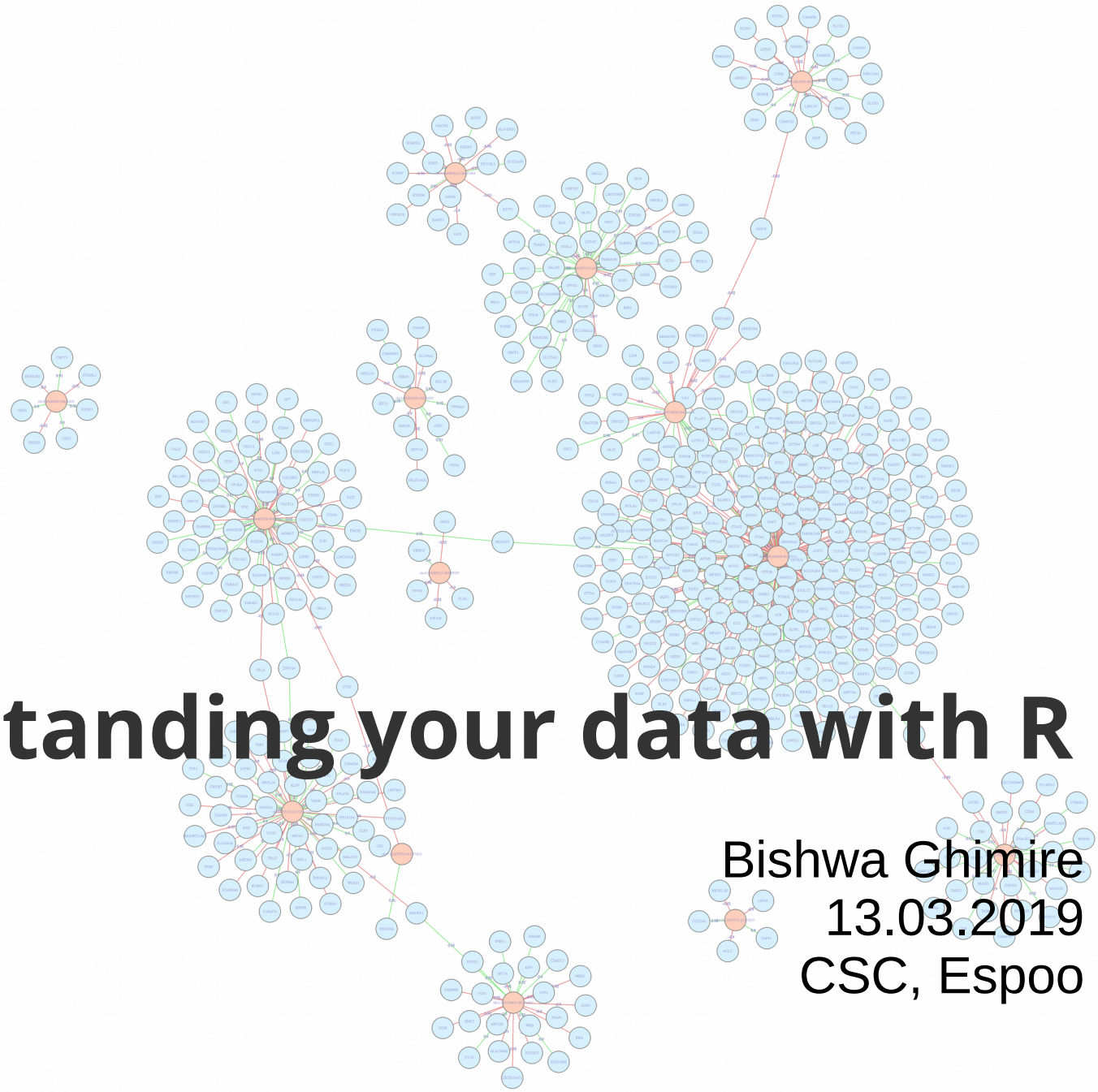


Understanding your data with R

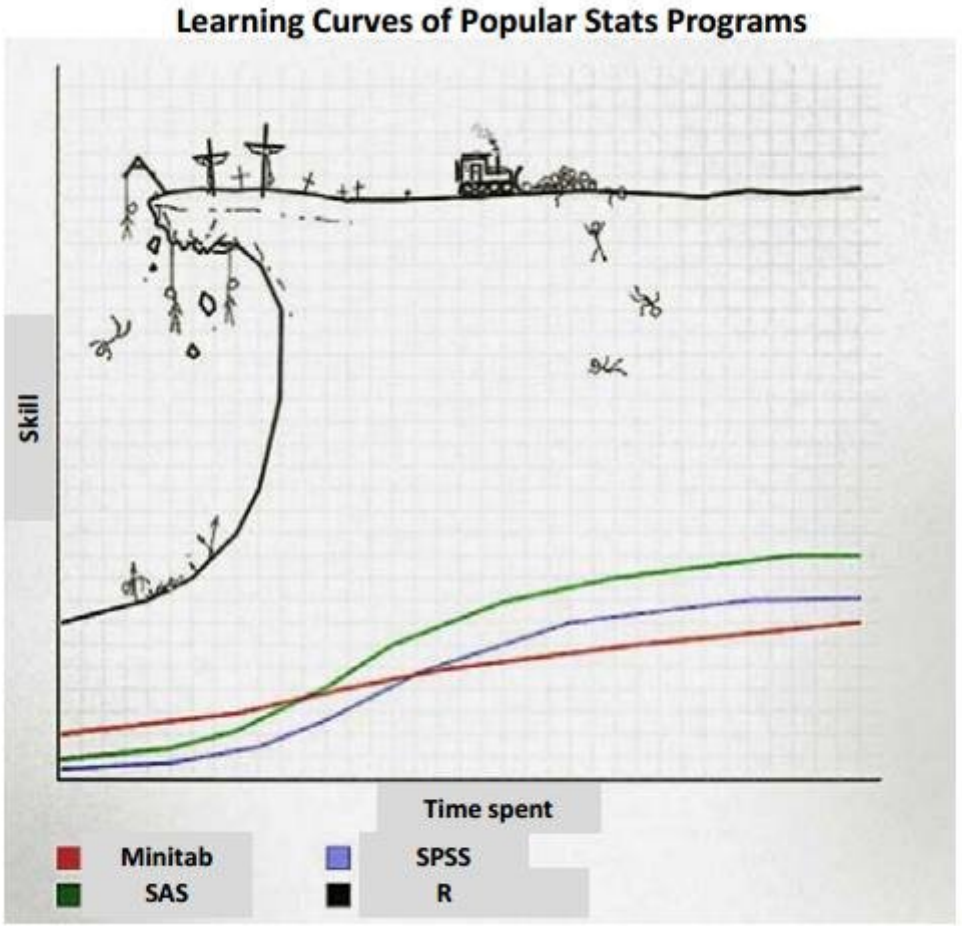
Bishwa Ghimire

13.03.2019

CSC, Espoo

Session 1: Getting started

R learning curve



Source: Rogier Kievit

Why R?

- Free and open source
- Platform independent
- Reproducible results
- Huge community

Data set 1

Developmental Cell
Article

Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals

Jung-Woong Kim,[1,2,6] Hyun-Jin Yang,[1,6] Adam Phillip Oel,[3,6] Matthew John Brooks,[1] Li Jia,[4] David Charles Plachetzki,[5] Wei Li,[4] William Ted Allison,[3,*] and Anand Swaroop[1,*]

[1]Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]Department of Life Science, College of Natural Sciences, Chung-Ang University, Seoul 156-756, Republic of Korea
[3]Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada
[4]Retinal Neurophysiology Section, National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA
[5]Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH 03824, USA
[6]Co-first author
*Correspondence: ted.allison@ualberta.ca (W.T.A.), swaroopa@nei.nih.gov (A.S.)
http://dx.doi.org/10.1016/j.devcel.2016.05.023

GEO accession: GSE74660

Data set 2

SCIENTIFIC REPORTS

Correction: Author Correction

OPEN

Transcriptomic and epigenetic responses to short-term nutrient-exercise stress in humans

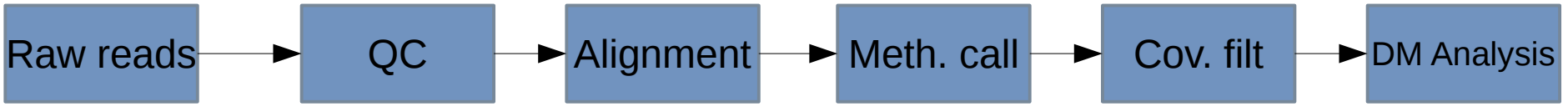
Received: 13 July 2017
Accepted: 27 October 2017
Published online: 09 November 2017

R. C. Laker[1], C. Garde[1], D. M. Camera[2], W. J. Smiles[2], J. R. Zierath[1,3], J. A. Hawley[2,4] & R. Barrès[1]

High fat feeding impairs skeletal muscle metabolic flexibility and induces insulin resistance, whereas exercise training exerts positive effects on substrate handling and improves insulin sensitivity. To identify the genomic mechanisms by which exercise ameliorates some of the deleterious effects of high fat feeding, we investigated the transcriptional and epigenetic response of human skeletal muscle to 9 days of a high-fat diet (HFD) alone (Sed-HFD) or in combination with resistance exercise (Ex-HFD), using genome-wide profiling of gene expression and DNA methylation. HFD markedly induced expression of immune and inflammatory genes, which was not attenuated by Ex. Conversely, Ex markedly remodelled expression of genes associated with muscle growth and structure. We detected marked DNA methylation changes following HFD alone and in combination with Ex. Among the genes that showed a significant association between DNA methylation and gene expression changes were PYGM, which was epigenetically regulated in both groups, and ANGPTL4, which was regulated only following Ex. In conclusion, while short-term Ex did not prevent a HFD-induced inflammatory response, it provoked a genomic response that may protect skeletal muscle from atrophy. These epigenetic adaptations provide mechanistic insight into the gene-specific regulation of inflammatory and metabolic processes in human skeletal muscle.

