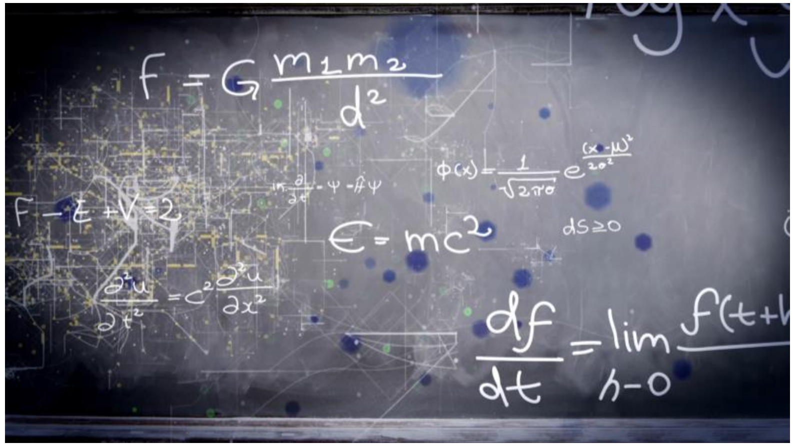
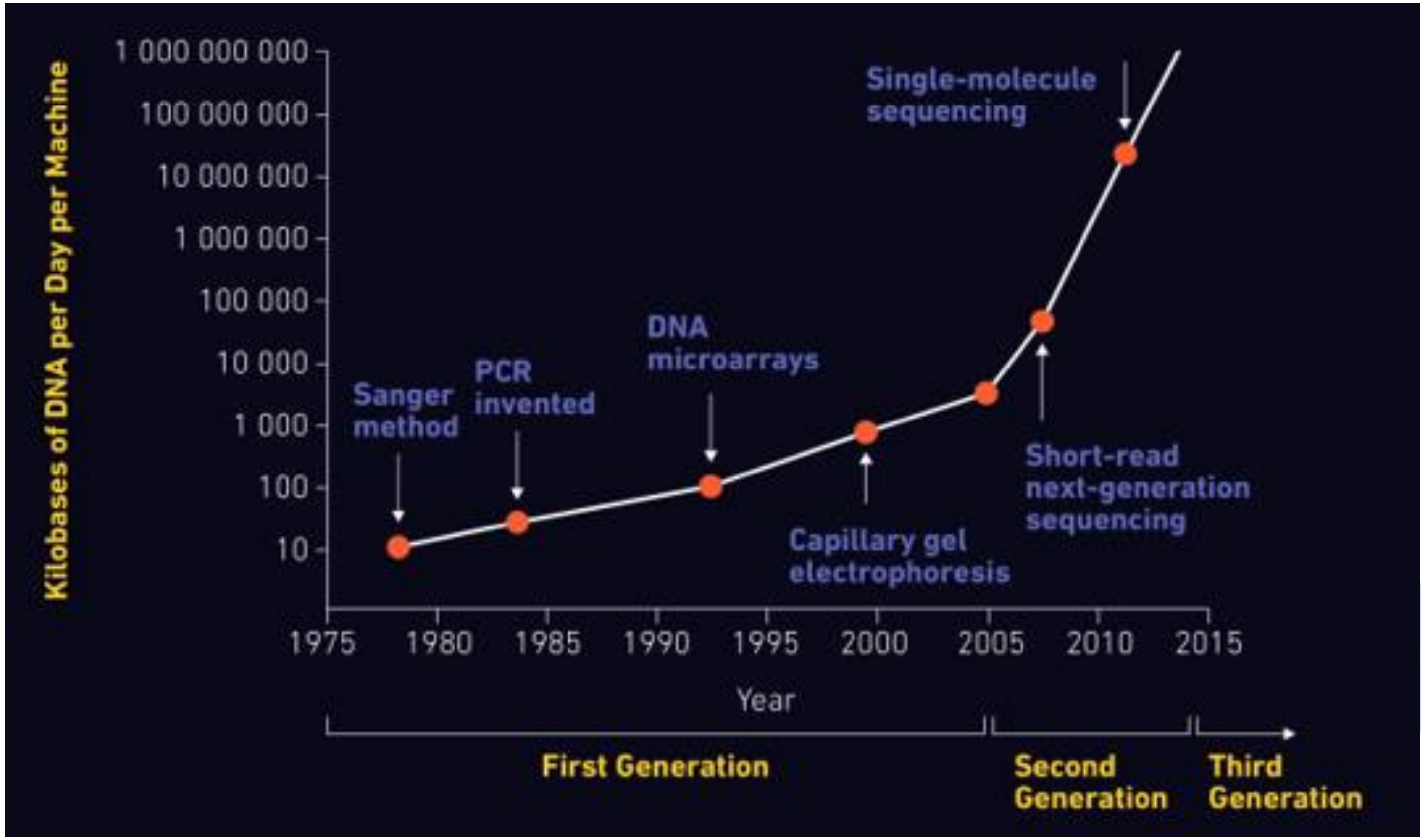
Programming Languages and Bioinformatics

