

Introducing R

Intro to R Workshop

Trainers – Dr Paul Harrison & Adele Barugahare

Helpers – Dr Sarah Williams & Dr Nick Wong



@MonashBioinfo



Bioinformatics.platform@monash.edu



<http://bioinformatics.erc.monash.edu>

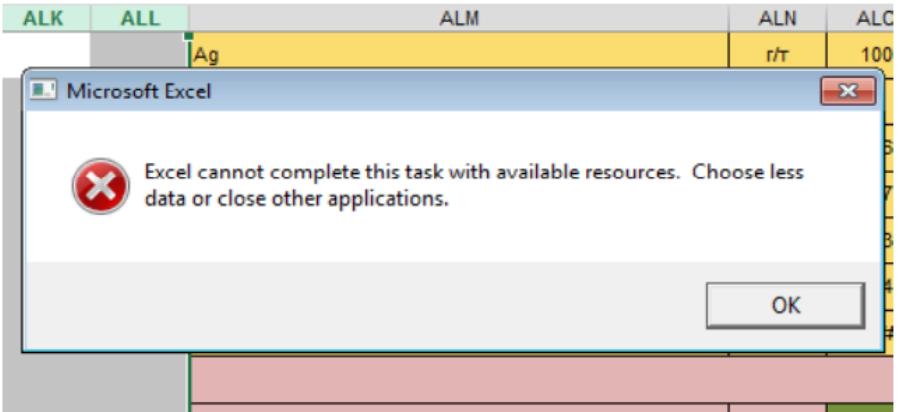
Today's Objective(s) – What we hope you will get out of today

- 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 specialized 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?)



ncomms14756-s5													
1 Supplementary Data 4: Summary of whole exome and low coverage whole genome sequencing metrics													
2 Whole Exome Sequencing Metrics													
Instrument	Sample	Run Type	Total reads	Mapped reads	% Reads mapped	% Duplicates	Paired in sequencing	Read1	Read2	% Reads OnTarget	OnTarget paired in sequencing	OnTarget read1	OnTarget read2
4 Round 1													
5 Hiseq2000	CLL004-GL	2 x 100 bp	79431487	79082800	99.56	3.08	76996153	38496883	38499270	64.82	49681321	2483	
6 Hiseq2000	CLL004-P1	2 x 100 bp	98940210	98877364	99.94	8.21	90824527	45396711	45427816	59.15	53682058	26875944	2680
7 Hiseq2000	CLL004-P6	2 x 100 bp	103366932	103219254	99.86	17.48	85320888	42632310	42668578	56.21	47874795	23964553	2391
8 Hiseq2000	CLL004-T1	2 x 100 bp	185655257	185682932	99.97	3.34	179450419	89716842	89733577	64.95	116516756	58278266	5821
9 Hiseq2000	CLL004-T2	2 x 100 bp	183426933	183333239	99.96	4.65	174522521	87246019	87273902	66.78	116516756	58278266	5820
10 Hiseq2000	CLL022-GL	2 x 100 bp	72740898	71705348	98.58	3.09	70524813	35206470	35206470	62.92	43722867	21861763	2186
11 Hiseq2000	CLL022-P1	2 x 100 bp	123437234	123316869	99.9	5.12	117126001	58449111	58676890	57.09	66794971	33446530	3334
12 Hiseq2000	CLL022-P4	2 x 100 bp	113377628	113161373	99.81	7.82	104530611	52202968	52327643	58.52	61049374	30568266	3048
13 Hiseq2000	CLL022-T1	2 x 100 bp	182341197	182296904	99.98	3.11	176669481	88326932	88342549	63.42	112008771	56020799	5598
14 Hiseq2000	CLL022-T2	2 x 100 bp	185750125	185707266	99.98	3.06	180063172	90027951	90035221	62.44	112395938	56206701	5618
15 Round 2													
16 Hiseq2000	CLL004-GL	2 x 100 bp	82525156	82189286	99.56	3.15	79966195	39981716	39984479	64.81	51587931	25800599	2578
17 Hiseq2000	CLL004-P1	2 x 75 bp	97448950	96845012	99.38	9.64	88115552	43995063	44120489	56.56	49499825	24810396	2468
18 NextSeq500	CLL004-P6	2 x 75 bp	101954250	100853739	98.92	18.55	83250437	41586527	41685910	53.45	43909044	22039561	2186
19 Hiseq2000	CLL004-T1	2 x 100 bp	191470710	19121103	99.97	3.39	184980677	92480067	92500617	64.92	120062256	60053696	6000
20 Hiseq2000	CLL004-T2	2 x 100 bp	189373610	189298122	99.96	4.94	180029458	89996328	90033130	68.82	123886599	61904058	6193
21 Hiseq2000	CLL022-GL	2 x 100 bp	79295202	7815932	98.56	3.26	78745090	38370968	38372802	62.98	47670464	23605296	2360
22 NextSeq500	CLL022-P1	2 x 75 bp	123816415	122900689	99.26	6.02	115684693	57726235	57958684	54.64	61049374	31443968	3126
23 NextSeq500	CLL022-P4	2 x 75 bp	110872834	109754466	99.26	9.04	100647775	90214311	90433484	55.85	55734473	27999965	2775