GEO accession: GSE99965

Data analysis (RRBS)



samples

Chr	Start	End	meth.level	noC	noT
chr1	1000233	1000233	0	0	26
chr1	1000239	1000239	6.25	2	30
chr1	100038192	100038192	8.6	2	21
chr1	100038210	100038210	0	0	23
chr1	100038220	100038220	0	0	23

Pre

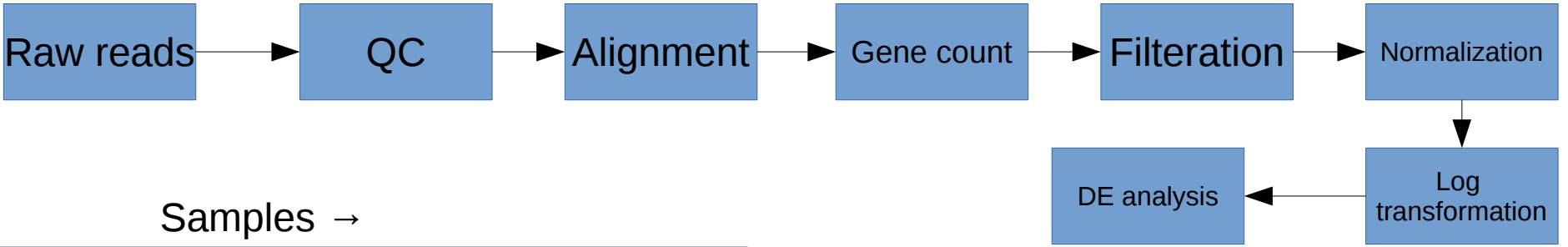
Post

Methylation call

Pre vs. post DM analysis

Chr	Start	End	meth1	meth2	meth .diff
chr1	1000233	1005233	0.5	1	-0.5
chr1	1000239	1000339	0.1	0.5	-0.4
chr1	100038192	100041192	0.9	0.2	0.7
chr1	100038210	100038510	0.85	0.21	0.64
chr1	100038220	100039720	0.1	0.8	-0.7

Data analysis (RNA-Seq)



Samples →

↓ Genes

	S13_pre	S14_post	S14_pre	S2_post	S2_pre
DDX11L1	0	0	0	0	0
WASH7P	7	13	17	2	11
MIR6859-3	0	0	0	0	0
MIR6859-2	0	0	0	0	0
MIR6859-1	0	0	0	0	0
MIR6859-4	0	0	0	0	0
MIR1302-2	0	0	0	0	0
MIR1302-11	0	0	0	0	0
MIR1302-9	0	0	0	0	0
MIR1302-10	0	0	0	0	0

	logFC	AvgExp	P.value	adj.P.value
METTL21EP	1.97	1.15	9.6E-11	6.7E-07
CA14	2.60	1.76	1.1E-10	6.7E-07
CALML6	3.43	2.48	1.7E-10	6.7E-07
FAM57B	1.98	1.24	4.2E-10	1.2E-06
FAM166B	4.83	3.94	1E-09	1.9E-06
TTN	18.54	18.40	3.1E-09	1.9E-06
ART3	5.59	4.82	3.8E-09	1.9E-06
MT1X	5.40	4.65	4E-09	1.9E-06
LOC646736	2.53	1.91	4.6E-09	1.9E-06

Pre vs. post DE analysis

1.0 Installation

- Go to https://ftp.acc.umu.se/mirror/CRAN/
- Download R specific to your operating system.
- Install the program.

CRAN
Mirrors
What's new?
Task Views
Search

About R

The Comprehensive R Archive Network

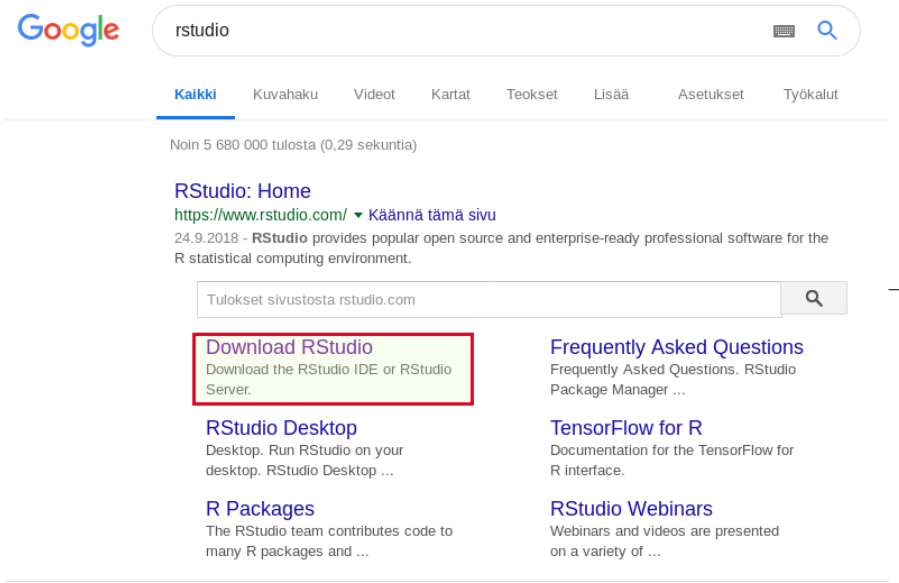
Download and Install R

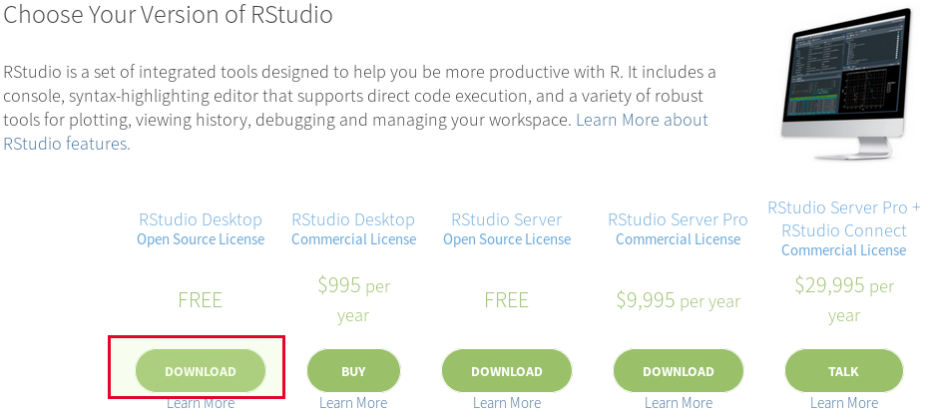
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- 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.

