

Introducing R

Intro to R Workshop

Trainers –

Helpers –



@MonashBioinfo



Bioinformatics.platform@monash.edu



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

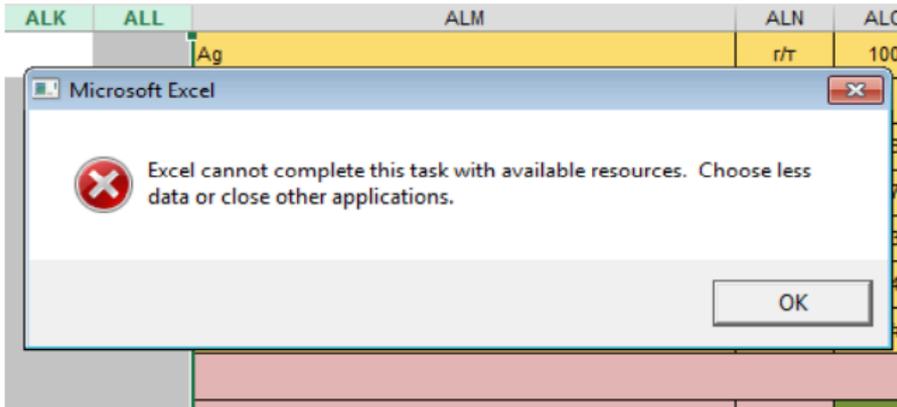
Today's Objective(s) – What we hope you will get out of today

- 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 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?)



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
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-P8	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	89718842	89733577	64.95	116516756	58278266	5833
Hiseq2000	CLL004-T2	2 x 100 bp	183420933	183353329	99.95	4.85	174522521	87249519	87272952	69.18	109998999	59998999	6005
Hiseq2000	CLL022-GL	2 x 100 bp	72742898	71705349	98.58	3.09	70526963	34968488	34981168	62.52	43722987	21881763	2188
Hiseq2000	CLL022-P1	2 x 100 bp	123437234	123318989	99.9	5.12	117126301	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	185702666	99.98	3.06	180063172	90027951	90035221	62.44	112395638	56206701	5618
Round 2													
Hiseq2000	CLL004-GL	2 x 100 bp	82552156	82198296	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	49499829	24810396	2468
NextSeq500	CLL004-P8	2 x 75 bp	101954253	100853739	98.92	18.55	83250437	41564527	41685910	53.45	43909044	22039561	2198
Hiseq2000	CLL004-T1	2 x 100 bp	191470710	191421103	99.97	3.39	184980677	92480060	92500617	64.92	120062296	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	76295202	76150732	99.56	3.26	76743790	38370988	38372802	62.98	47810464	23805296	2380
NextSeq500	CLL022-P1	2 x 75 bp	123818416	122950689	99.26	6.62	119694989	57726235	57958654	54.64	62713841	31443965	3128
NextSeq500	CLL022-P4	2 x 75 bp	110572834	109754468	99.26	9.04	109647775	50214311	50433464	56.85	58754473	27699565	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.

