MONASH
University

Introducing R

Presented by Haroon Naeem, Ph.D.
Bioinformatician
Monash Bioinformatics Platform

30.11.15

MONASH University

Introduce R well enough to understand some basics and feel comfortable trying things on their own

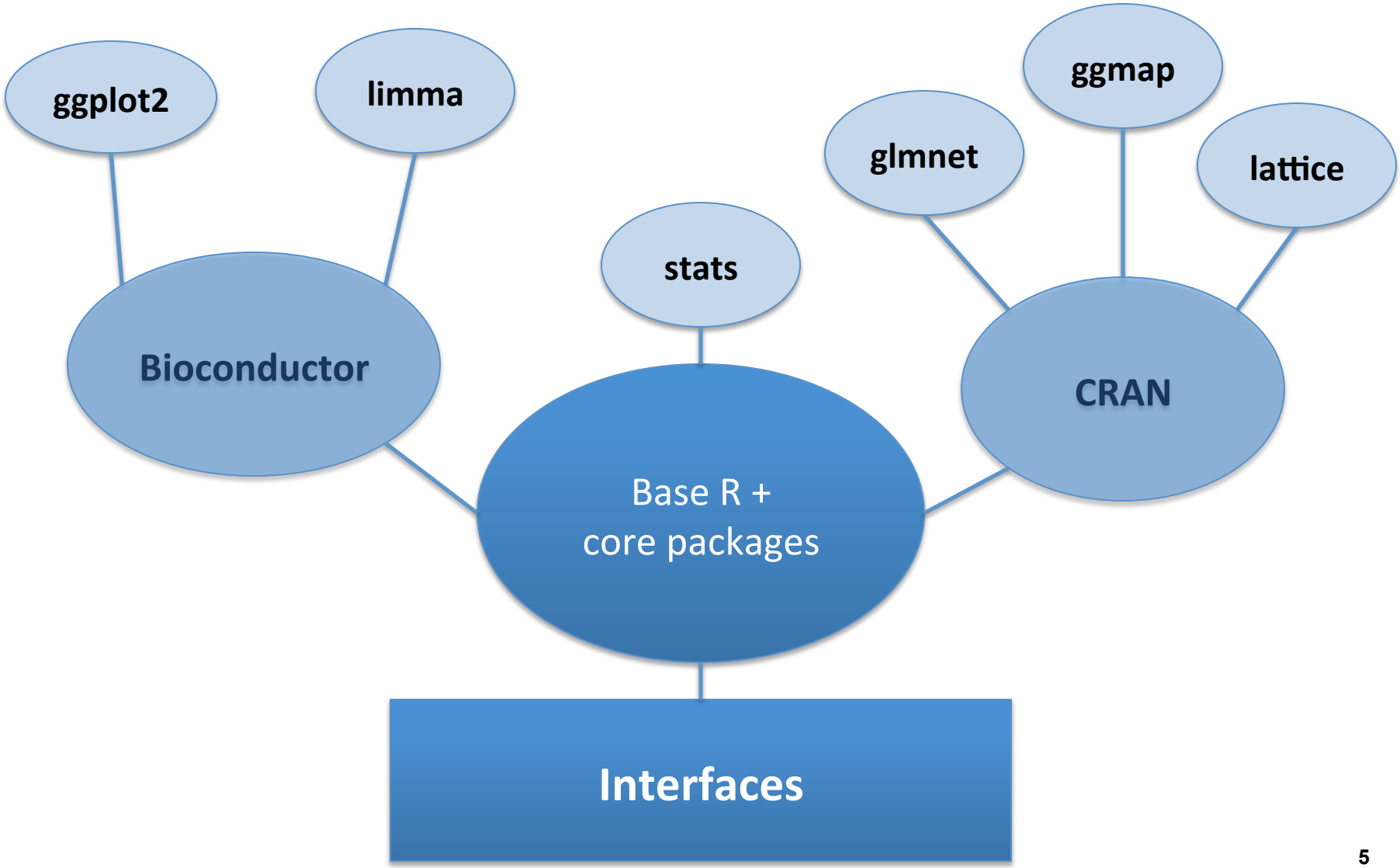
Overview

- **R**
  - A programming environment for statistical computing and graphics, and for rapid development of new tools
  - https://cran.r-project.org/
  - Open source, cross platform
  - Mainly command-line driven
  - software packages written in R for bioinformatics application
  - Currently more than 6000 packages