1.0 Installation

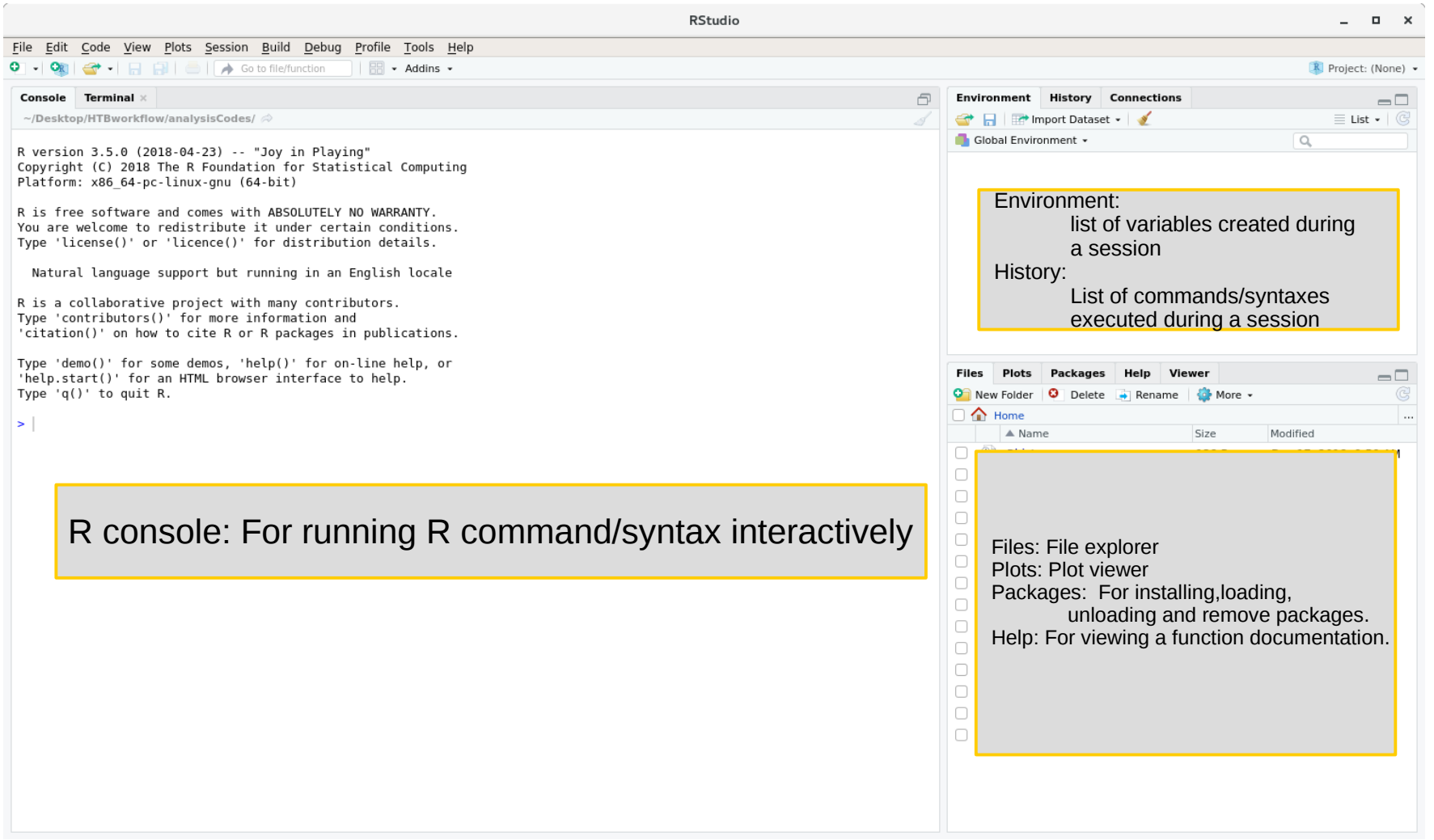




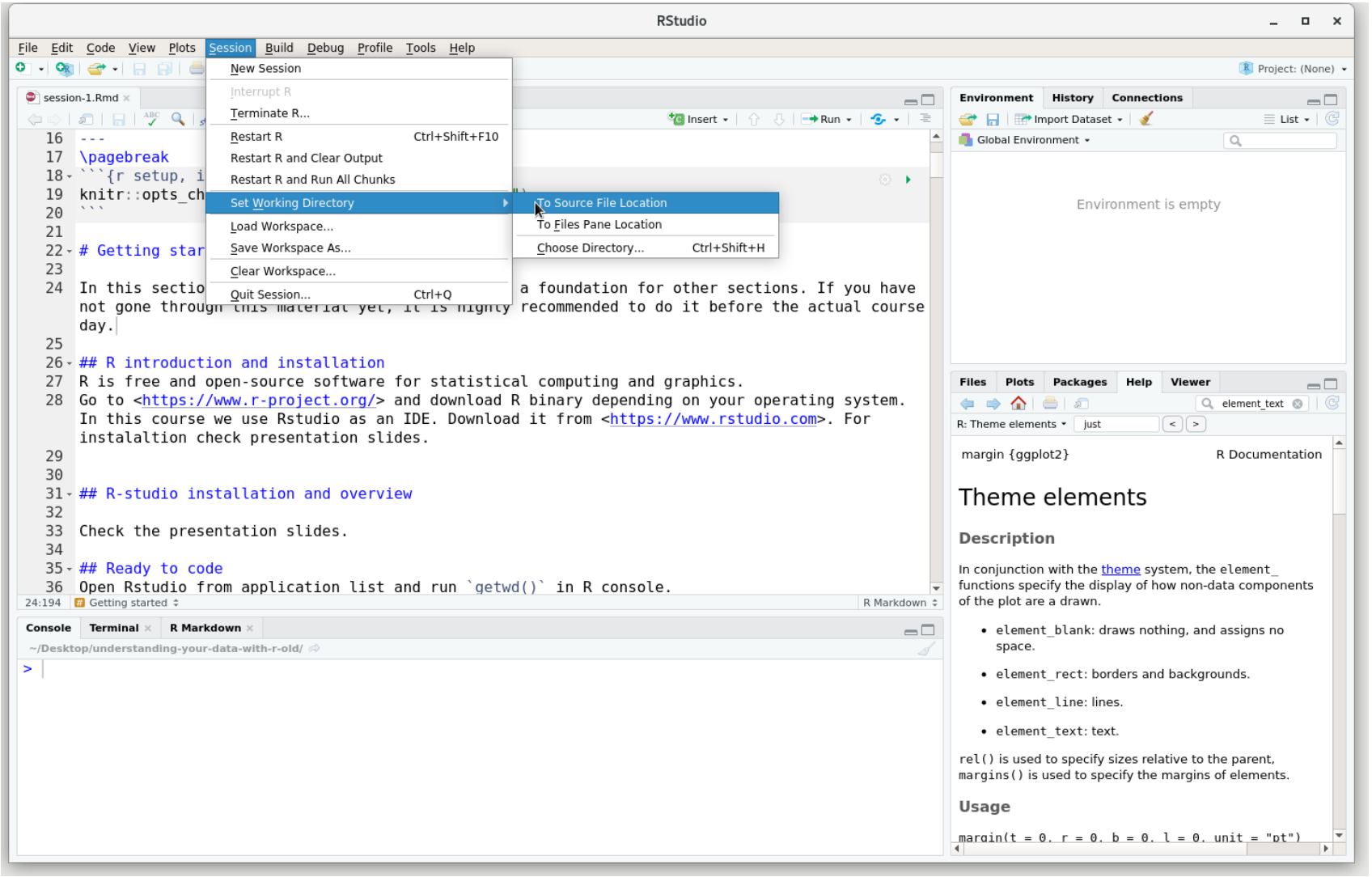
Installers for Supported Platforms

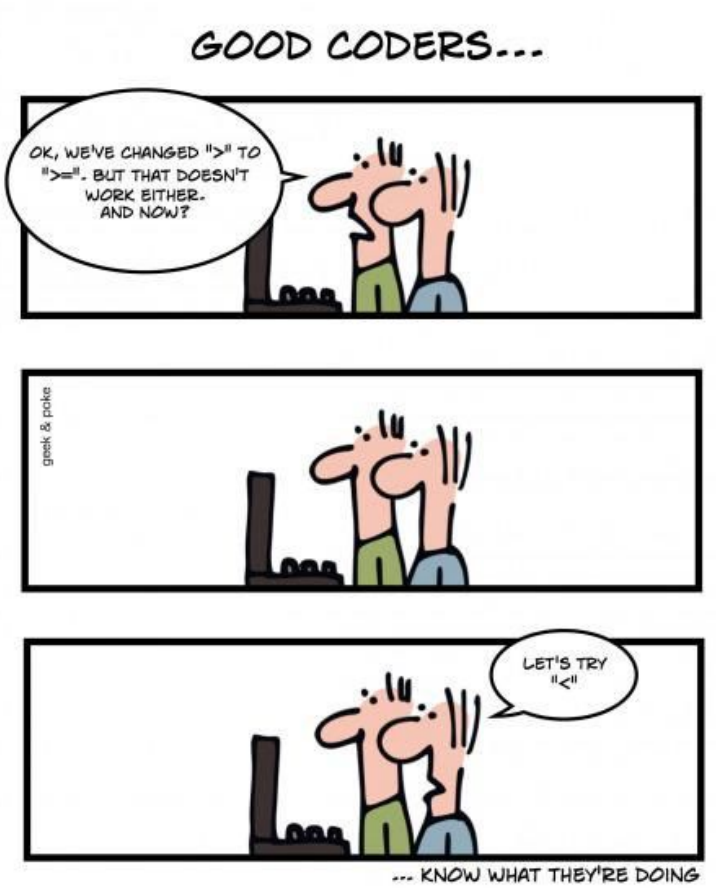
Installers	Size	Date	MD5
RStudio 1.1.463 - Windows Vista/7/8/10	85.8 MB	2018-10-29	58b3d796d8cf96fb8580c62f46ab64d4
RStudio 1.1.463 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-10-29	a79032ba4d7daaa86a8da01948278d94
RStudio 1.1.463 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-10-29	8a6755fa9fae2bafce289df3358aaf63
RStudio 1.1.463 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-10-29	bc50d6bd34926c1cc3ae4a209d67d649
RStudio 1.1.463 - Ubuntu 16.04+/Debian 9+ (64-bit)	65 MB	2018-10-29	cfd659db18619cc78d1592fefaa7c753
RStudio 1.1.463 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-10-29	742f0bad60dfeaa3281576e14ad6699e
RStudio 1.1.463 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-10-29	c7303067a0ca99deea7e427b856952d1

1.2 R-studio overview



1.3 Ready to code





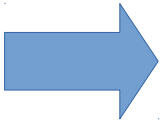
Session 2: Data Manipulation

Data frame slicing

my.df

	c1	c2	c3	c4	c5
r1					
r2					
r3					
r4					
r5					

- `my.df[ ,c('c1','c2')]`
- `my.df[ ,c(TRUE, TRUE, FALSE, FALSE)]`
- `my.df[ ,1:2]`
- `my.df[ ,c(1,2)]`



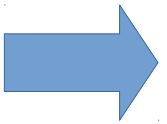
	c1	c2
r1		
r2		
r2		
r3		
r4		

Data frame slicing

my.df

	c1	c2	c3	c4	c5
r1					
r2					
r3					
r4					
r5					

- `my.df[c('r1', 'r2', 'r3'), ]`
- `my.df[,c(TRUE,TRUE,TRUE,FALSE,FALSE,FALSE)]`
- `my.df[1:3, ]`
- `my.df[c(1, 2, 3), ]`