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 Ziemann¹, 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

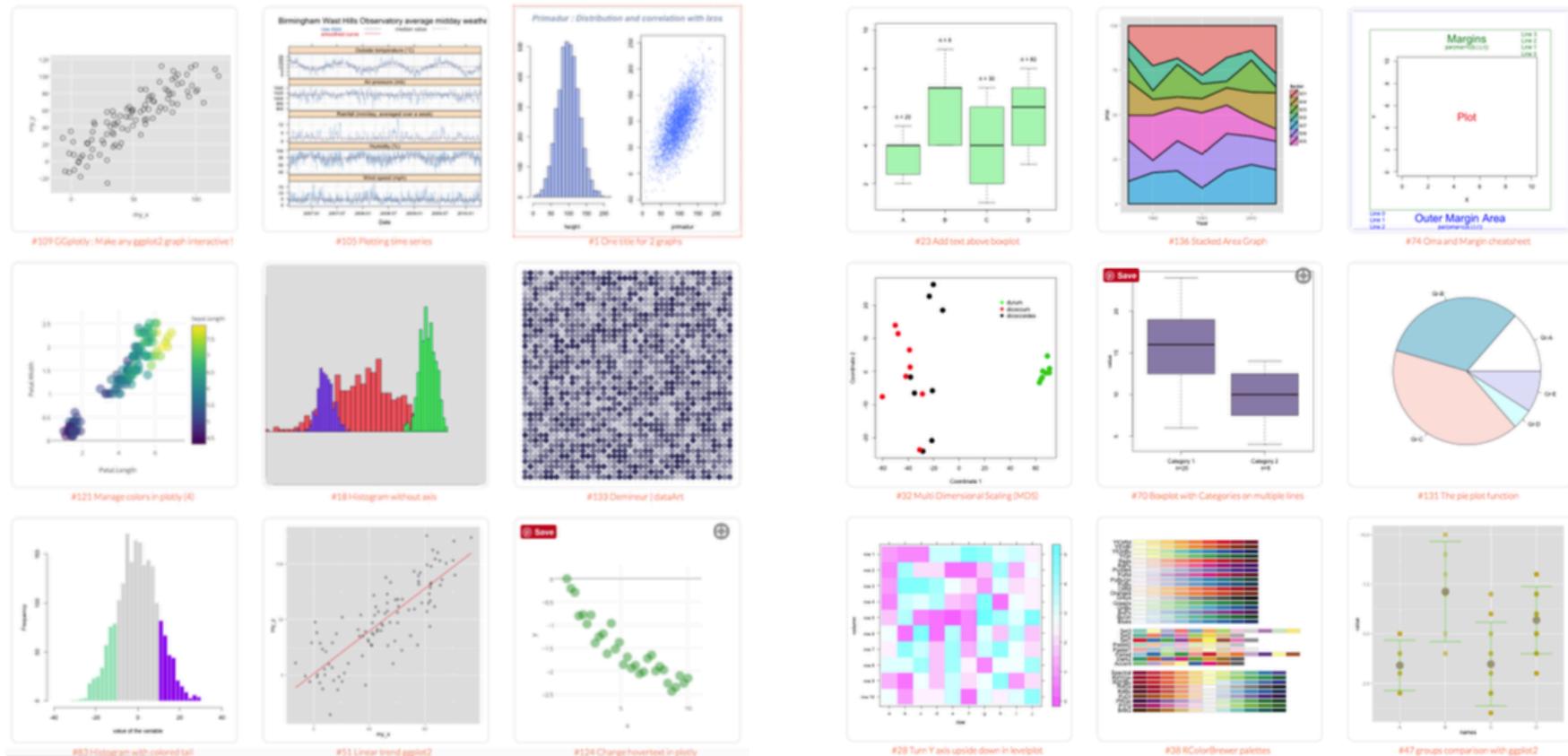
frequently reused. Our aim here is to raise awareness of the problem.