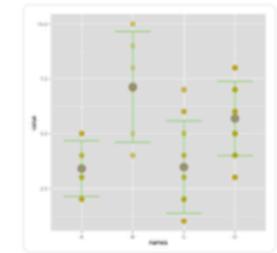
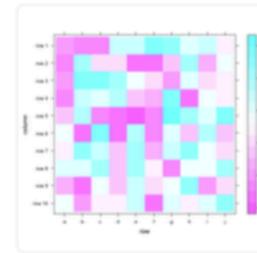
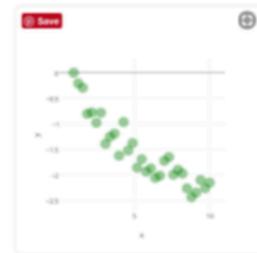
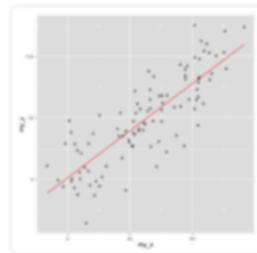
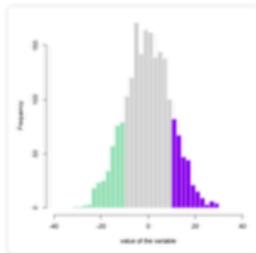
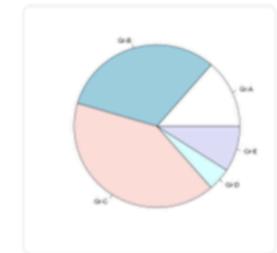
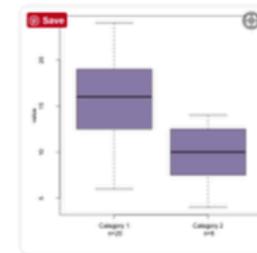
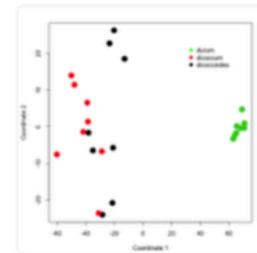
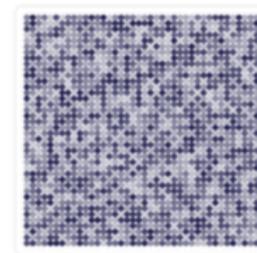
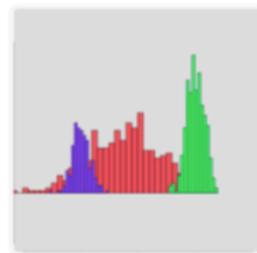
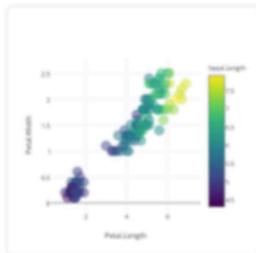
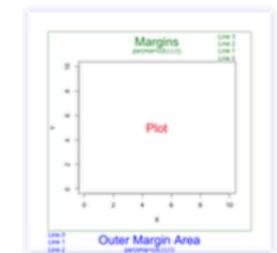
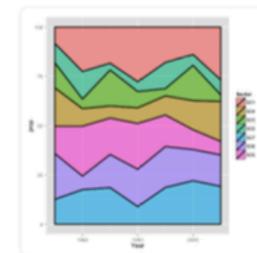
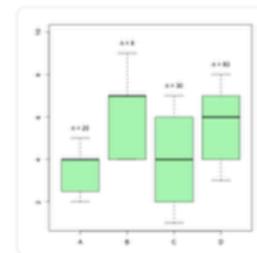
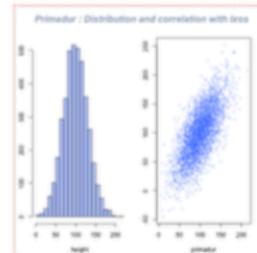
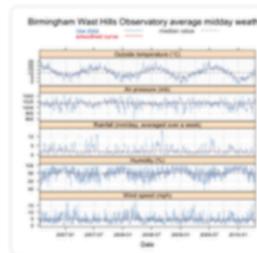
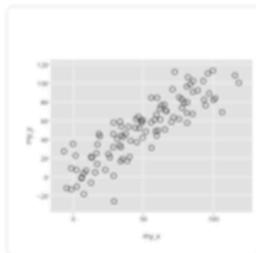
We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (tsv) with sconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv 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 *Aspidosiphonia*, *Caenorhabditis elegans*, *Drosophila*

Typo and saved it

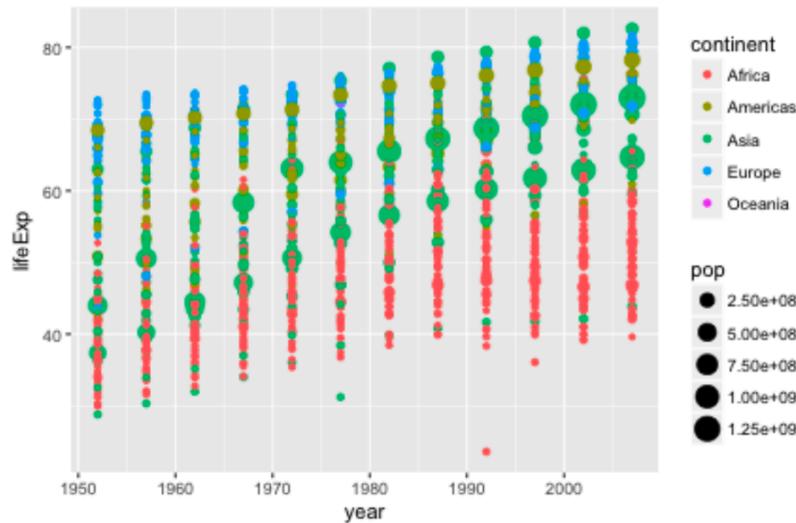
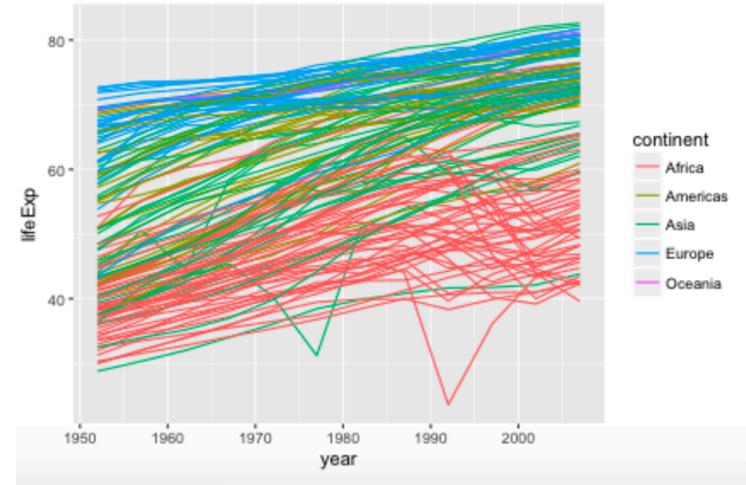
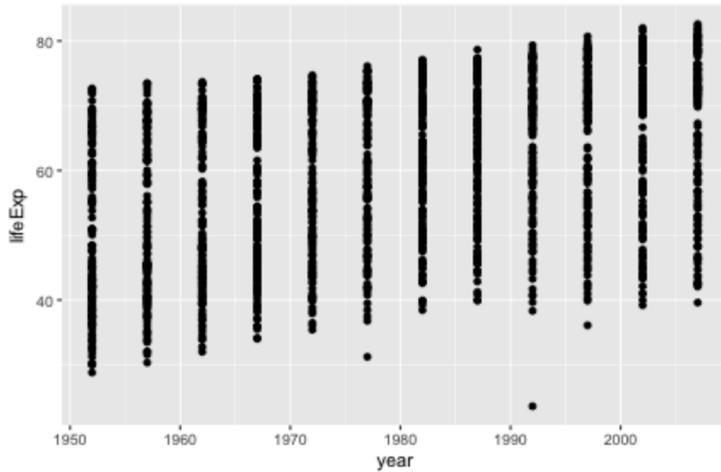