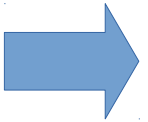
	c1	c2	c3	c4	c5
r1					
r2					
r3					

Data frame slicing

my.df

	c1	c2	c3	c4	c5
r1					
r2					
r3					
r4					
r5					

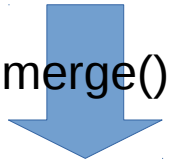
- `my.df[c('r1', 'r2'), c('c3', 'c4', 'c5') ]`
- `my.df[c(TRUE,TRUE,FALSE,FALSE,FALSE,FALSE), c(FALSE,FALSE,TRUE,TRUE,TRUE,FALSE)]`
- `my.df[1:2, 3:5]`
- `my.df[c(1,2), c(3,4,5) ]`



	c3	c4	c5
r1			
r2			

Merge

	cell1	cell2	cell3
g1			
g2			
g3			
g4			
g5			
g6			



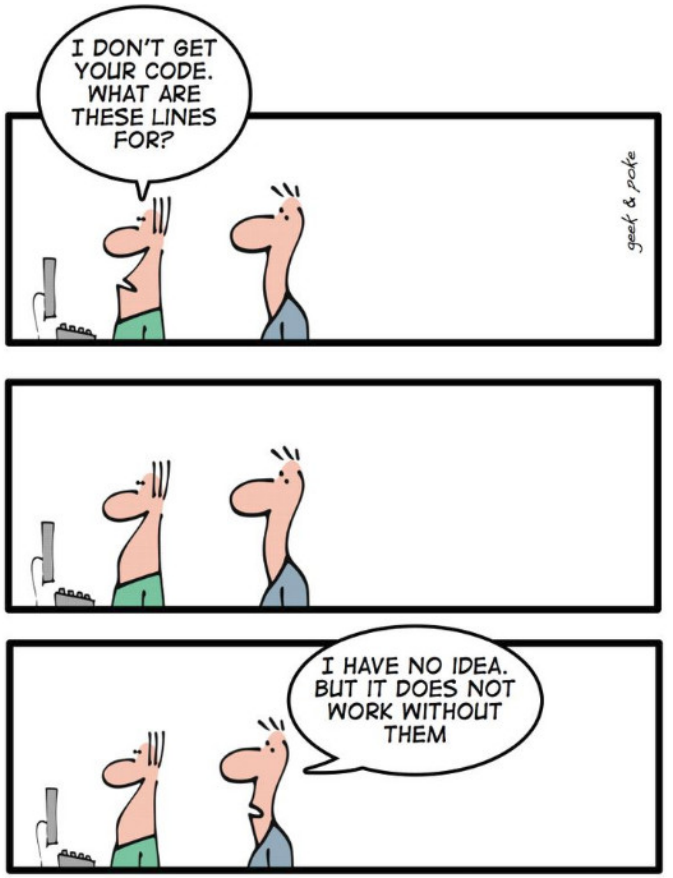
	cell10	cell20	cell30	cell40	cell50
g1					
g2					
g3					
g4					
g5					

	cell1	cell2	cell3	cell10	cell20	cell30	cell40	cell50
g1								
g2								
g3								
g4								
g5								
g6	NA	NA	NA					

Tidy format

	smp1	smp2	smp3	smp4
gene1	7	0	2	5
gene2	7	0	3	4
gene3	9	0	2	7
gene4	10	0	2	5
gene5	7	0	2	5

gene	variable	value
gene1	smp1	7
gene2	smp1	7
gene3	smp1	9
gene4	smp1	10
gene5	smp1	7
gene1	smp2	0
gene2	smp2	0
gene3	smp2	0
gene4	smp2	0
gene5	smp2	0
gene1	smp3	2
gene2	smp3	3
gene3	smp3	2
gene4	smp3	2
gene5	smp3	2
gene1	smp4	5
gene2	smp4	4
gene3	smp4	7
gene4	smp4	5
gene5	smp4	5

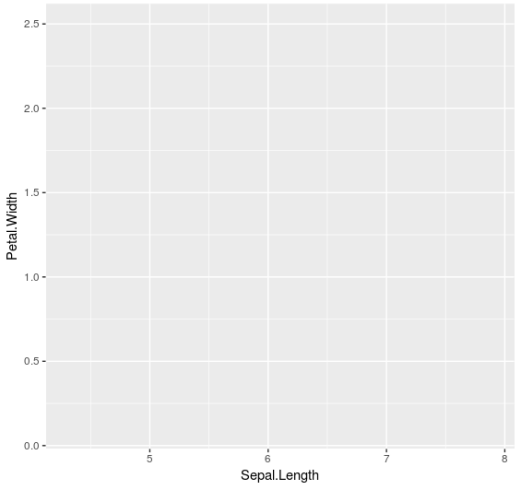


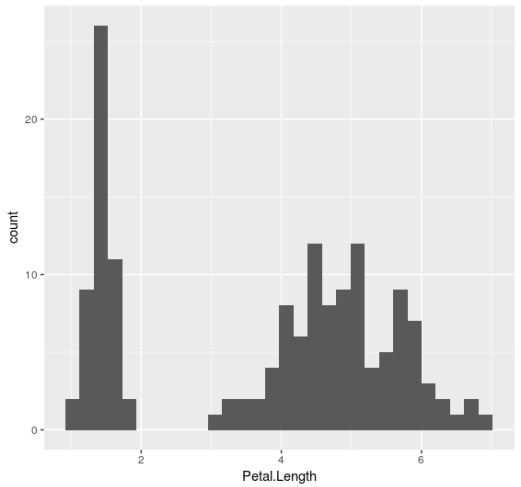
Session 3: Basic visualization

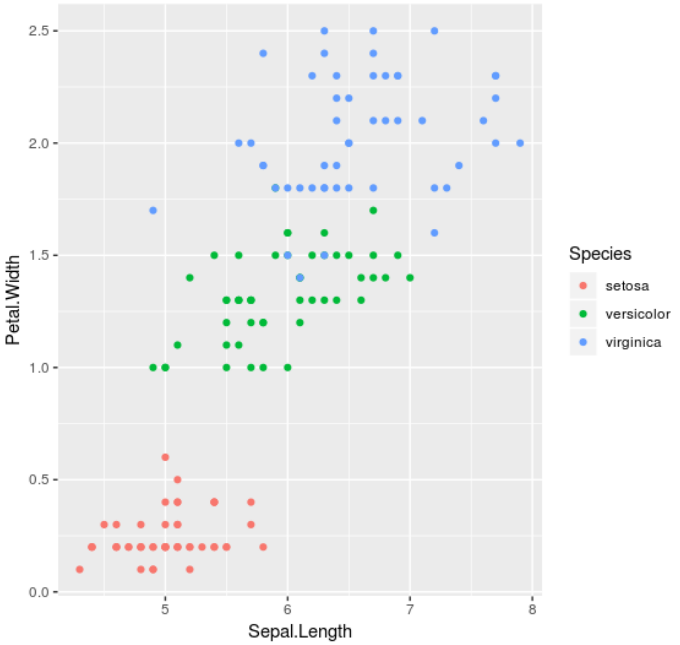
Ggplot

Syntax

```
library("ggplot2")
data("iris")

## making x-axis and y-axis
ggplot(iris, aes(x=Sepal.Length, y=Petal.Width))

## when only one x-axis is used
ggplot(iris, aes(x=Petal.Length)) +
  geom_histogram()

## add a layer of points
ggplot(iris, aes(x=Sepal.Length, y=Petal.Width, color=Species)) +
  geom_point()
```



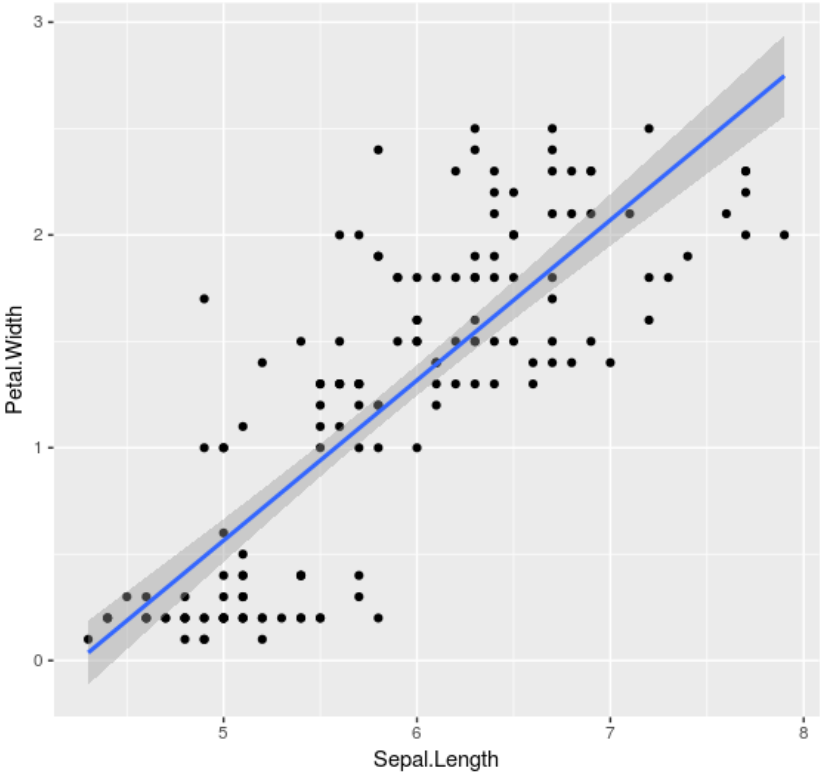




Ggplot

Syntax

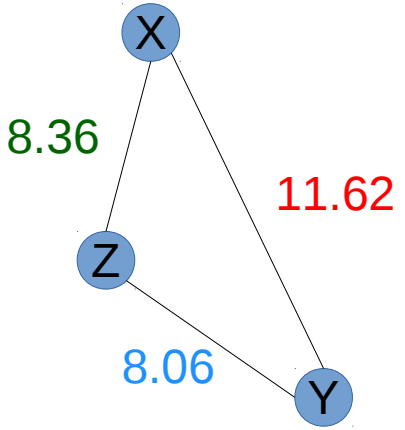
```
## add two layers
ggplot(iris, aes(x=Sepal.Length, y=Petal.Width)) +
  geom_point() +
  geom_smooth(method="lm")

## save to variable
p1 <- ggplot(iris, aes(x=Sepal.Length, y=Petal.Width)) +
  geom_point()

p2 <- p1 + geom_smooth(method="lm")

## save plot
ggsave(file ="histogram.png",
       width = 300,
       heigh = 300,
     ! dpi = 300,
       units="mm")
```



Distance matrix

Table: Expression matrix toy data.