Overview

- **Advantages**
  - Cross platform (Linux, windows, Mac)
  - Covers various phases of data analysis in a single environment (for microarray analysis, NGS analysis)
  - Easy to incorporate new methods and functions
  - Algorithms have undergone evaluation by statisticians and researchers
  - Comprehensive manuals, notes and course materials

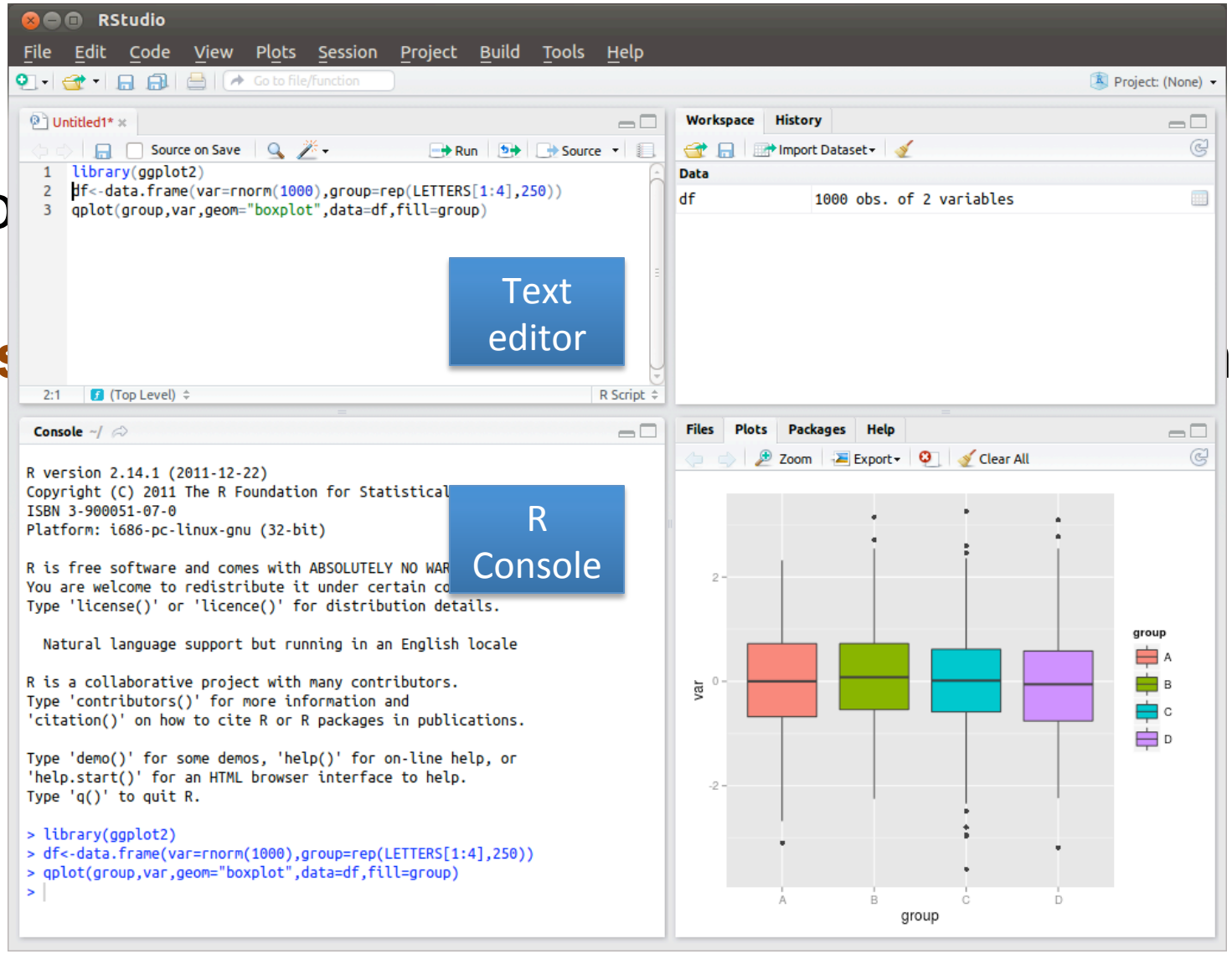
Interacting with R



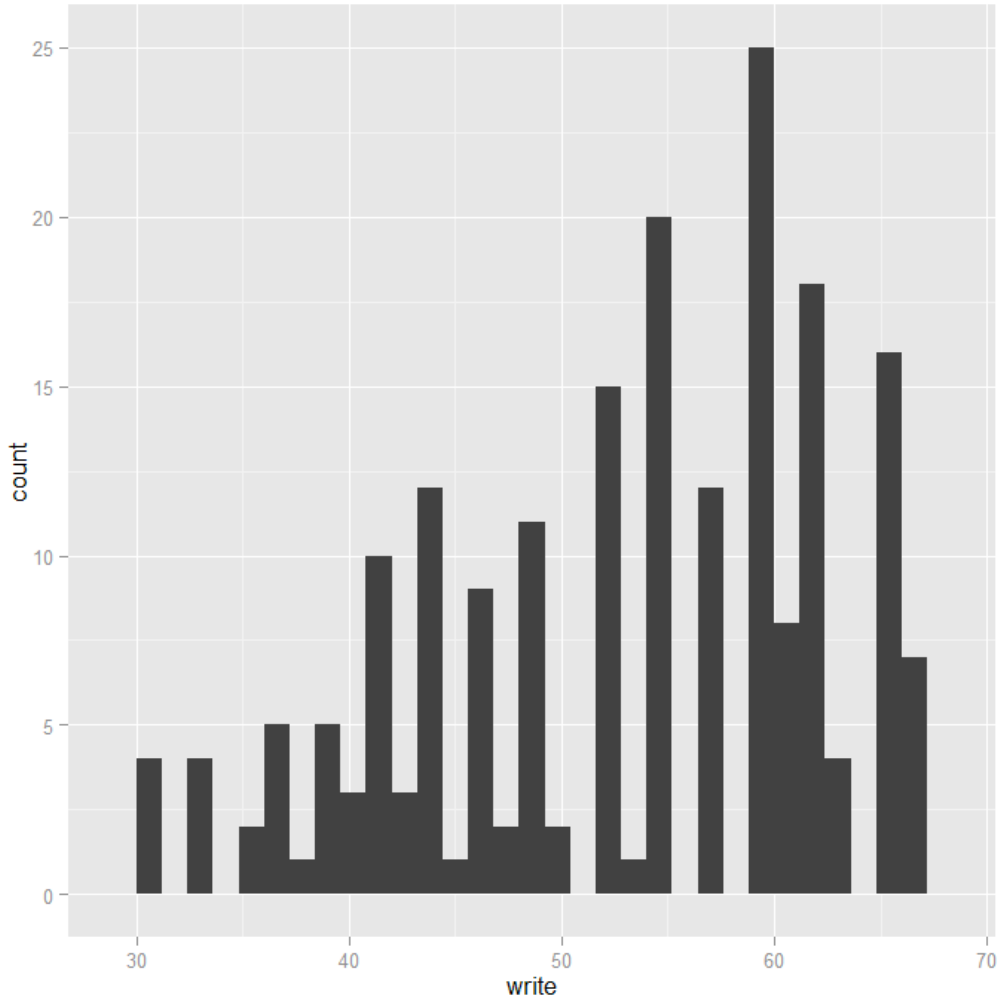
Interacting with R - RStudio

You
grap

- Rs ent



R as a Graphical Tool



Data representation in R

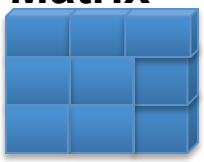
MONASH University

Vector

Collection of data of the same basic type

- 2, 3, 5 - numeric
- TRUE, FALSE – logical
- ”a", ”b” - character

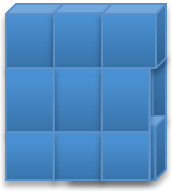