Wisam Hazim Salo

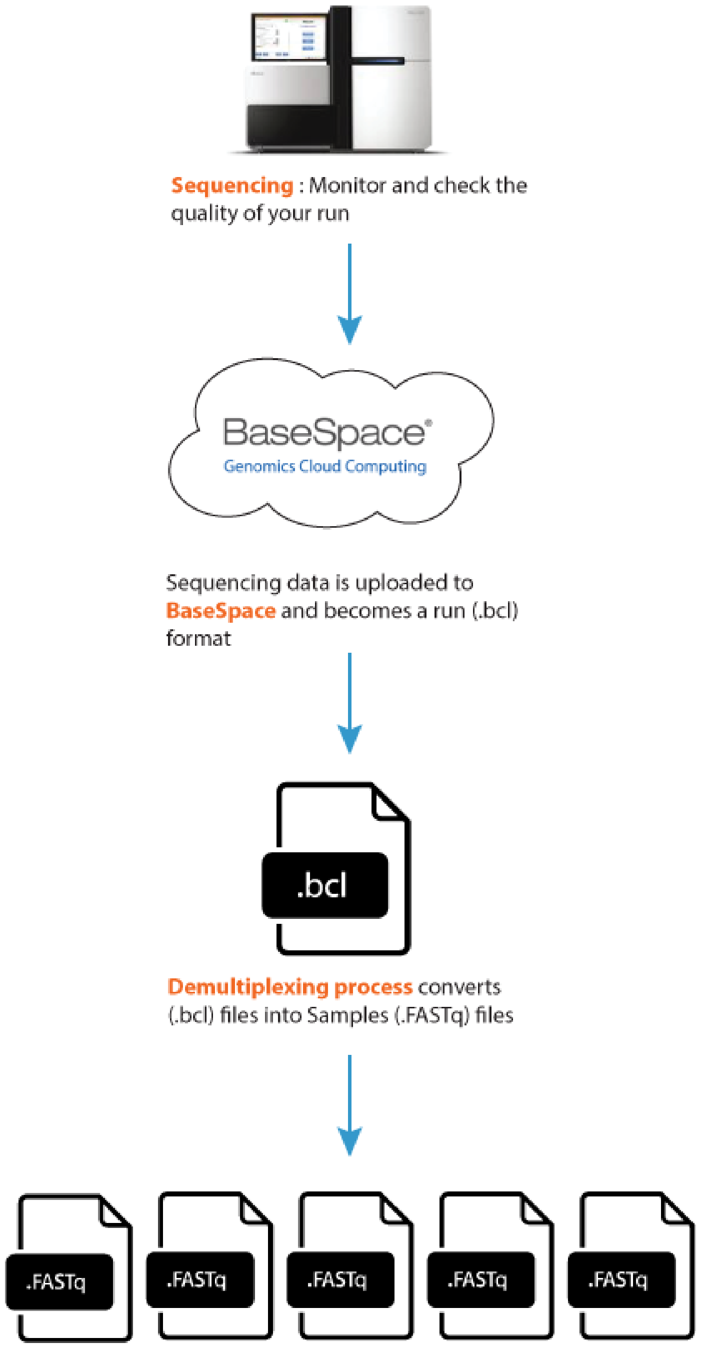


Introduction

- As the Next generation sequencing (NGS) emerged, biological data sets have become larger, the requirements for analysis have also become more complex.



Introduction



Sequencing : Monitor and check the quality of your run

Sequencing data is uploaded to **BaseSpace** and becomes a run (.bcl) format

Demultiplexing process converts (.bcl) files into Samples (.FASTq) files

Hence each FASTQ file represent one sample and ranging between 5-50 Gb’s

Obviously, that’s need unconventional computer skills and methods.

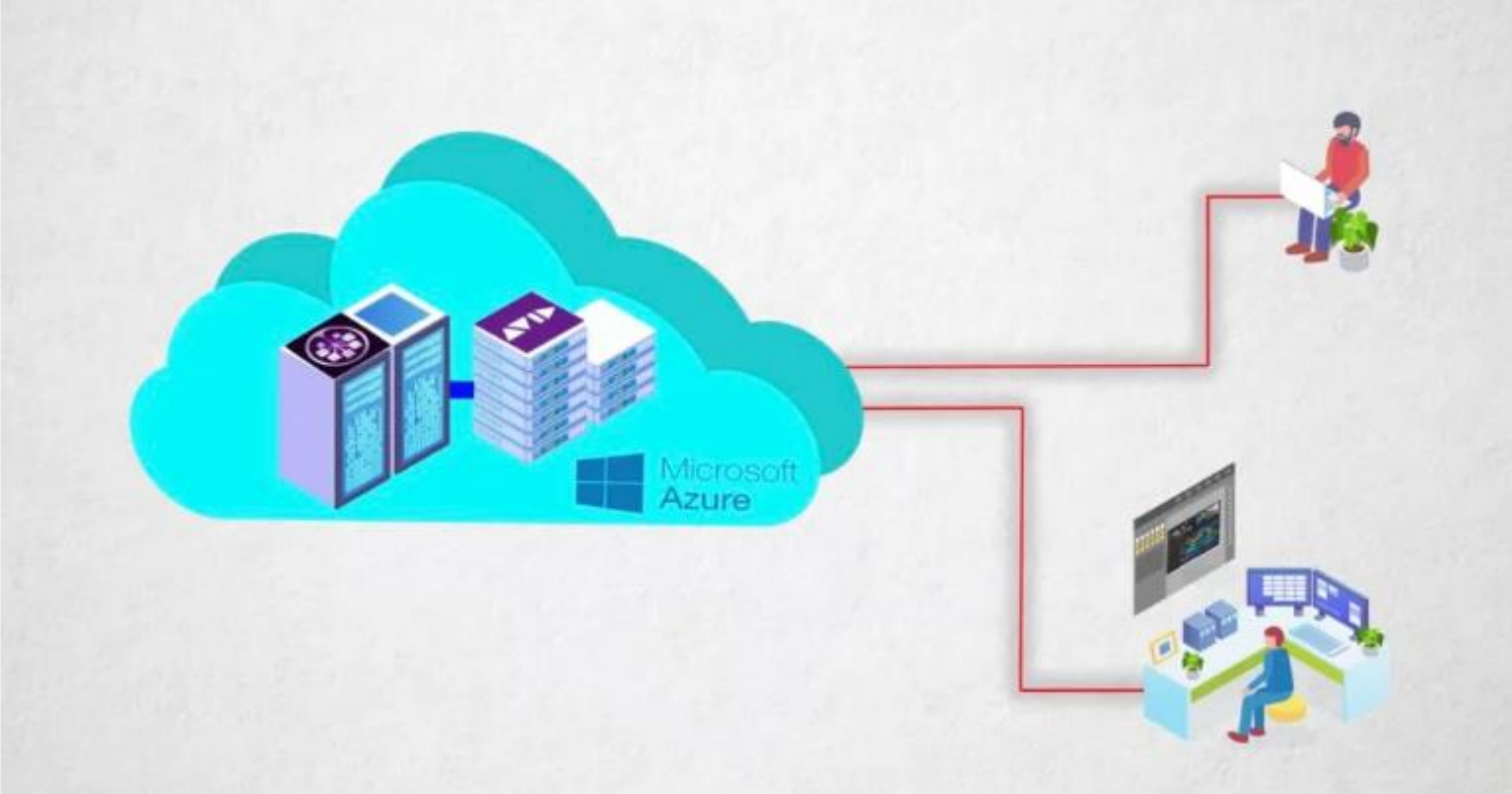
THAT’S BECAUSE OF THE DATA SIZE QUITE BIG TO BE MANAGED IN NORMAL COMPUTERS.