	SampX	SampY	SampZ
Gene A	4	1	5
Gene B	3	4	1
Gene C	10	0	2
Gene D	6	1	7



Euclidean distance $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$

$XY = \sqrt{(4-1)^2 + (3-4)^2 + (10-0)^2 + (6-1)^2} = 11.62$

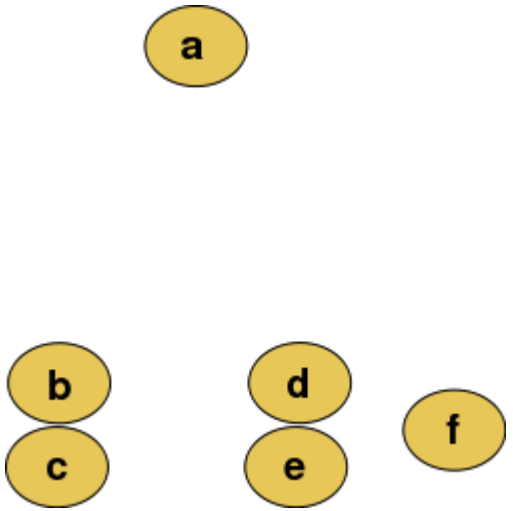
$XZ = 8.36$

$YZ = 8.06$

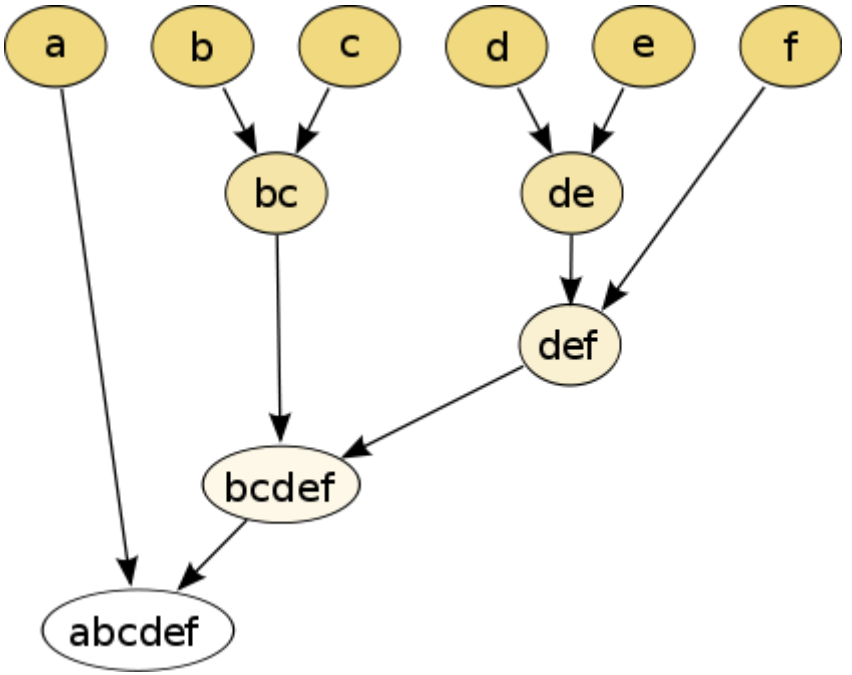
Distance matrix

	X	Y	Z
X			
Y	11.62		
Z	8.36	8.06	

Hierarchical clustering



Raw data



Hierarchical clustering

Image source: Wikipedia

Session 4: Advanced visualization

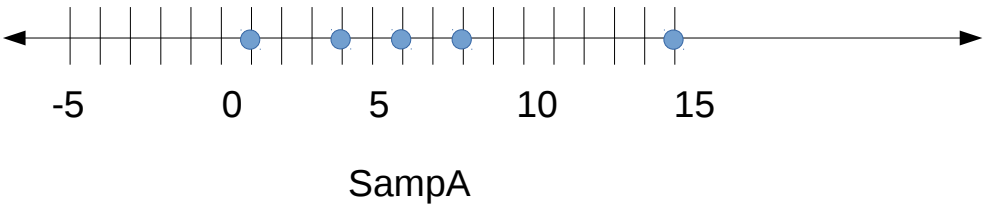
PCA

- Dimension reduction technique
- Idea is to preserve most of the variation when reduced to lower dimensions.
- Does not work well with the data
  - having non linear relationship with the variables
  - having low variation

PCA (*Cont.*)

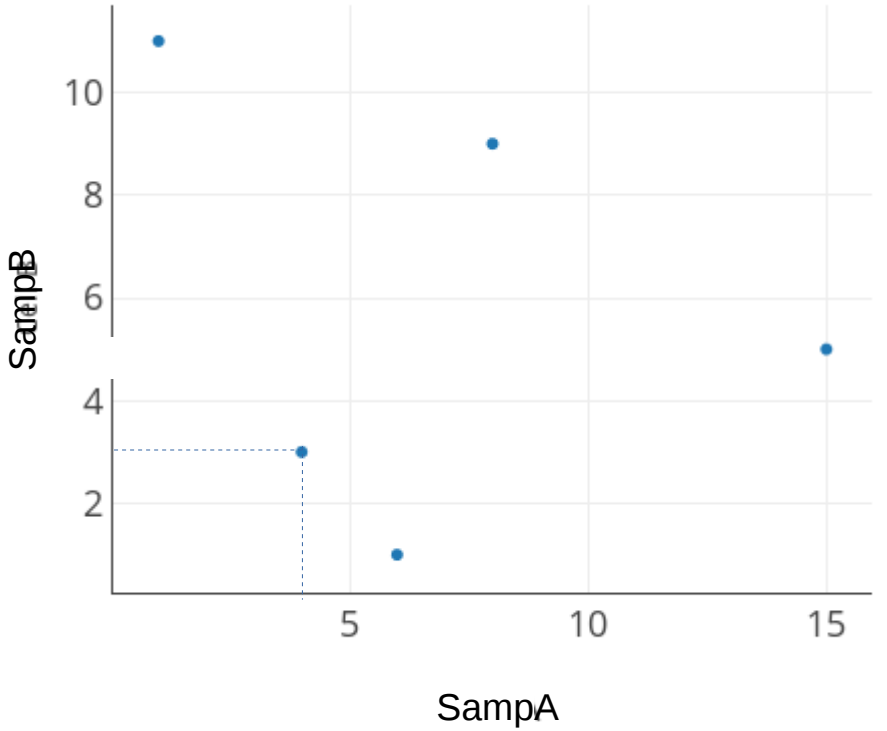
About dimension.

Gene	SampA
g1	4
g2	6
g3	8
g4	1
g5	15



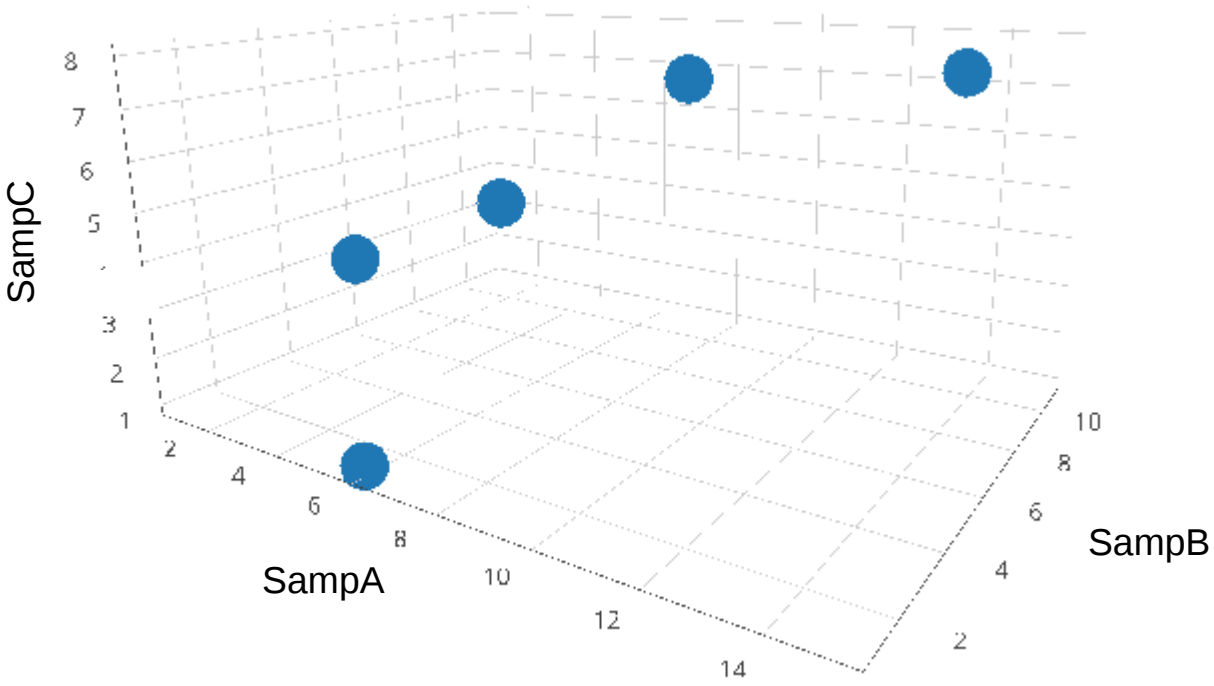
PCA (*Cont.*)