Matrix



Collection of data of the same type in 2D rectangular layout

Samples1	Samples2	Samples2
3.3	4.7	1.4
4.4	5.6	2.0
-6.4	6.5	4.0

Data frame

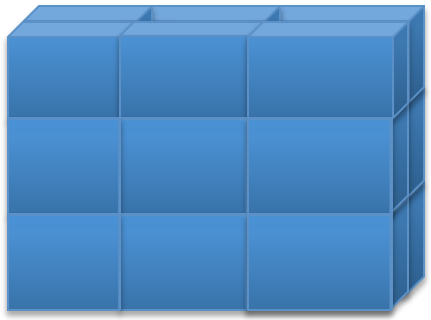


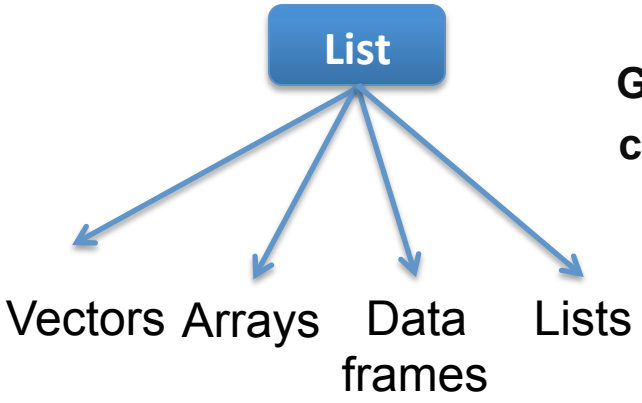
Used for keeping data tables

Columns can be of different types

Chr	Start	End	Coverage
chr1	3016214	3016215	4
chr1	3016215	3016216	2
chr1	3016283	3016284	4

Array





Generic vector contains other objects

Working with Data in R

Data Files (text)

Typically values in data files are separated or delimited

- **by tabs or spaces**

Chr	Start	End	MethyCoverage	TotalCoverage	Strand
chr1	3016214	3016215	4	5	F
chr1	3016215	3016216	2	2	R
chr1	3016283	3016284	4	4	F
chr1	3016284	3016285	2	3	R

- **by commas**

```
