MONASH University

MONASH BIOINFORMATICS PLATFORM

Introducing R

Intro to R Workshop – Alfred Campus October 19 2017

Trainers – Adele Barugahare, Haroon Naeem, Nick Wong

Helpers – Kevin Gillinder, Antony Kaspi, Graham Magor, Haloom Rafehi, Mark Ziemann

@MonashBioinfo

Bioinformatics.platform@monash.edu

http://bioinformatics.erc.monash.edu

GROUP OF EIGHT AUSTRALIA

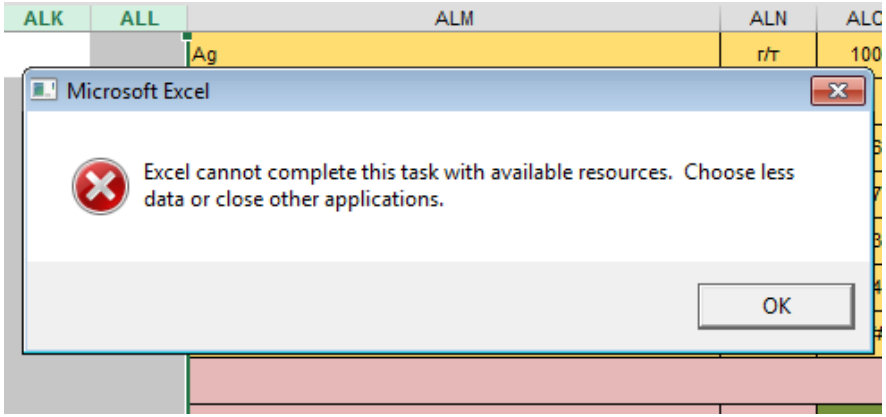
Today's Objective(s) – What we hope you will get out of today

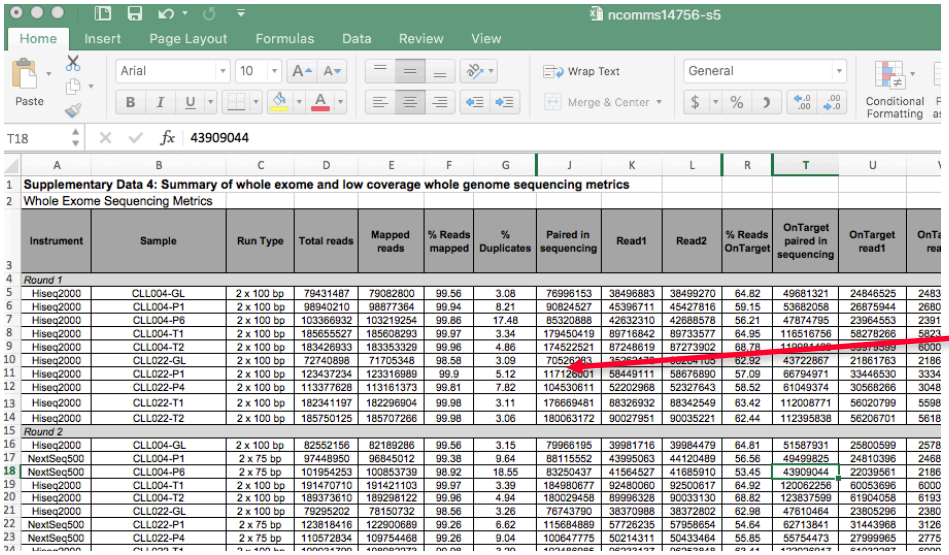
- Introduction to R
- The basics of R – data objects
- Data plotting
- How to find more information for a specific task
- The start of the journey, a network of R learners/users around you.
- Entry into more advanced and specialised uses of R for your research.

Why R? (as opposed to Excel?)

- It's free
- Reproducible analysis
  - Documenting what you have done with your data.
  - Come back to it 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 written in the code.
- 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?)