Gene	sampA	SampB
g1	4	3
g2	6	1
g3	8	9
g4	1	11
g5	15	5



PCA (*Cont.*)

Gene	SampA	SampB	SampC
g1	4	3	4
g2	6	1	1
g3	8	9	7
g4	1	11	3
g5	15	5	8



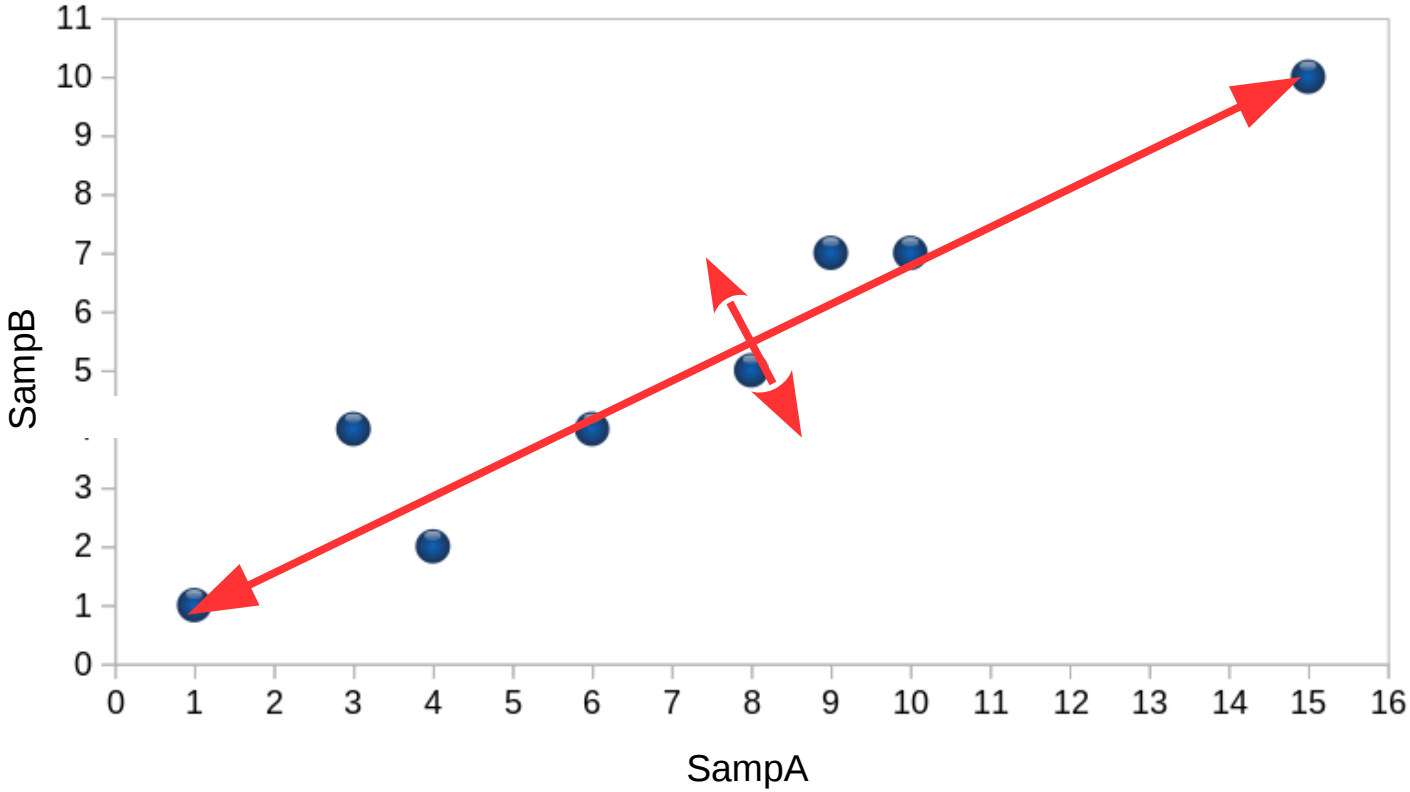
PCA (*Cont.*)

Gene	Samp A	SampB	SampC	SampD
g1	4	3	4	2
g2	6	1	1	1
g3	8	9	7	0
g4	1	11	3	4
g5	15	5	8	9

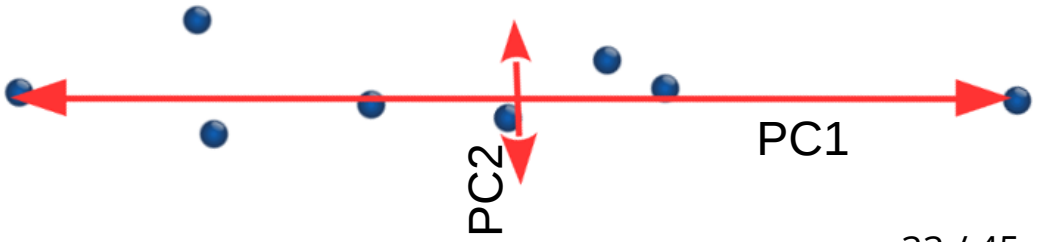
How does it look? I have even 200 samples.

PCA (*Cont.*)

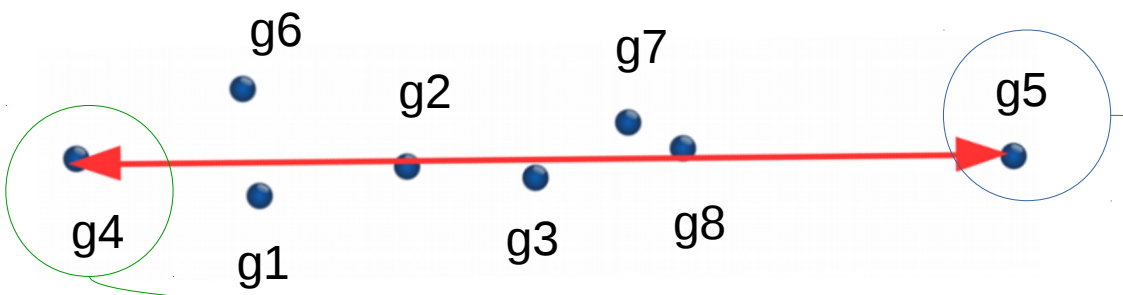
Gene	Samp A	Samp B
g1	4	2
g2	6	4
g3	8	5
g4	1	1
g5	15	10
g6	3	4
g7	9	7
g8	10	7



Here we are plotting **genes** not samples.



PCA (*Cont.*)



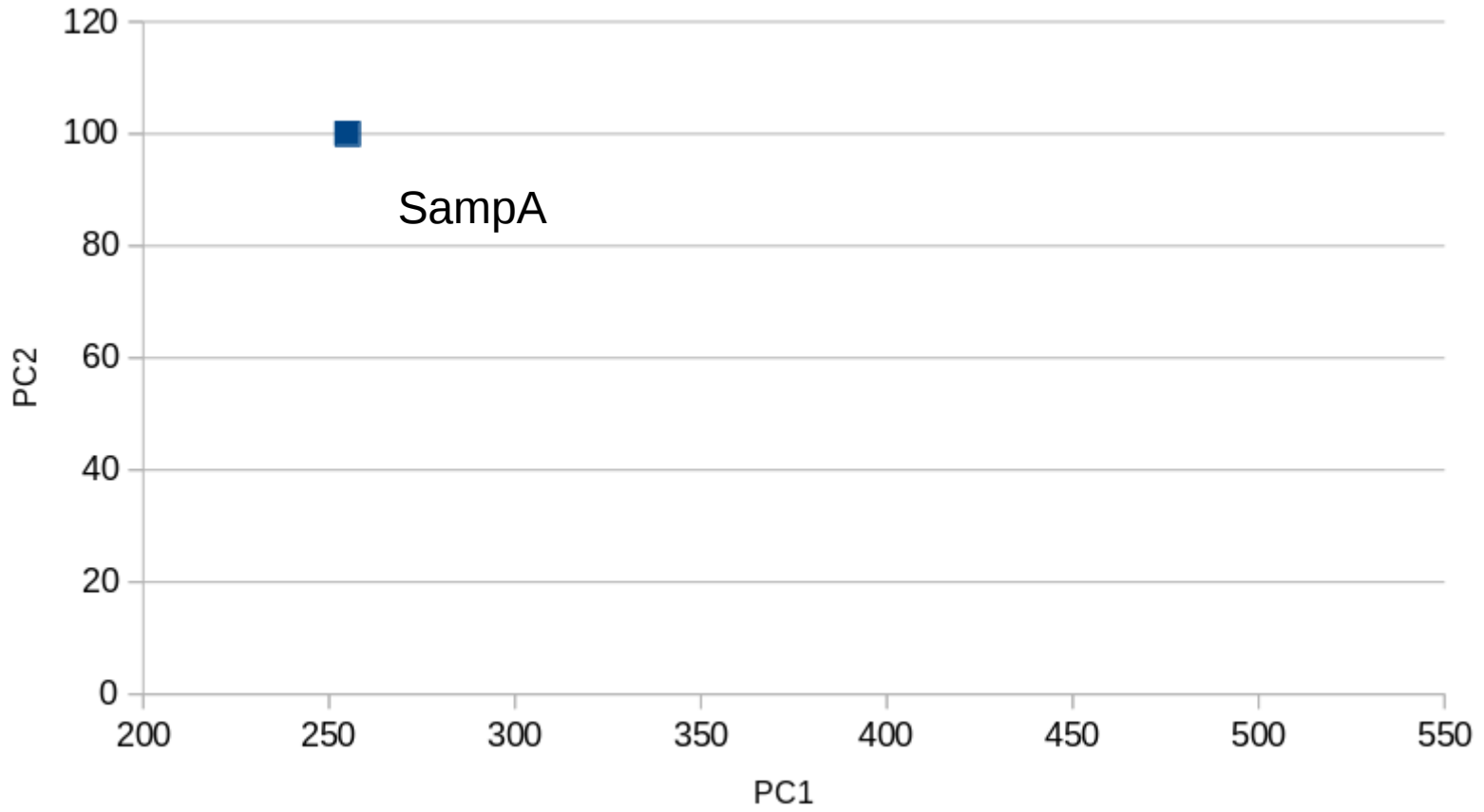
Gene	Influence in PC1	In numbers
g1	high	-9
g2	low	4
g3	low	0
g4	high	-14
g5	high	15
g6	high	-10
g7	low	4
g8	low	5