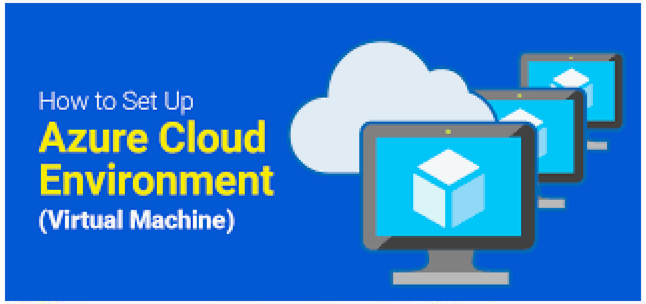
USUALLY, THE MAXIMUM POWER FOR LAPTOP PROCESSORS ARE QUADE CORES AND OCTA CORE IN SOME DESKTOP.

Okay what we have to do then

Cloud Computing

- The Solution will be solved using Cloud Virtual Machine





- That's it, but what is the roll of programming languages?

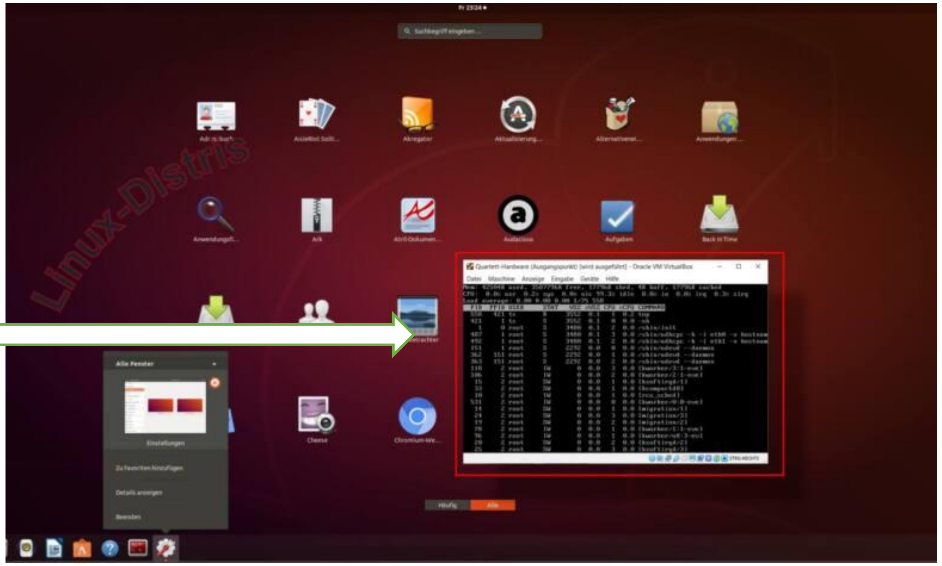
Linux Command Line

- The Virtual Machin Computers generally use the Linux/ Unix operating system and specifically Bash language Command line

Linux has:

Graphic User Interface GUI

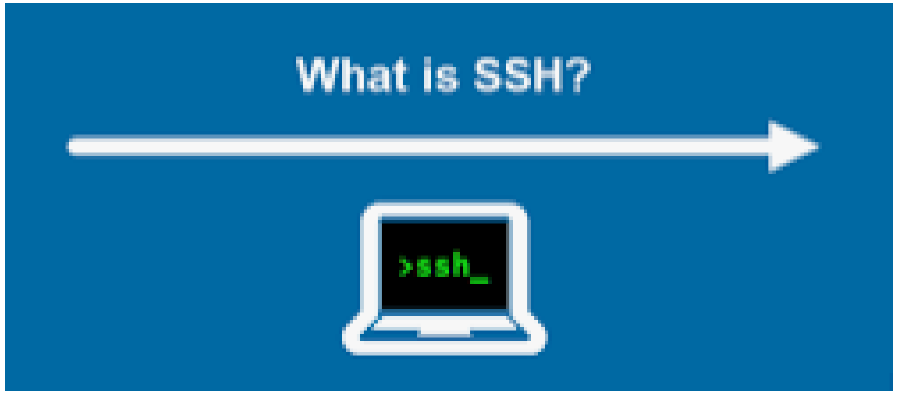
and Command Line CL



- But why?

Why Linux command line?

- 1- To log to the cloud computers (Virtual Machine), we need a Secure shell (SSH Key)