Instrument	Sample	Run Type	Total reads	Mapped reads	% Reads mapped	% Duplicates	Paired in sequencing	Read1	Read2	% Reads OnTarget	OnTarget paired in sequencing	OnTarget read1	OnTa rea
Round 1													
Hiseq2000	CLL004-GL	2 x 100 bp	79431487	79082800	99.56	3.08	76996153	38496883	38499270	64.82	49681321	24846525	2483
Hiseq2000	CLL004-P1	2 x 100 bp	98940210	98877364	99.94	8.21	90824527	45396711	45427816	59.15	53682058	26875944	2680
Hiseq2000	CLL004-P6	2 x 100 bp	103366932	103219254	99.86	17.48	85320888	42632310	42688578	56.21	47874795	23964553	2391
Hiseq2000	CLL004-T1	2 x 100 bp	185655527	185608293	99.97	3.34	179450419	89716842	89733577	64.95	116516756	58278266	5823
Hiseq2000	CLL004-T2	2 x 100 bp	183426933	183353329	99.96	4.86	174522521	87248619	87273902	68.78	119941667	59979589	6000
Hiseq2000	CLL022-GL	2 x 100 bp	72740898	71705348	98.58	3.09	70526203	35264098	35254105	62.92	43722867	21861763	2186
Hiseq2000	CLL022-P1	2 x 100 bp	123437234	123316989	99.9	5.12	117126001	58449111	58676890	57.09	66794971	33446530	3334
Hiseq2000	CLL022-P4	2 x 100 bp	113377628	113161373	99.81	7.82	104530611	52202968	52327643	58.52	61049374	30568266	3048
Hiseq2000	CLL022-T1	2 x 100 bp	182341197	182296904	99.98	3.11	176669481	88326932	88342549	63.42	112008771	56020799	5598
Hiseq2000	CLL022-T2	2 x 100 bp	185750125	185707266	99.98	3.06	180063172	90027951	90035221	62.44	112395838	56206701	5618
Round 2													
Hiseq2000	CLL004-GL	2 x 100 bp	82552156	82189286	99.56	3.15	79966195	39981716	39984479	64.81	51587931	25800599	2578
NextSeq500	CLL004-P1	2 x 75 bp	97448950	96845012	99.38	9.64	88115552	43995063	44120489	56.56	49499825	24810396	2468
NextSeq500	CLL004-P6	2 x 75 bp	101954253	100853739	98.92	18.55	83250437	41564527	41685910	53.45	43909044	22039561	2186
Hiseq2000	CLL004-T1	2 x 100 bp	191470710	191421103	99.97	3.39	184980677	92480060	92500617	64.92	120062256	60053696	6000
Hiseq2000	CLL004-T2	2 x 100 bp	189373610	189298122	99.96	4.94	180029458	89996328	90033130	68.82	123837599	61904058	6193
Hiseq2000	CLL022-GL	2 x 100 bp	79295202	78150732	98.56	3.26	76743790	38370988	38372802	62.98	47610464	23805296	2380
NextSeq500	CLL022-P1	2 x 75 bp	123818416	122900689	99.26	6.62	115684889	57726235	57958654	54.64	62713841	31443968	3126
NextSeq500	CLL022-P4	2 x 75 bp	110572834	109754468	99.26	9.04	100647775	50214311	50433464	55.85	55754473	27999965	2775

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

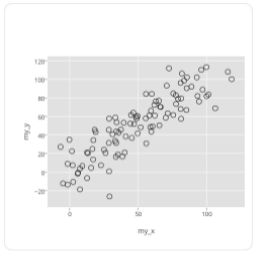
Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

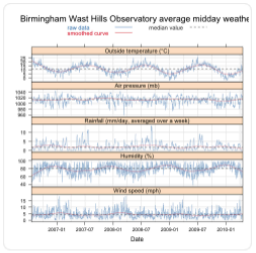
frequently reused. Our aim here is to raise awareness of the problem.

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 *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosoph*

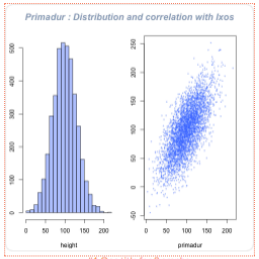
Why R? (as opposed to Excel?)