Original data

Gene	SampA	SampB
g1	4	2
g2	6	4
g3	8	5
g4	1	1
g5	15	10
g6	3	4
g7	9	7
g8	10	7

PC1 score = (4 * -9) + (6 * 4) …… = 255
PC2 score = (4 * influence of g1 in PC2)
+ (6 * influence of g2 in PC2) …
= 50 (lets say)

PCA (*Cont.*)


120
100
80
60
40
20
0
PC2
SampA
200
250
300
350
400
450
500
550
PC1

PCA (*Cont.*)

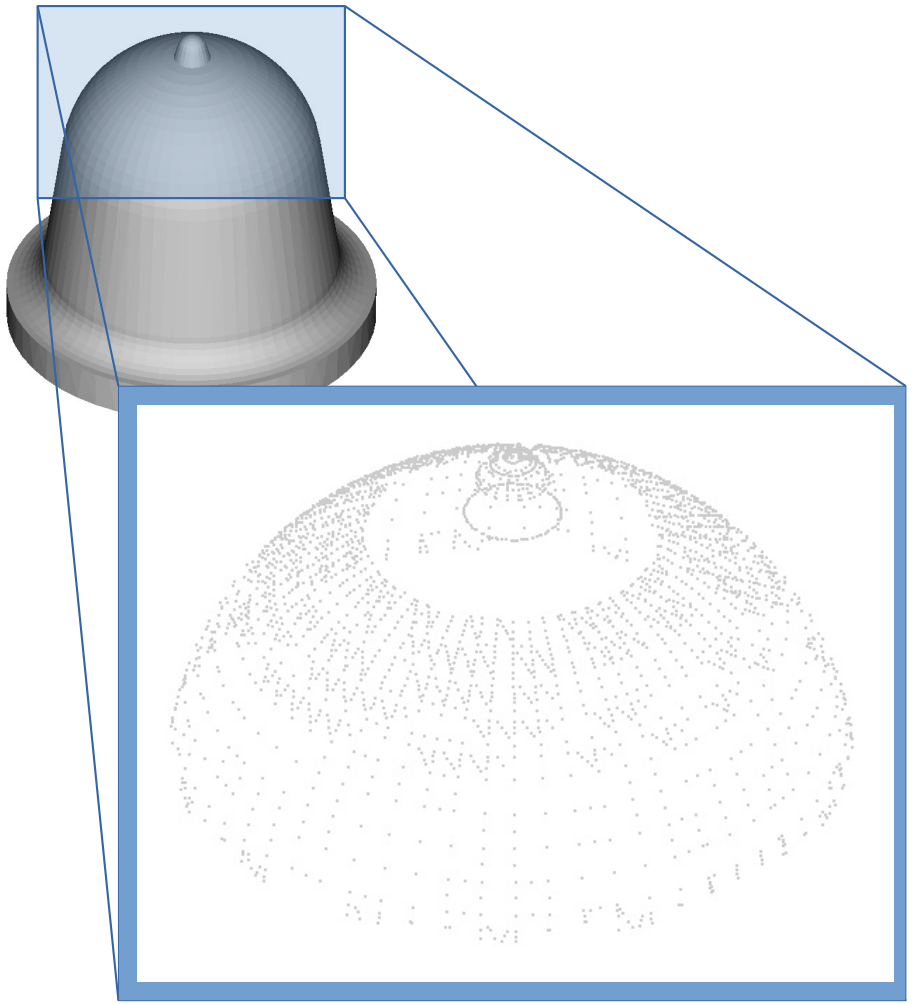
Raw data

PCA reduced data

Fig: PCA analogy

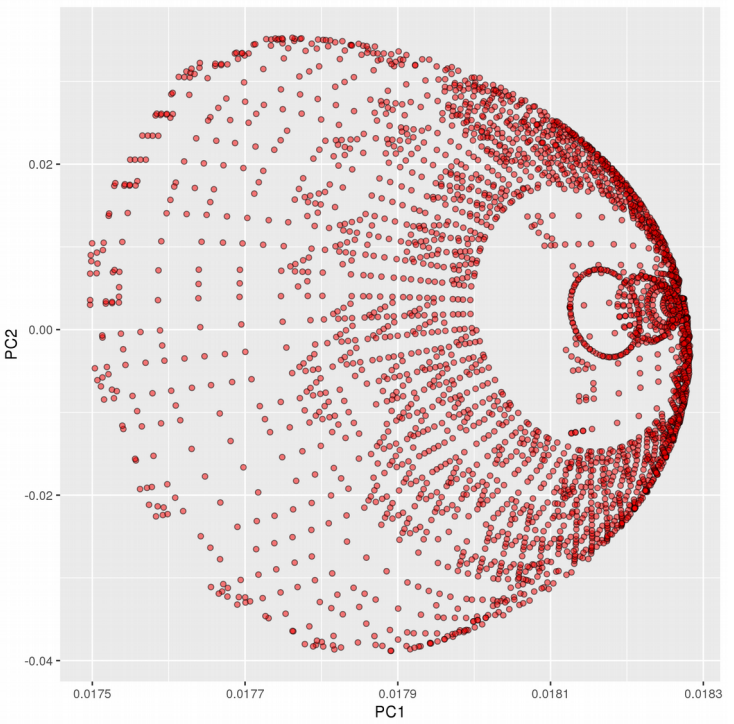
PCA (Cont.)

3D model



Sampled points

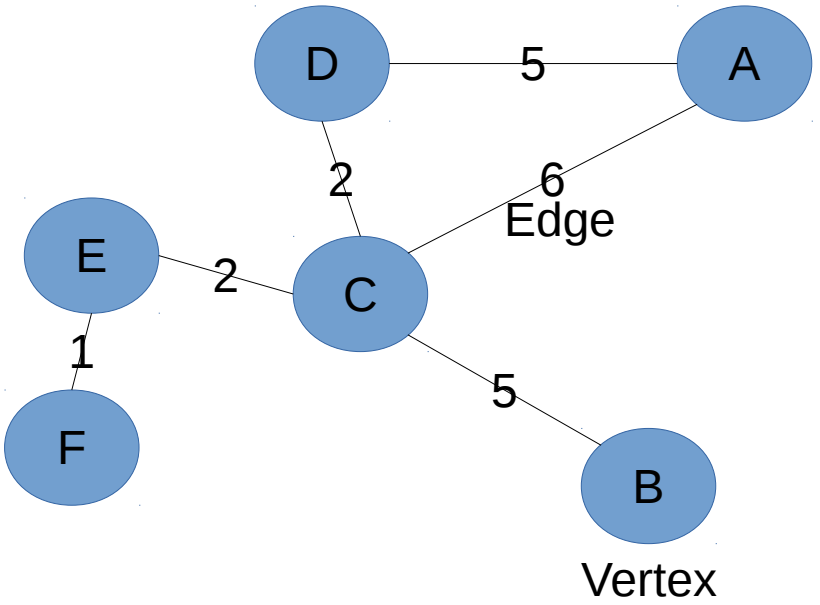
PCA



Multidimensional scaling (MDS)