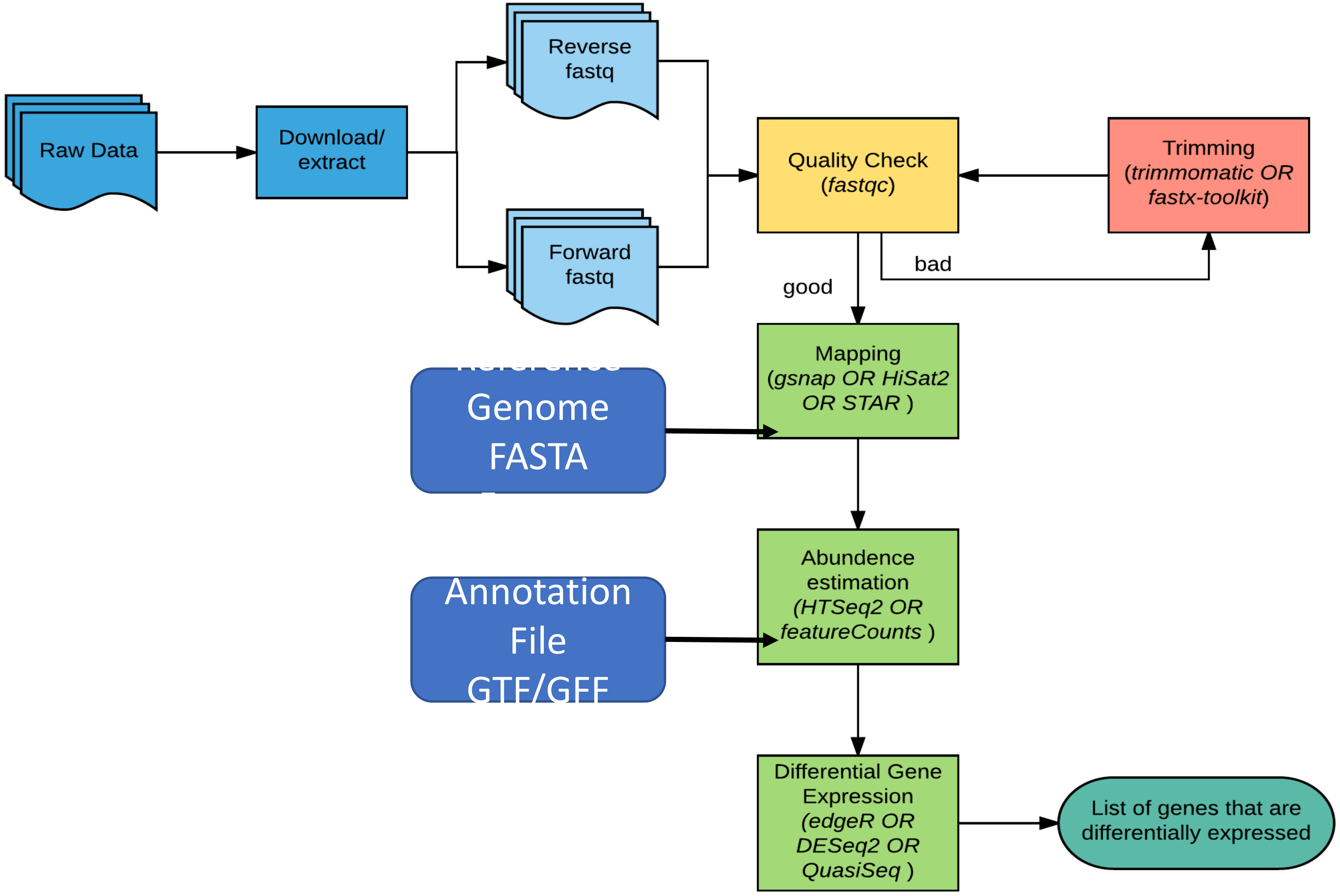


- 2. preparing our environment and setting the pipelines

a. Installing all the algorithms that used in bioinformatic analysis depending on the sequencing data.

b. Arranging the algorithms in a workflow to achieve the analysis

Workfllow or pipe line



Raw Data

Download/ extract

Reverse fastq

Forward fastq

Quality Check (*fastqc*)

Trimming (*trimmomatic OR fastx-toolkit*)

bad

good

Mapping (*gsnap OR HiSat2 OR STAR*)

Genome FASTA

Annotation File GTF/GFF

Abundence estimation (*HTSeq2 OR featureCounts*)

Differential Gene Expression (*edgeR OR DESeq2 OR QuasiSeq*)

List of genes that are differentially expressed

- 3. preparing and organizing the data

Programming language

So why we will use programming languages, and what are the mostly common once?

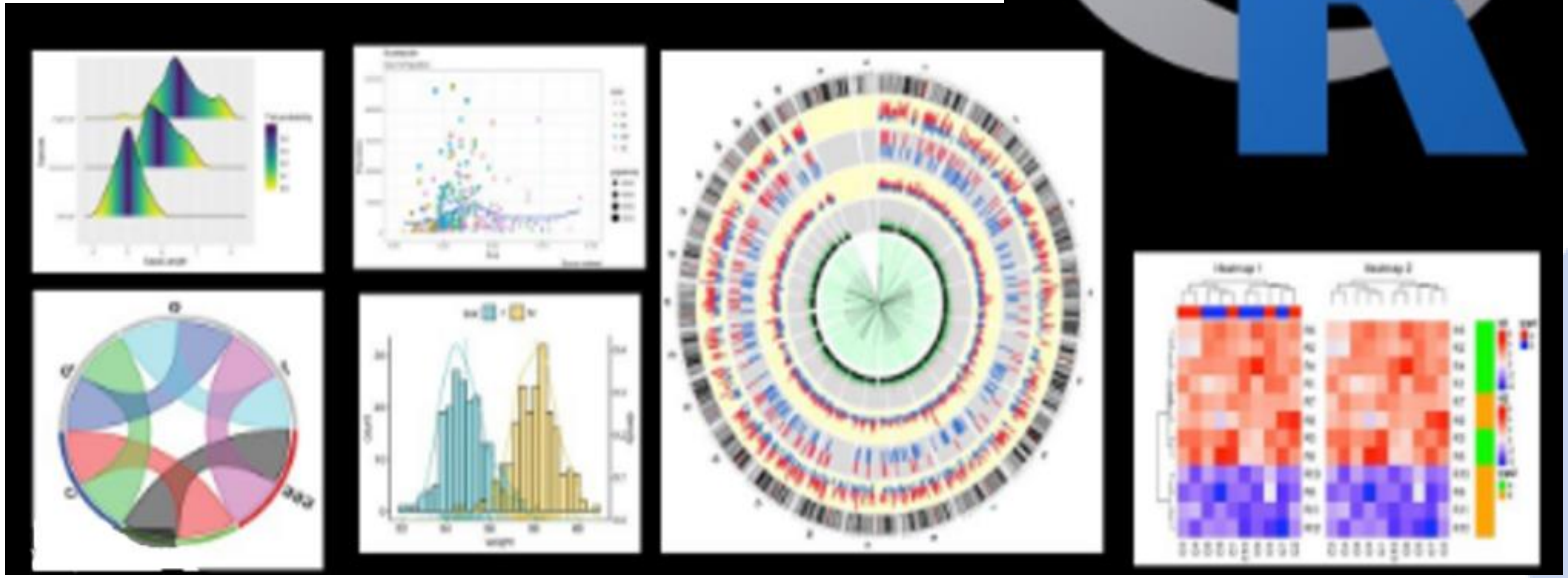
R Language

R is the main programming language that tandemly used in bioinformatics analysis.