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 <http://www.ensembl.org>.

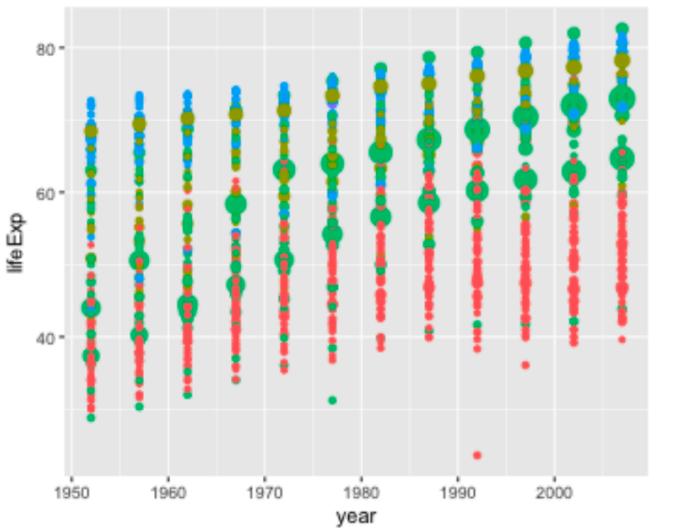
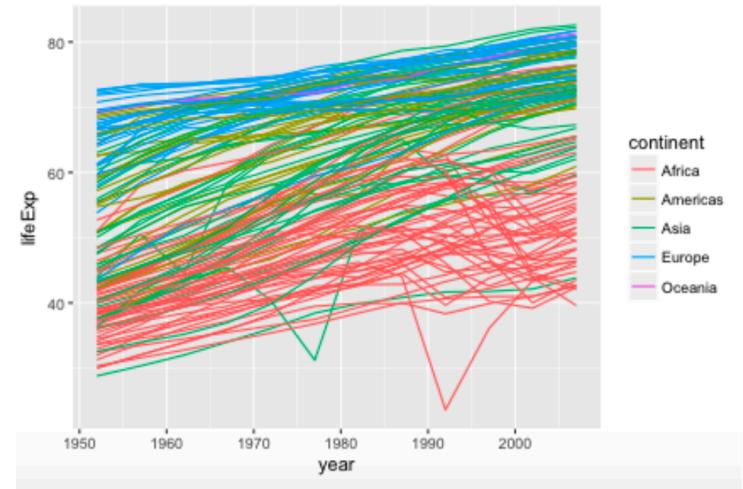
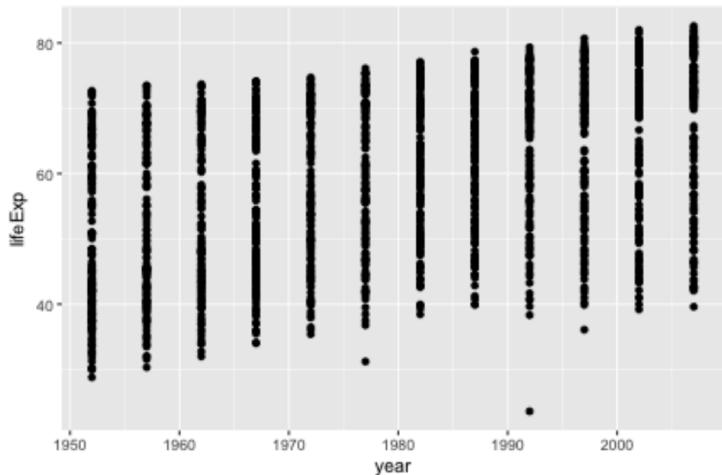
Typo and saved it