Chr,  Start ,  End, MethyCoverage, TotalCoverage, Strand
chr1, 3016214, 3016215, 4,  5, F
```

- R provides a number of formats to read and save our data

Viewing Data in R

- Typically datasets stored as data frames in R.
- Individual rows, columns and cells in a data frame can be accessed via object[row,column] notation

Chr	Start	End	MethyCoverage	TotalCoverage	Strand
chr1	3016214	3016215	4	5	F
chr1	3016215	3016216	2	2	R
chr1	3016283	3016284	4	4	F

Single cell value : data[1,3]

```
3016215
```

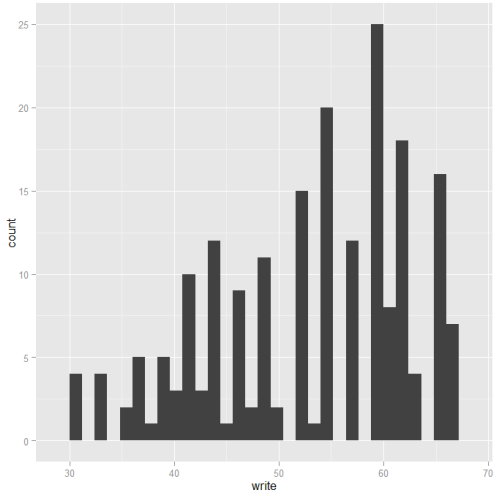
All columns in row 3: dat[3,]

Chr	Start	End	Methylatedcoverage	TotalCoverage	Strand
chr1	3016283	3016284	4	4	F

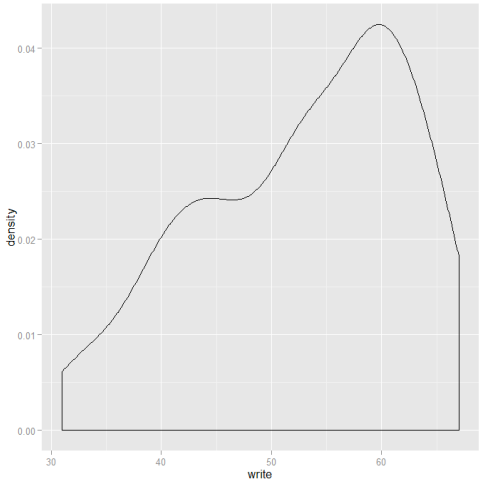
- **R** offers numbers of ways for simple (or complex) calculations