It’s been invented in 2000.

Normally used in biological data analysis:

Statistics
visualisation



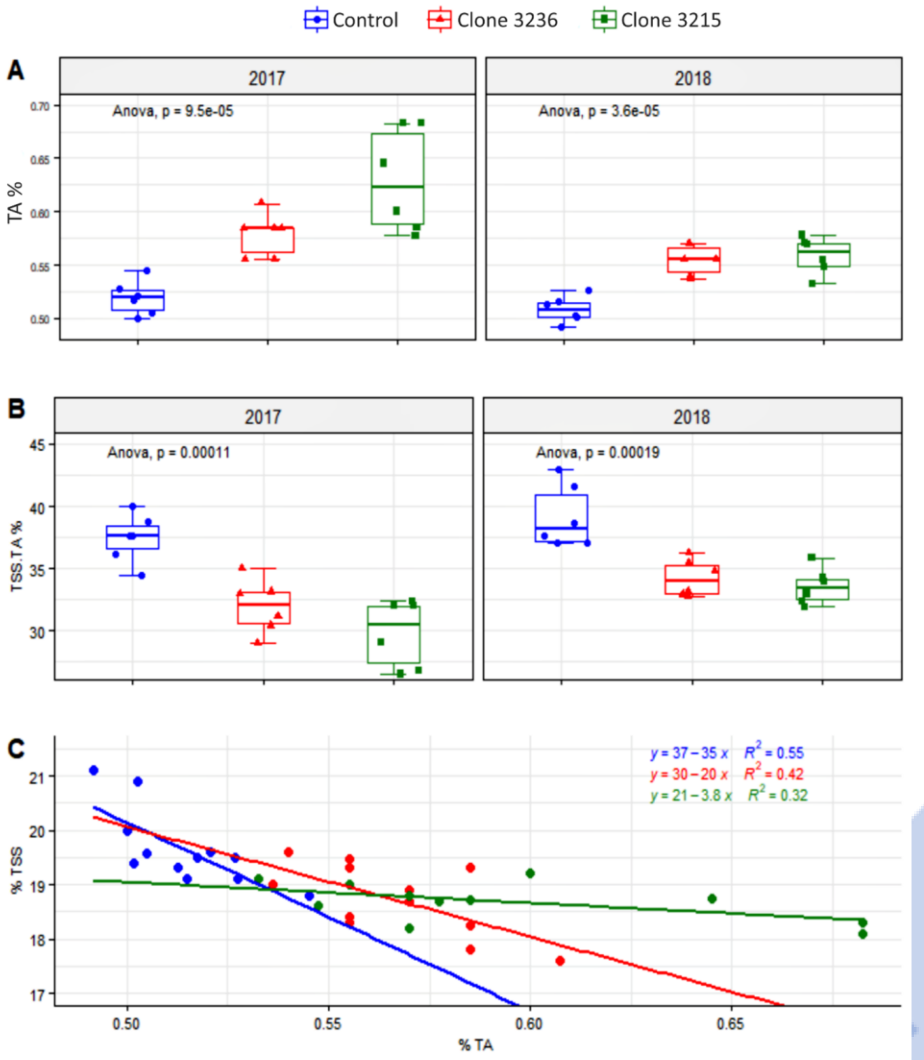
R language

- In addition to the R basic language there is many packages can be added to accomplished more sophisticated analysis:
- ggplot2 for plotting graphs
- dplyer for arranging and manipulating data
- Markdown for editing and composing complicated text

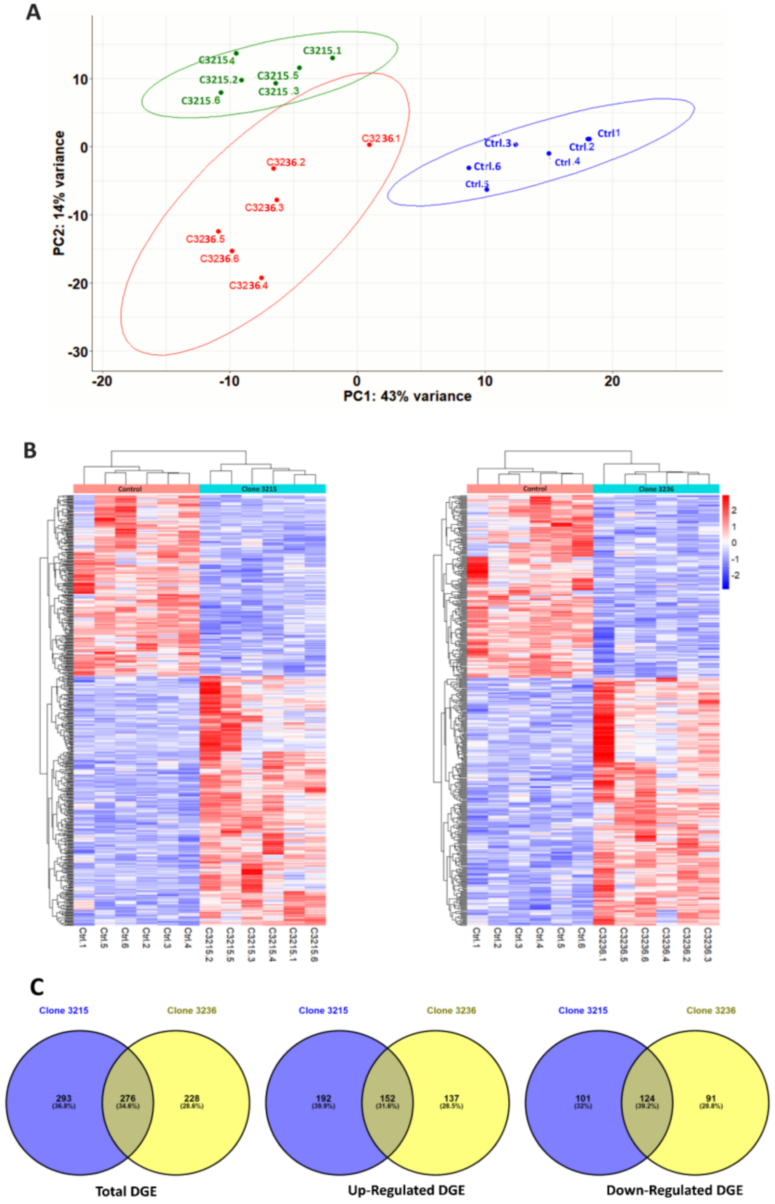
- Moreover, there is a big library of bioinformatics packages known as: Bioconductor