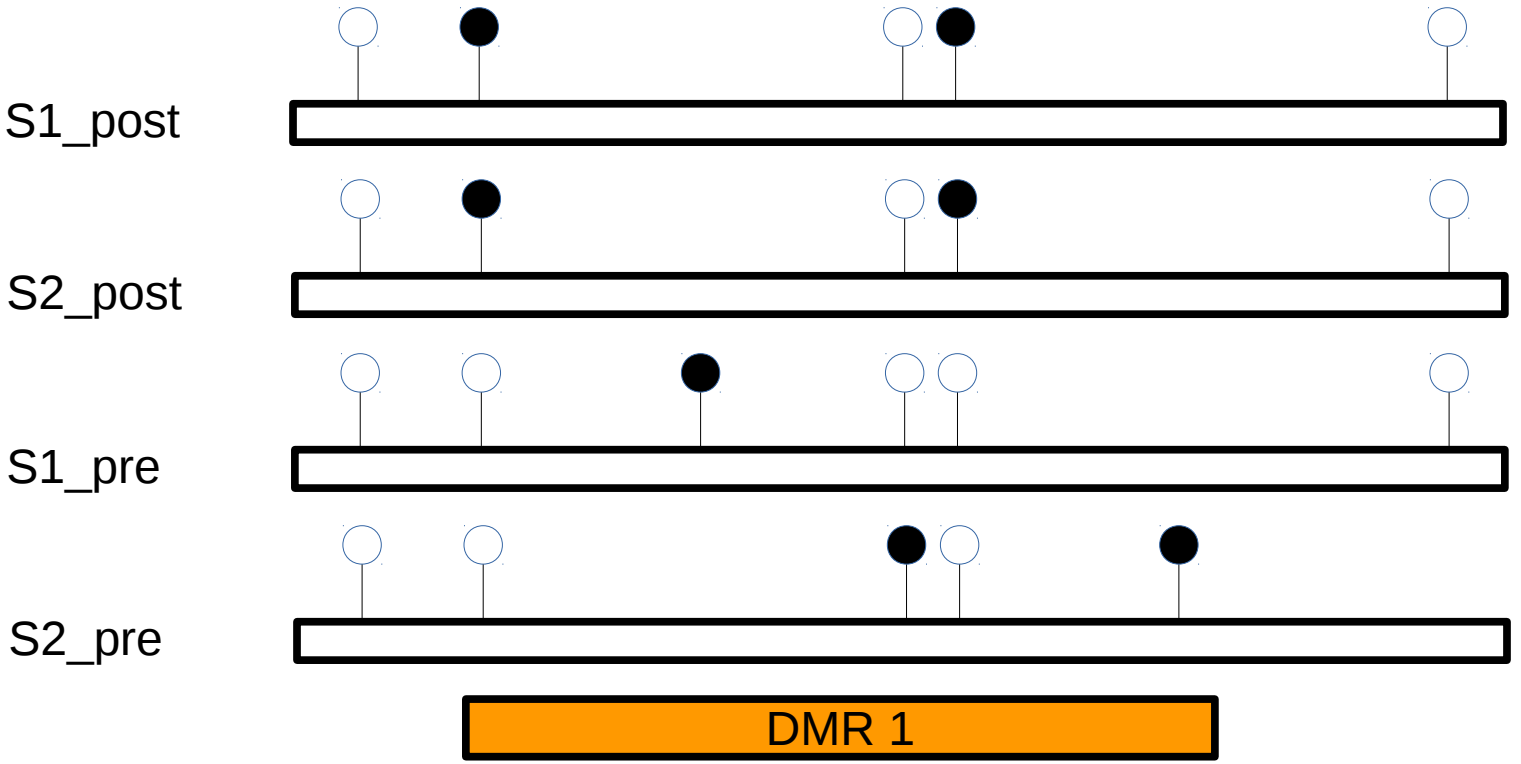
- Another common dimension reduction technique.
- Projects data of high dimension to lower dimension thereby preserving distances (similarity) between observation as much as possible.

Graph



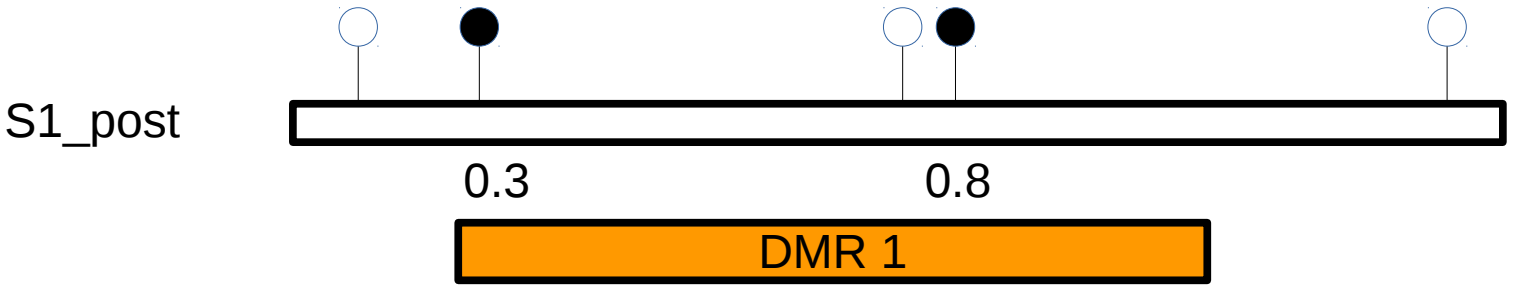
Correlation graph

- DMR



Correlation graph

- DMR average methylation

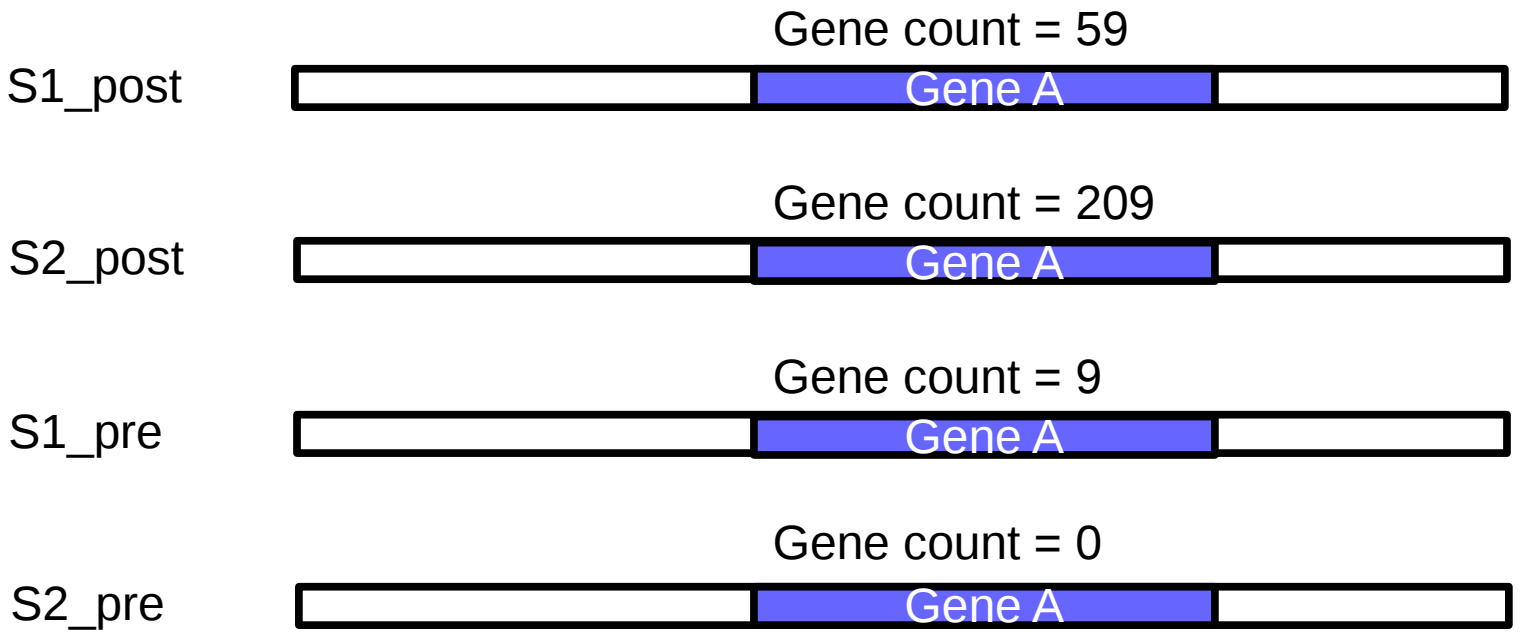


Avg. meth for DMR1 in S1_post = $\frac{(0.3 + 0.8)}{2}$

				S1_post	S1_pre	S2_post	S2_pre	S3_post
DMR 1	Chr1	100,289	109289	0.55	NA	0.37	0.16	0.89
DMR 2	Chr1	456,869	458,850	0.77	0.32	NA	NA	0.35

Correlation graph

- DEG



Gene A is differentially expressed

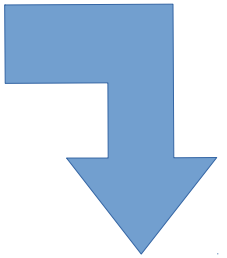
Correlation graph

DEG counts

	S1_post	S1_pre	S2_post	S2_pre	S3_post
Gene A	59	9	209	0	5
Gene B	30	332	9	45	0
Gene C	45	87	9	23	345

DMR avg. meth.

	S1_post	S1_pre	S2_post	S2_pre	S3_post
Chr1_100,289_109289	0.55	NA	0.37	0.16	0.89
Chr1_456,869_458,850	0.77	0.32	NA	NA	0.35



Correlation

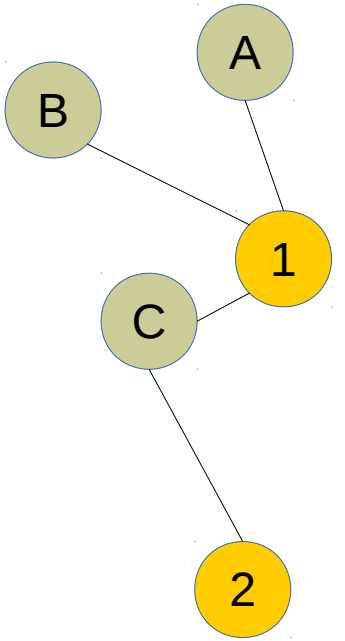
	Gene A	Gene B	Gene C	DMR 1	DMR2
Gene A				0.90	0.78
Genen B				0.60	0.66
Gene C				0.98	0.50
DMR1					
DMR2					

Correlation graph

Correlation matrix

	Gene A	Gene B	Gene C	DMR 1	DMR 2
Gene A				0.96	0.78
Genen B				0.95	0.66
Gene C				0.98	0.96
DMR 1					
DMR 2					

Correlation graph



Resources

- http://www.r-tutor.com/
- https://www.r-bloggers.com/
- https://stackoverflow.com/
- https://stats.stackexchange.com/
- https://ggplot2.tidyverse.org/
- https://github.com/jokergoo/ComplexHeatmap