Exploring Data

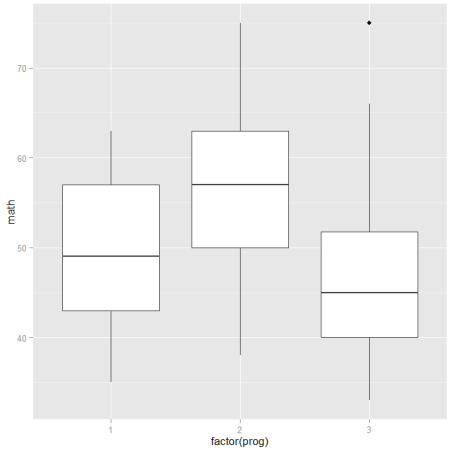
R plots



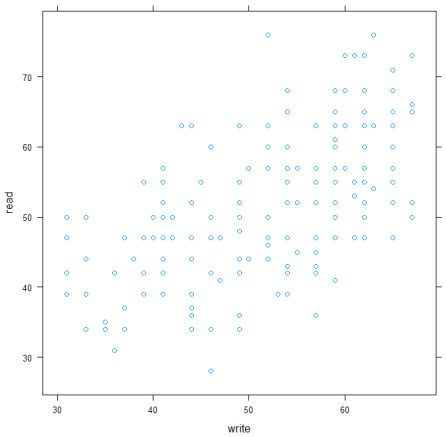
Histogram



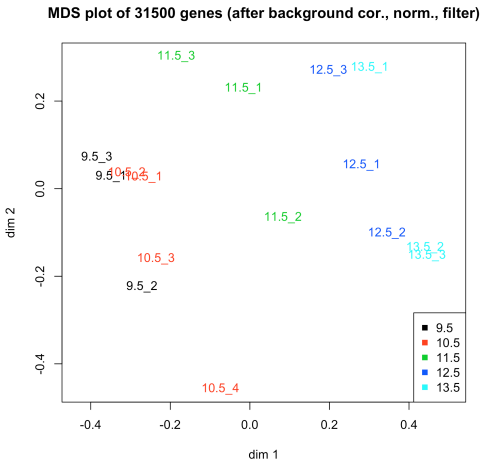
Density plot



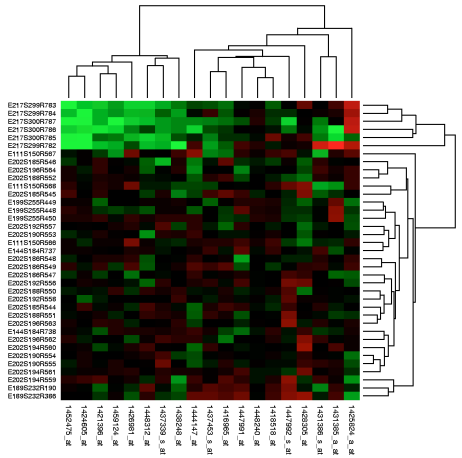
Boxplot



Scatter



MDS/PCA plot



Heatmap

Correlation

cor function tests for a relationship between two numerical variables/vectors.

Gene ID	benign	Primary	Metastasis
1	3.1	-3.43	6.62
2	-4.5	1.93	3.61
3	2.7	3.61	-1.98
4	1.9	2.57	2.63

cor

	benign	Primary	Metastasis
benign	1	0.59	0.66
Primary	0.59	1	0.61
Metastasis	0.66	0.61	1
	0.63	0.57	0.63

Pairwise correlation among samples in columns 2 through 4

Modifying Data

Sorting

arrange function sorts rows by variable/column name

Gene ID	benign	Primary	Metastasis
2	3.1	-3.43	6.62
4	-4.5	1.93	3.61
3	2.7	3.61	-1.98
1	1.9	2.57	2.63

arrange

Gene ID	benign	Primary	Metastasis
1	1.9	2.57	2.63
2	3.1	-3.43	6.62
3	2.7	3.61	-1.98
4	-4.5	1.93	3.61

Sort the table by column (Gene ID)

Subsetting data

subset function splits the data into 2 datasets

Chr	Start	End	Coverage	TotalCoverage	Strand
chr1	3016214	3016215	4	5	F
chr1	3016215	3016216	2	2	R
chr1	3016283	3016284	4	4	F