R Language application

- 1- Statistic analysis and visualisation

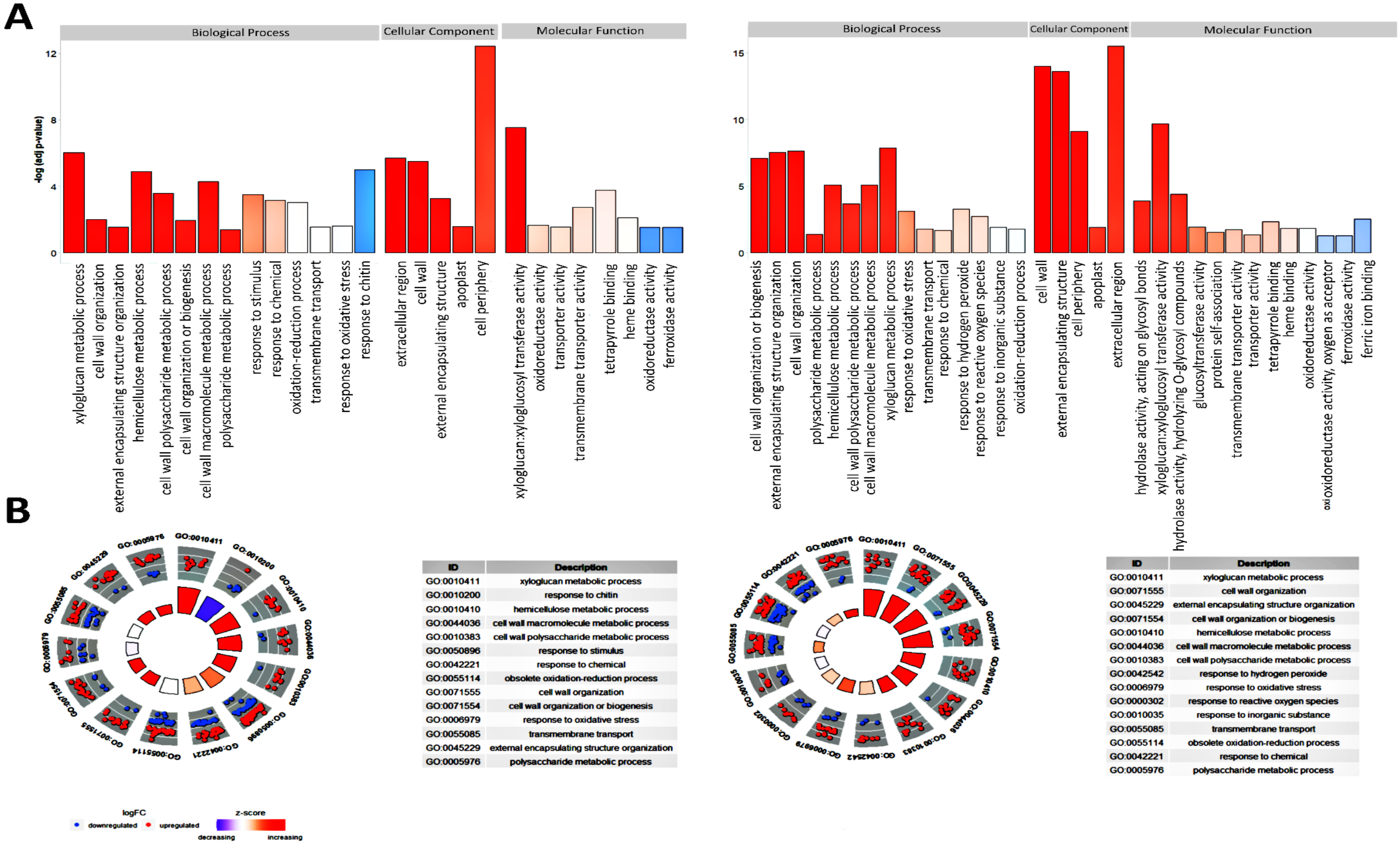


- 2- Clustering of the treatments



R Language application

- 3- Dashboard preparing the for figures and tables

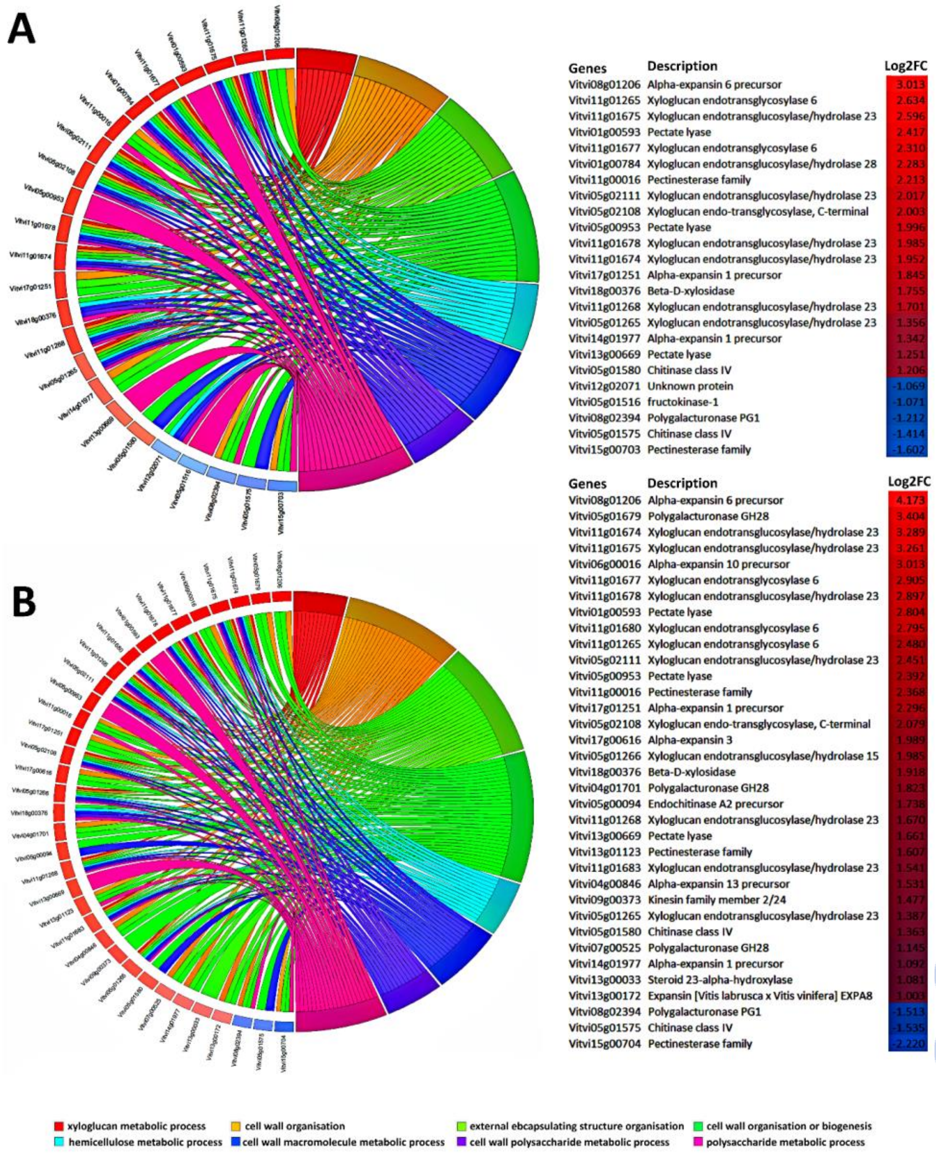


ID	Description
GO:0010411	xyloglucan metabolic process
GO:0010200	response to chitin
GO:0010410	hemicellulose metabolic process
GO:0044036	cell wall macromolecule metabolic process
GO:0010383	cell wall polysaccharide metabolic process
GO:0050896	response to stimulus
GO:0042221	response to chemical
GO:0055114	obsolete oxidation-reduction process
GO:0071555	cell wall organization
GO:0071554	cell wall organization or biogenesis
GO:0006979	response to oxidative stress
GO:0055085	transmembrane transport
GO:0045229	external encapsulating structure organization
GO:0005976	polysaccharide metabolic process

ID	Description
GO:0010411	xyloglucan metabolic process
GO:0071555	cell wall organization
GO:0045229	external encapsulating structure organization
GO:0071554	cell wall organization or biogenesis
GO:0010410	hemicellulose metabolic process
GO:0044036	cell wall macromolecule metabolic process
GO:0010383	cell wall polysaccharide metabolic process
GO:0042542	response to hydrogen peroxide
GO:0006979	response to oxidative stress
GO:0000302	response to reactive oxygen species
GO:0010035	response to inorganic substance
GO:0055085	transmembrane transport
GO:0055114	obsolete oxidation-reduction process
GO:0042221	response to chemical
GO:0005976	polysaccharide metabolic process