Why R? (as opposed to Excel?)

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



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



continent
Africa
Americas
Asia
Europe
Oceania

pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09

Learn to encode data
In a graph

How to get R?



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)

The Comprehensive R Archive Network

Download and Install R

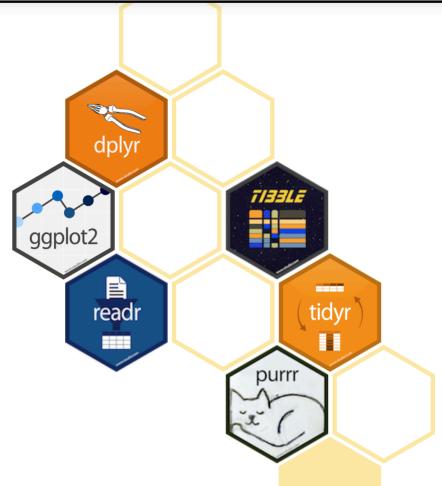
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

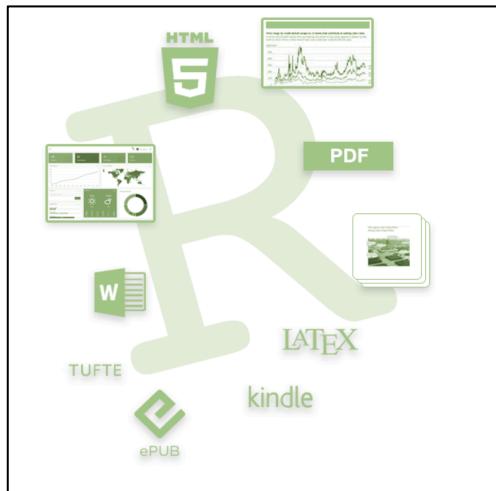
<https://cran.r-project.org>

Many specialized tools/libraries for specific purposes

Tidyverse: Data Science



Rmarkdown: Documents



Bioconductor: Bioinformatics



R and the interfaces to R

▪ R on the command line

```
nwon0008 — R — 80x24
Last login: Thu Oct  5 11:06:23 on ttys001
[MU00105304X:~ nwon0008$ R]

R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

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.

> ]
```

R and the interfaces to R

▪ RStudio – RStudio Server, a graphical interface to R

The screenshot shows the RStudio interface. On the left, the Console tab displays the standard R startup message. On the right, the Environment tab shows R code related to the Iris dataset, specifically setting ranges for Petal Width and Petal Length. Below the code, the Viewer tab displays a page titled "Linear Models for Microarray Data" from the LIMMA package, featuring the R logo and navigation links for user guides and package vignettes.

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

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.