Chr	Start	End	Coverage	TotalCoverage	Strand
chr1	3016214	3016215	4	5	F
chr1	3016283	3016284	4	4	F

Chr	Start	End	Coverage	TotalCoverage	Strand
chr1	3016215	3016216	2	2	R

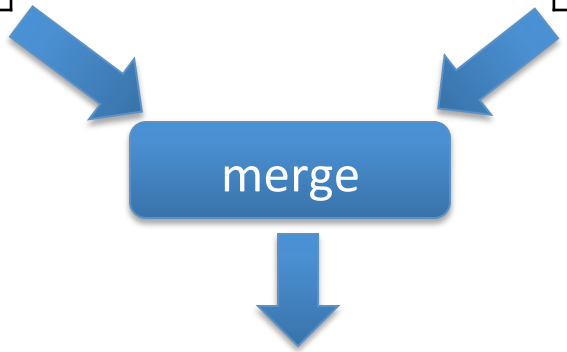
- rbind function appends the datasets row-wise

Merging data

merge function combines data by common columns

Gene ID	benign
2	3.1
4	-4.5
3	2.7
1	1.9

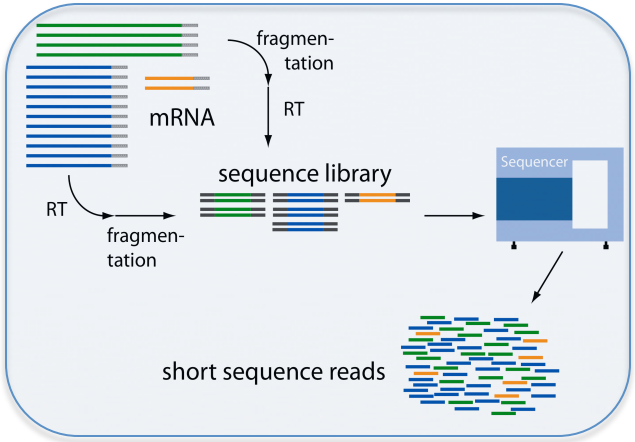
Gene ID	Primary	Metastasis
2	-3.43	6.62
4	1.93	3.61
3	3.61	-1.98
1	2.57	2.63



Gene ID	benign	Primary	Metastasis
2	3.1	-3.43	6.62
4	-4.5	1.93	3.61
3	2.7	3.61	-1.98
1	1.9	2.57	2.63

Analyzing Genomics Data

RNA-seq data analysis workflow



Pre-processing

Mapping

Differential expression

Downstream analysis

- A typical RNA-seq experiment

- Check short raw sequence reads
- Trim/filter raw sequence reads for minimum quality