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

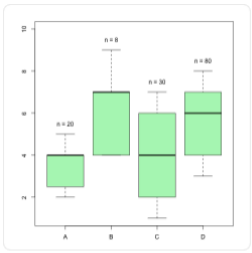

#109 GGplotly : Make any ggplot2 graph interactive !

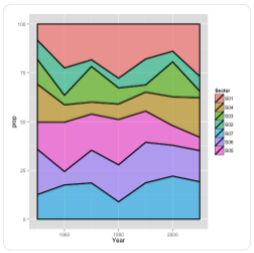


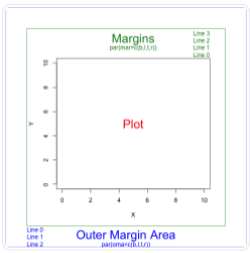
#105 Plotting time series



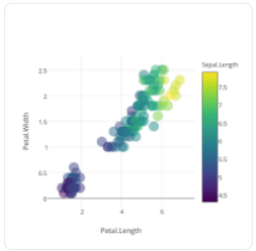
#1 One title for 2 graphs

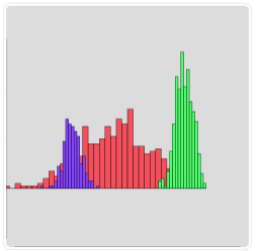

#23 Add text above boxplot

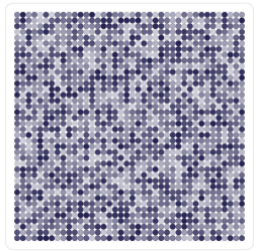

#136 Stacked Area Graph

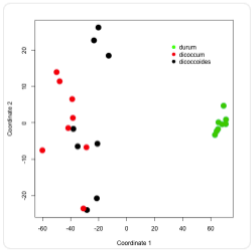


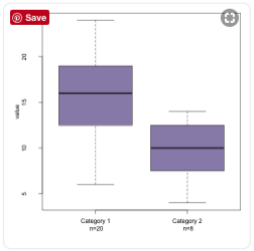
#74 Oma and Margin cheatsheet


#121 Manage colors in plotly (4)

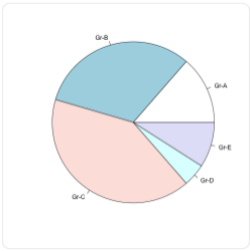

#18 Histogram without axis

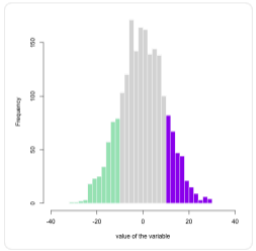

#133 Demineur | dataArt

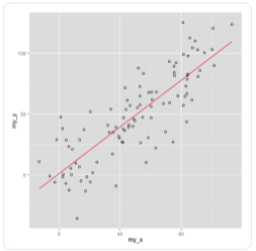

#32 Multi Dimensional Scaling (MDS)

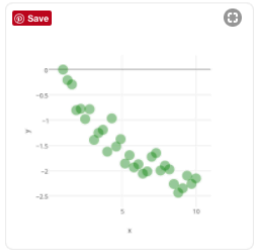


#70 Boxplot with Categories on multiple lines

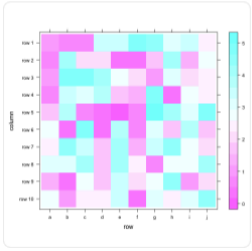

#131 The pie plot function

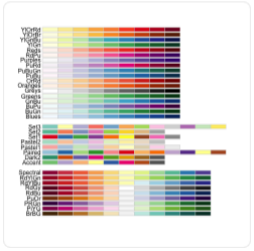

#83 Histogram with colored tail

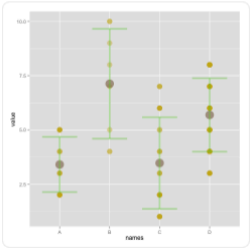

#51 Linear trend ggplot2



#124 Change hovertext in plotly


#28 Turn Y axis upside down in levelplot


#38 RColorBrewer palettes


#47 groups comparison with ggplot2

How to get R?