> |
```

list(range = c(0,2.5),
label = 'Petal Width', values = ~petal_width),
list(range = c(1,7),
label = 'Petal Length', values = ~petal_length)
)
)
p
q("no")
q("no")
q("no")

Linear Models for Microarray Data

User Guides and Package Vignettes

- [LIMMA User's Guide \(pdf\)](#). This is the main documentation for the package.
- [LIMMA Introduction \(pdf\)](#). One page introduction.
- [LIMMA Change Log \(txt\)](#). Historical record of changes.

[Package Contents]

Today's agenda

- Starting out in R
- Working with tidy 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:

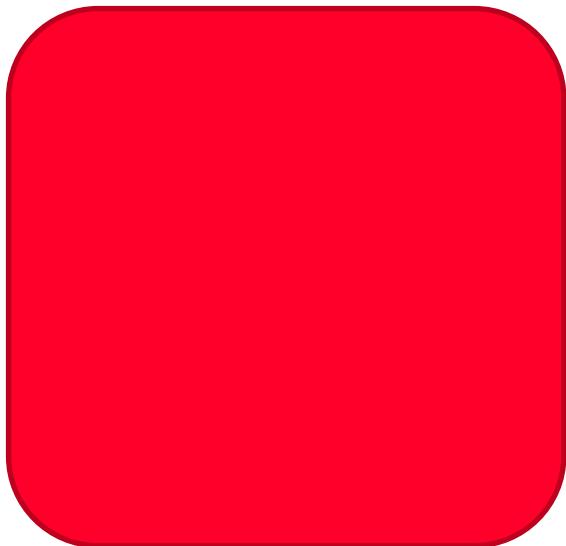
- Base R
- Writing your own functions & loops
- Statistical modelling (lm, glm.....).

- Advanced R workshops

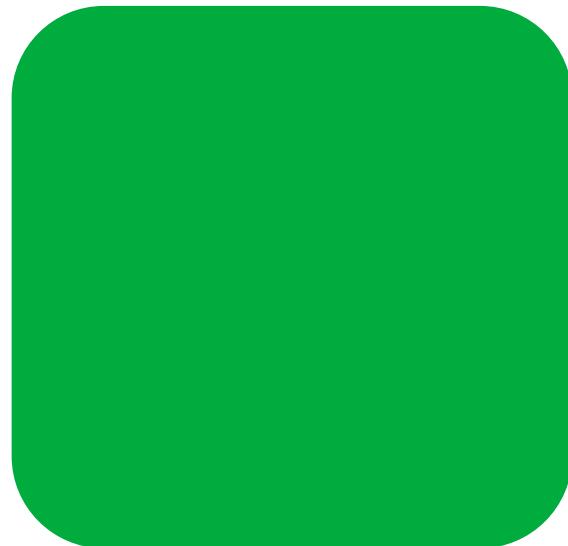


How today's workshop works – signalling for help

- Workshop notes
- Etherpad
- Sticky notes



I need help please



I am all good

The R workbook

Explanatory text

Under the hood, a data frame is a list of column vectors. This is why `mode` told us `gap` was a list. This means we can use `$` to retrieve columns. (Occasionally it is also useful to use `[[]]` to retrieve columns, for example if the column name we want is stored in a variable.)

```
head( gap$lifeExp )
```

```
## [1] 28.801 30.332 31.997 34.020 36.088 38.438
```

```
head( gap[["lifeExp"]] )
```

```
## [1] 28.801 30.332 31.997 34.020 36.088 38.438
```

The code

So to get just the `lifeExp` value of the third row as above, but unwrapped, we can use:

The R workbook – Challenges and tips

Challenge

What is the average gdpPercap for each continent in 2007?

Advanced: Use `weighted.mean` to calculate this correctly.



We will give you challenges and tips through the workshop to work with the example data



Tip

A data frame has column names (`colnames`), and base R data frames can also have row names (`rownames`). However the modern convention, which the Tidyverse enforces, is for a data frame to use column names but not row names. Typically a data frame contains a collection of items (rows), each having various properties (columns). If an item has an identifier such as a unique name, this would be given as just another column.

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)
- Friday help sessions at Clayton 3:00pm.

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 `functionName()`
- Subset `[x , y]`
- Run code from R script `command/control <enter>`
- Swap between panes `control 1/2`