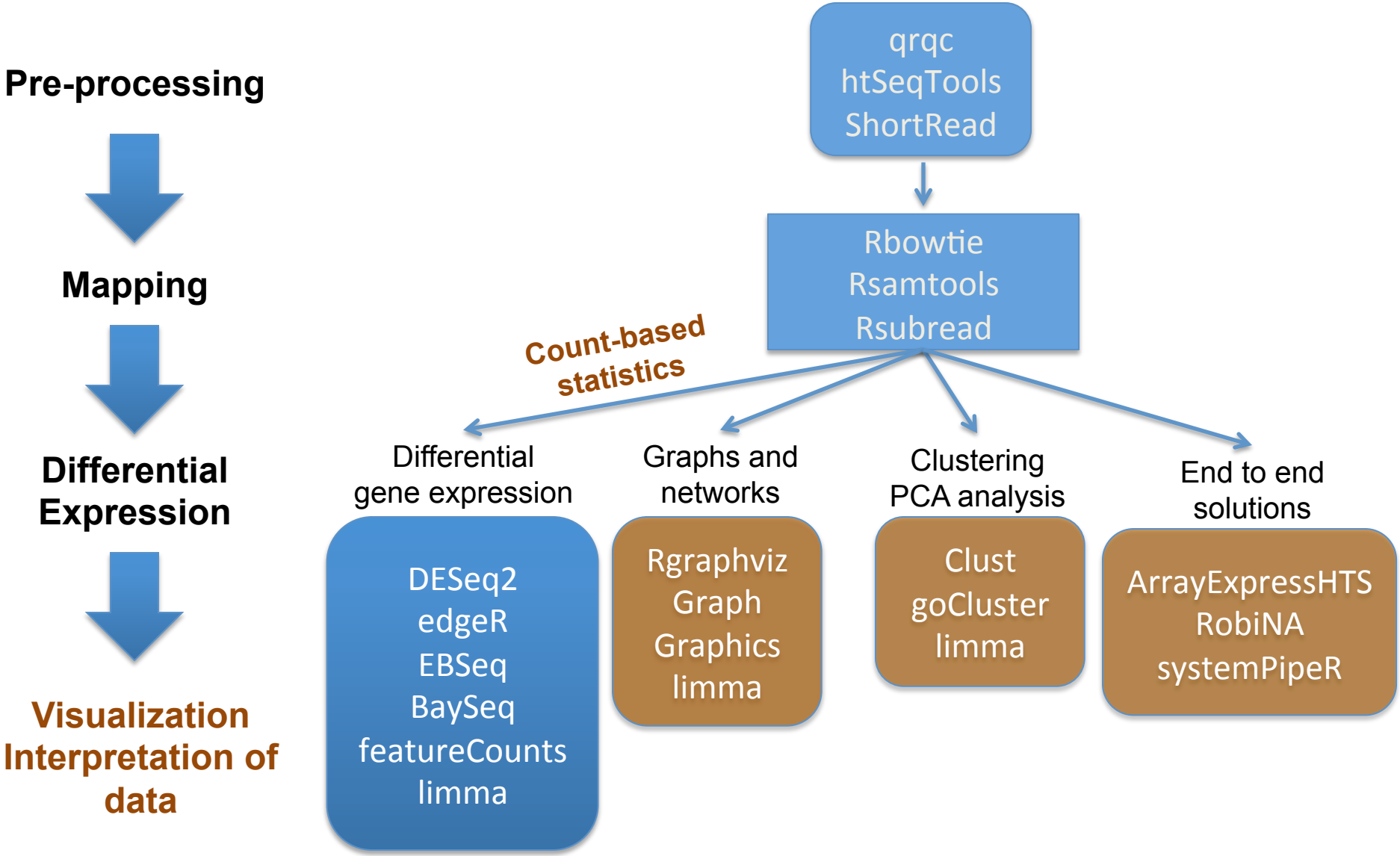
- Many options for short read alignment tools (Bowtie, STAR)

- Many options for count-based statistics

- Interpretation of results
- Data visualization, GO and Pathway analysis

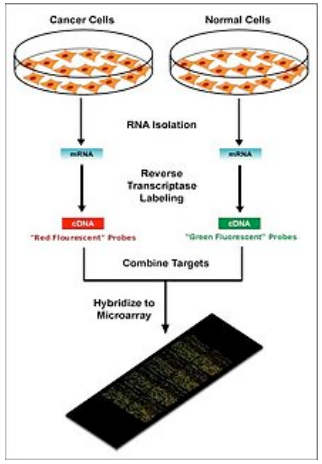
RNA-seq data analysis in R

MONASH University

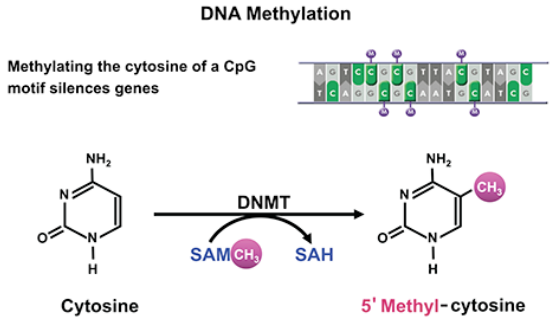


Other technologies

R also provides statistical frameworks and tools for the analysis and comprehension of



Microarray data



DNA methylation data

Web links

- Introducing R (UCLA Institute for Digital Research & Education) - http://www.ats.ucla.edu/stat/r/seminars/intro.htm
- Getting Started with the R Data Analysis Package - http://heather.cs.ucdavis.edu/~matloff/r.html
- R tutorial from the O' Reilly book series - http://tryr.codeschool.com/levels/1/challenges/1
- R Tutorial – An R Introduction to Statistics - http://www.r-tutor.com/r-introduction/matrix

Thank you very much for your attention