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

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!

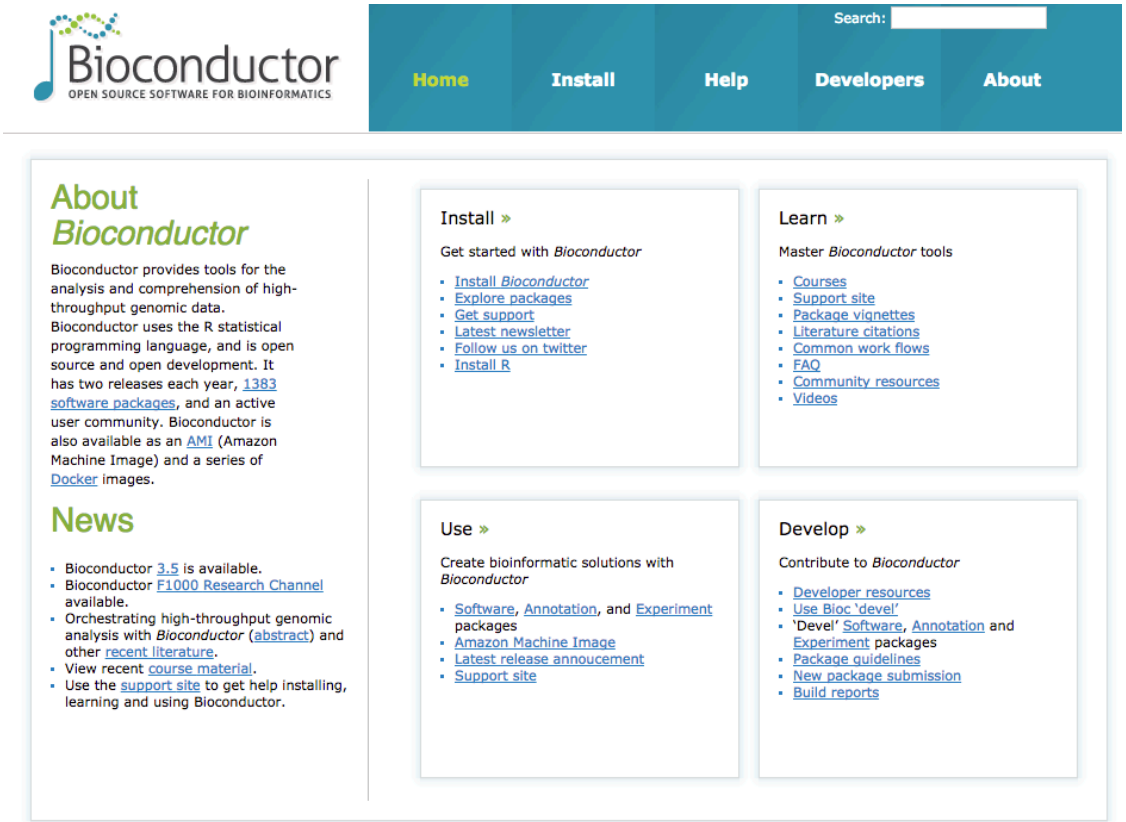
- The latest release (Thursday 2017-09-28, Short Summer) R-3.4.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available 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?