Why R? (as opposed to Excel?)

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



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



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!

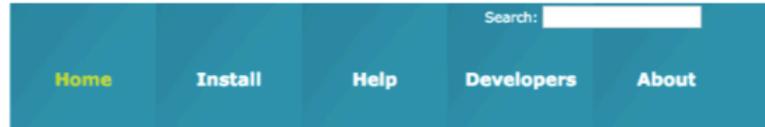
- The latest release (Thursday 2017-09-28, Short Summer) [R-3.4.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions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



About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1383 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

- Bioconductor [3.5](#) is available.
- Bioconductor [F1000 Research Channel](#) available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* ([abstract](#)) and other [recent literature](#).
- View recent [course material](#).
- Use the [support site](#) to get help installing, learning and using Bioconductor.

Install >

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn >

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use >

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment](#) packages
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

Develop >

Contribute to *Bioconductor*

- [Developer resources](#)
- [Use Bioc 'devel'](#)
- ['Devel' Software](#), [Annotation](#) and [Experiment](#) packages
- [Package guidelines](#)
- [New package submission](#)
- [Build reports](#)

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

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

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.

```
> |
```

Environment History

```
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")
```

Files Plots Packages Help Viewer

LIMMA User's Guides Find in Topic

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.

[\[Package Contents\]](#)

<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.

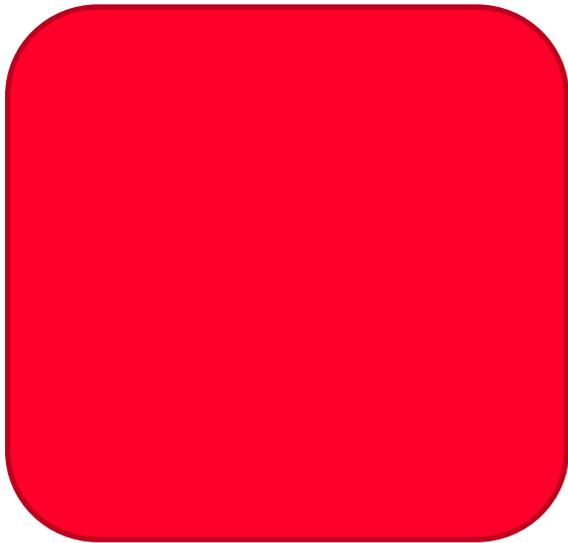
- Things we won't be covering but are of note:
 - Statistical modelling (lm, glm.....).
 - Packages specific to bioinformatics. (RNA Seq)
 - Tidyverse and TidyR (except ggplots)

- Advanced R workshops

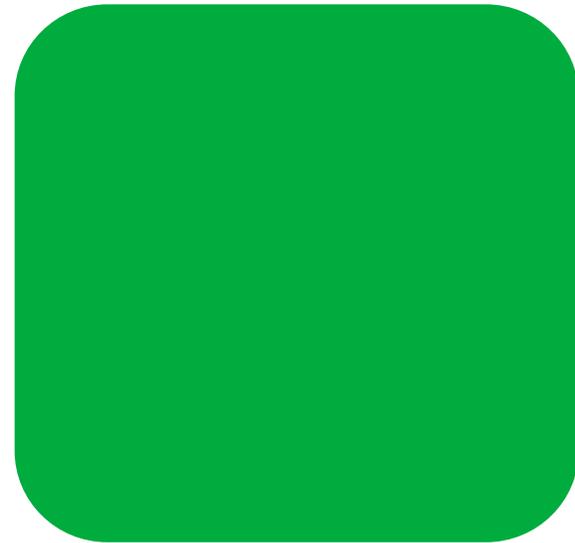


How today's workshop works – signalling for help

- Workshop notes
- Sticky notes



I need help please



I am all good

Working with data in a matrix

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)
```

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.

Explanatory text



The code



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