Tip

The location of the file is given relative to your "working directory". You can see the location of your working directory in the title of the console pane in RStudio. It is most likely "~", indicating your personal home directory. You can change working directory with `setwd`.

The filename "r-intro-files/pvc.csv" means from the current working directory, in the sub-directory "r-intro-files", the file "pvc.csv".

You can check that the file is actually in this location using the "Files" pane in the bottom right corner of RStudio.

The R workbook – Challenges and extra homework

```
avg_operator <- apply(mat, 2, mean)
```

Since the second argument to `apply` is `MARGIN`, the above command is equivalent to `apply(dat, MARGIN = 2, mean)`.

Tip

Some common operations have more concise alternatives. For example, you can calculate the row-wise or column-wise means with `rowMeans` and `colMeans`, respectively.

Challenge - summarizing the matrix

How would you calculate the standard deviation for each resin?

Advanced: How would you calculate the values two standard deviations above and below the mean for each resin?

We will give you challenges through the workshop to work with the example data, also homework

Rounding up

- This is the start of your R journey, many others are at the same stage, share your questions.
- Attend this workshop again, and others we offer
  - Advanced R
  - Specialised use-cases with R (eg: RNASeq)
- MBP have Friday help sessions at Clayton 3:30pm.
  - Alfred to be determined, contact Nick for questions.

Ice breaker activity

- Introduce yourself to your neighbour(s) and tell them why you are here and what you want to achieve.

Basic concepts of R and take home messages

- R is a programming language and your code can be recorded in an R script file.

- Variable
- Assignment operator **<-**
- Functions **name()**
- Subset **[x , y]**

- Run code from R script **command/control <enter>**
- Swap between panes **control 1/2**