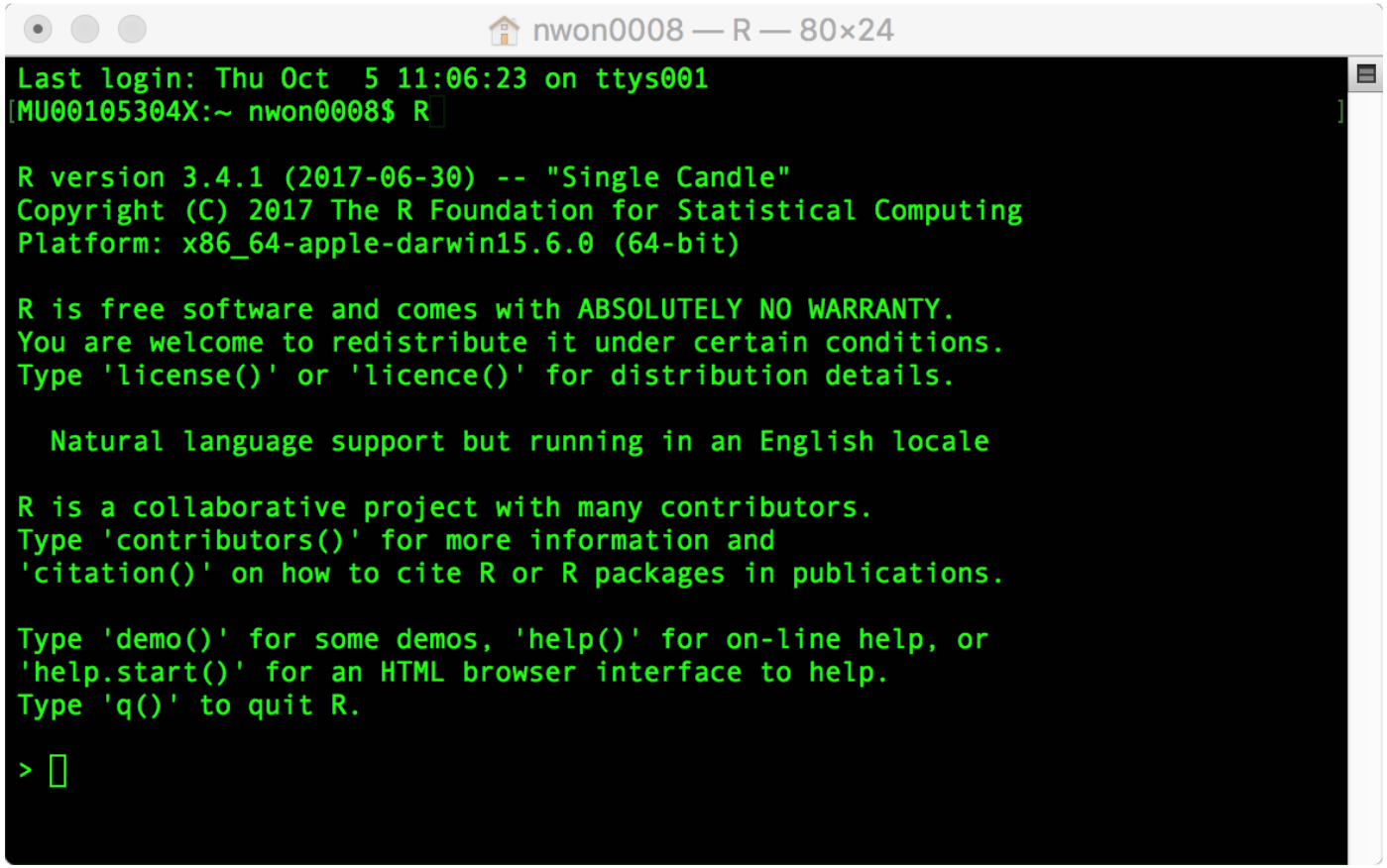
- Bioconductor, source for most bioinformatics libraries and tools for R. Home of LIMMA.



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

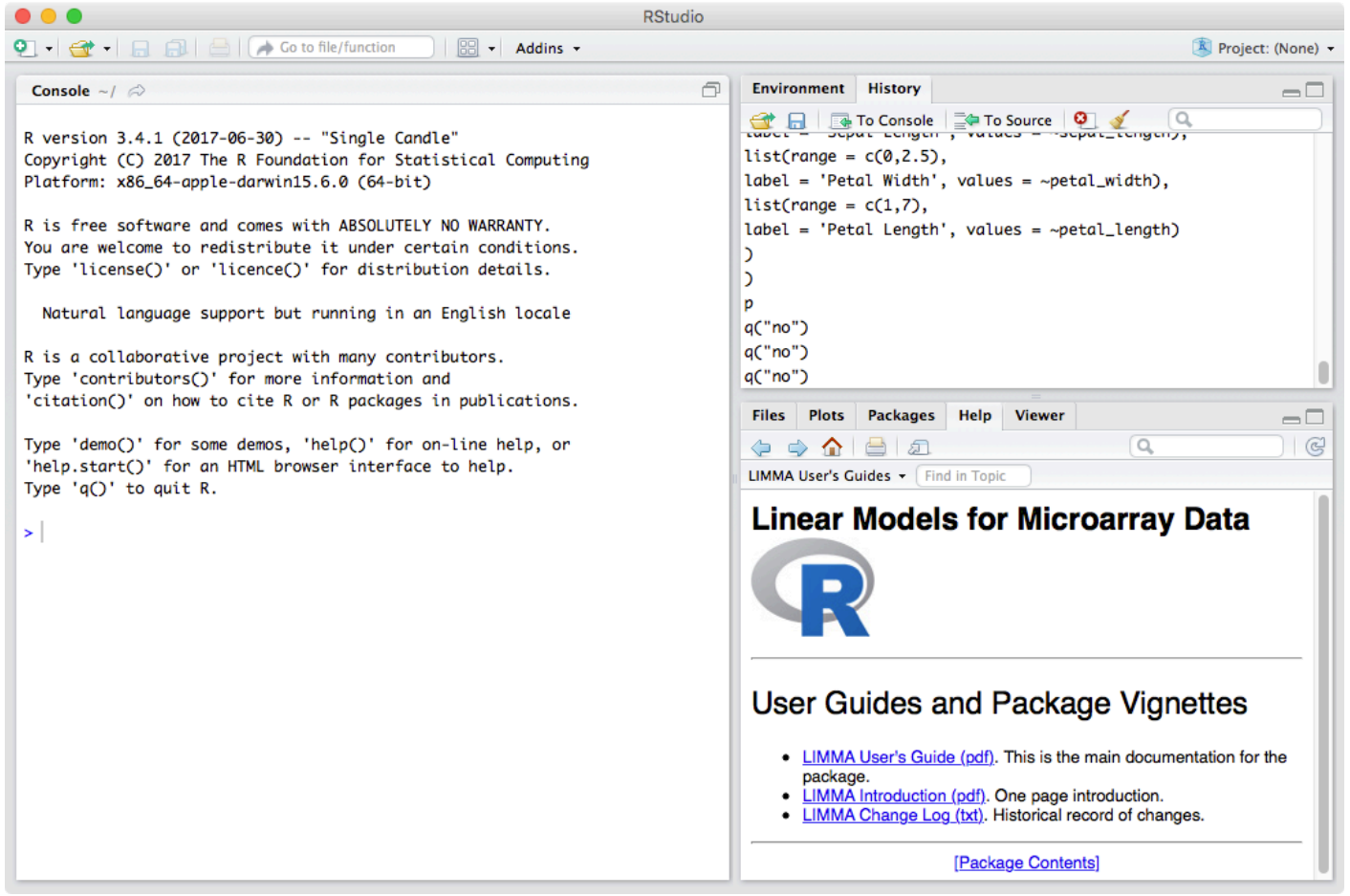
R and the interfaces to R

- R on the command line



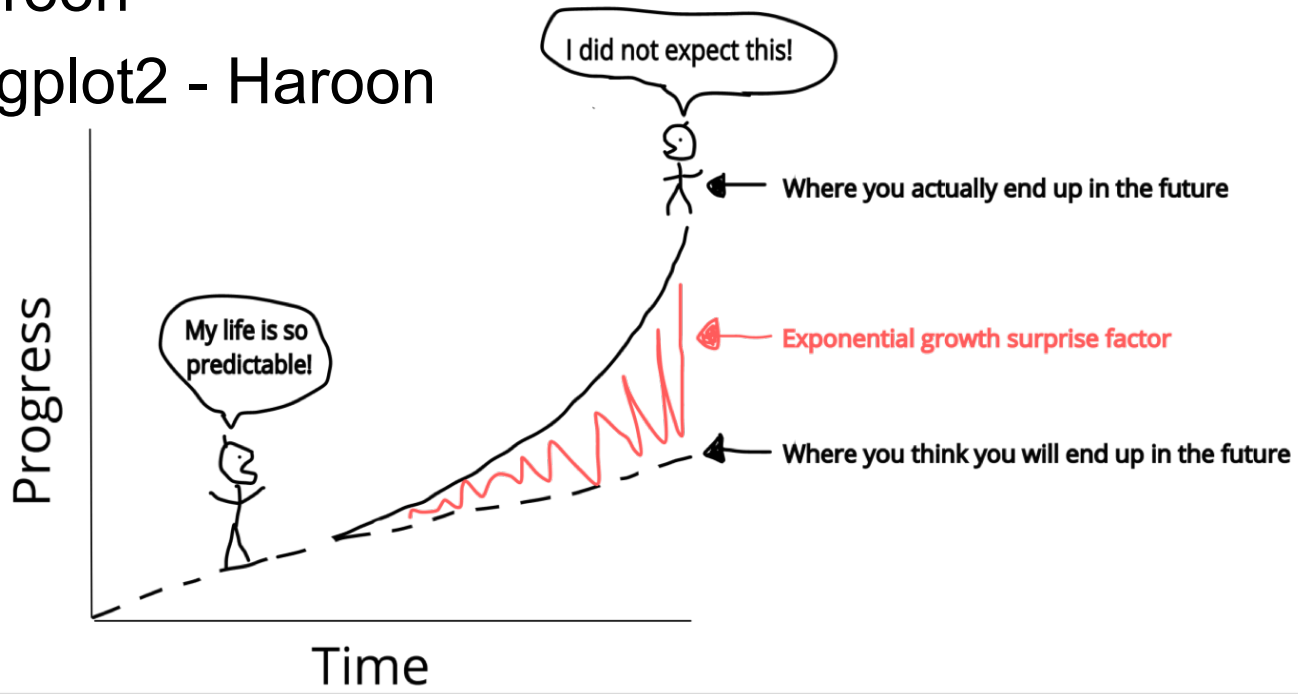
R and the interfaces of R

- RStudio – RStudio Server, a graphical interface to R



Today – Intro to R

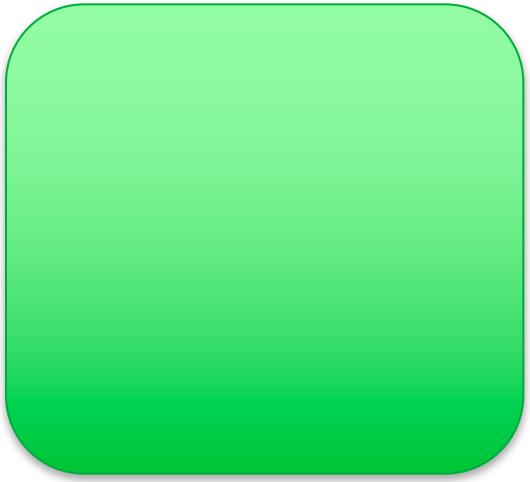
- Starting out in R - Adele
- Working with data in a matrix - Adele
- Working with data in a data frame - Nick
- For loops - Haroon
- Plotting with ggplot2 - Haroon



How the workshop works – signaling for help

Sticky notes

I need help!



I am all good.

Google Docs for communication and submitting answer to challenges

Rounding up

- This is the start of your R journey, many others are at the same stage, share your questions.
- Encourage to attend this workshop again and others we offer.
  - Advanced R
  - Specialised use cases with R (eg: RNA-Seq)
- MBP offer a Friday help session at Clayton @3:30pm for drop in questions. To be determined for Alfred. (or, contact Nick).
- Vamos!