R Language application

4- Chord clustering analysis and figures

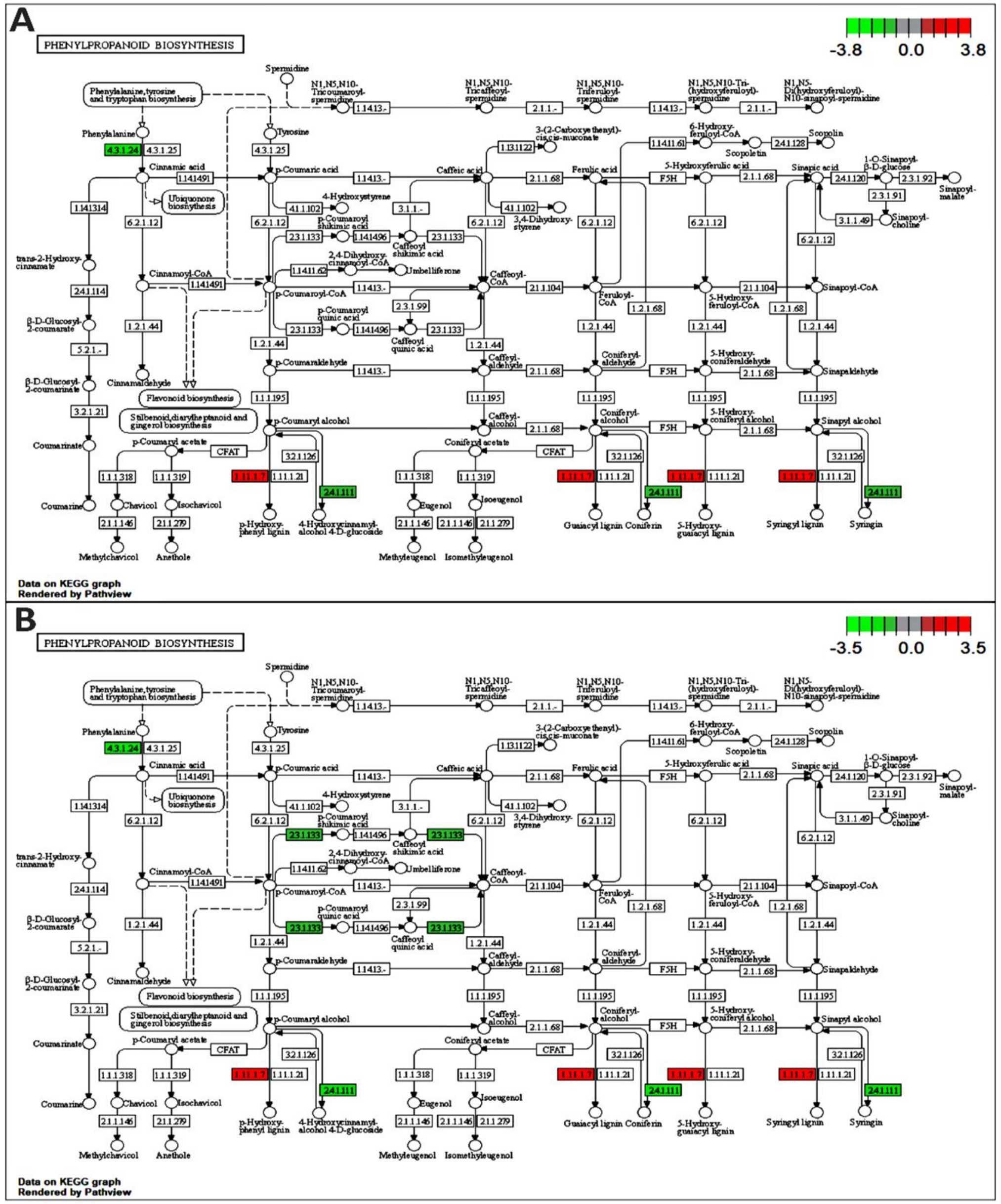


Genes	Description	Log2FC
Vitvi08g01206	Alpha-expansin 6 precursor	3.013
Vitvi11g01265	Xyloglucan endotransglycosylase 6	2.634
Vitvi11g01675	Xyloglucan endotransglucosylase/hydrolase 23	2.596
Vitvi01g00593	Pectate lyase	2.417
Vitvi11g01677	Xyloglucan endotransglycosylase 6	2.310
Vitvi01g00784	Xyloglucan endotransglucosylase/hydrolase 28	2.283
Vitvi11g00016	Pectinesterase family	2.213
Vitvi05g02111	Xyloglucan endotransglucosylase/hydrolase 23	2.017
Vitvi05g02108	Xyloglucan endo-transglycosylase, C-terminal	2.003
Vitvi05g00953	Pectate lyase	1.996
Vitvi11g01678	Xyloglucan endotransglucosylase/hydrolase 23	1.985
Vitvi11g01674	Xyloglucan endotransglucosylase/hydrolase 23	1.952
Vitvi17g01251	Alpha-expansin 1 precursor	1.845
Vitvi18g00376	Beta-D-xylosidase	1.755
Vitvi11g01268	Xyloglucan endotransglucosylase/hydrolase 23	1.701
Vitvi05g01265	Xyloglucan endotransglucosylase/hydrolase 23	1.356
Vitvi14g01977	Alpha-expansin 1 precursor	1.342
Vitvi13g00669	Pectate lyase	1.251
Vitvi05g01580	Chitinase class IV	1.206
Vitvi12g02071	Unknown protein	-1.069
Vitvi05g01516	fructokinase-1	-1.071
Vitvi08g02394	Polygalacturonase PG1	-1.212
Vitvi05g01575	Chitinase class IV	-1.414
Vitvi15g00703	Pectinesterase family	-1.602

Genes	Description	Log2FC
Vitvi08g01206	Alpha-expansin 6 precursor	4.173
Vitvi05g01679	Polygalacturonase GH28	3.404
Vitvi11g01674	Xyloglucan endotransglucosylase/hydrolase 23	3.289
Vitvi11g01675	Xyloglucan endotransglucosylase/hydrolase 23	3.261
Vitvi06g00016	Alpha-expansin 10 precursor	3.013
Vitvi11g01677	Xyloglucan endotransglycosylase 6	2.905
Vitvi11g01678	Xyloglucan endotransglucosylase/hydrolase 23	2.897
Vitvi01g00593	Pectate lyase	2.804
Vitvi11g01680	Xyloglucan endotransglycosylase 6	2.795
Vitvi11g01265	Xyloglucan endotransglycosylase 6	2.480
Vitvi05g02111	Xyloglucan endotransglucosylase/hydrolase 23	2.451
Vitvi05g00953	Pectate lyase	2.392
Vitvi11g00016	Pectinesterase family	2.368
Vitvi17g01251	Alpha-expansin 1 precursor	2.296
Vitvi05g02108	Xyloglucan endo-transglycosylase, C-terminal	2.079
Vitvi17g00616	Alpha-expansin 3	1.989
Vitvi05g01266	Xyloglucan endotransglucosylase/hydrolase 15	1.985
Vitvi18g00376	Beta-D-xylosidase	1.918
Vitvi04g01701	Polygalacturonase GH28	1.823
Vitvi05g00094	Endochitinase A2 precursor	1.738
Vitvi11g01268	Xyloglucan endotransglucosylase/hydrolase 23	1.670
Vitvi13g00669	Pectate lyase	1.661
Vitvi13g01123	Pectinesterase family	1.607
Vitvi11g01683	Xyloglucan endotransglucosylase/hydrolase 23	1.541
Vitvi04g00846	Alpha-expansin 13 precursor	1.531
Vitvi09g00373	Kinesin family member 2/24	1.477
Vitvi05g01265	Xyloglucan endotransglucosylase/hydrolase 23	1.387
Vitvi05g01580	Chitinase class IV	1.363
Vitvi07g00525	Polygalacturonase GH28	1.145
Vitvi14g01977	Alpha-expansin 1 precursor	1.092
Vitvi13g00033	Steroid 23-alpha-hydroxylase	1.081
Vitvi13g00172	Expansin [Vitis labrusca x Vitis vinifera] EXPA8	1.003
Vitvi08g02394	Polygalacturonase PG1	-1.513
Vitvi05g01575	Chitinase class IV	-1.535
Vitvi15g00704	Pectinesterase family	-2.220

R Language application

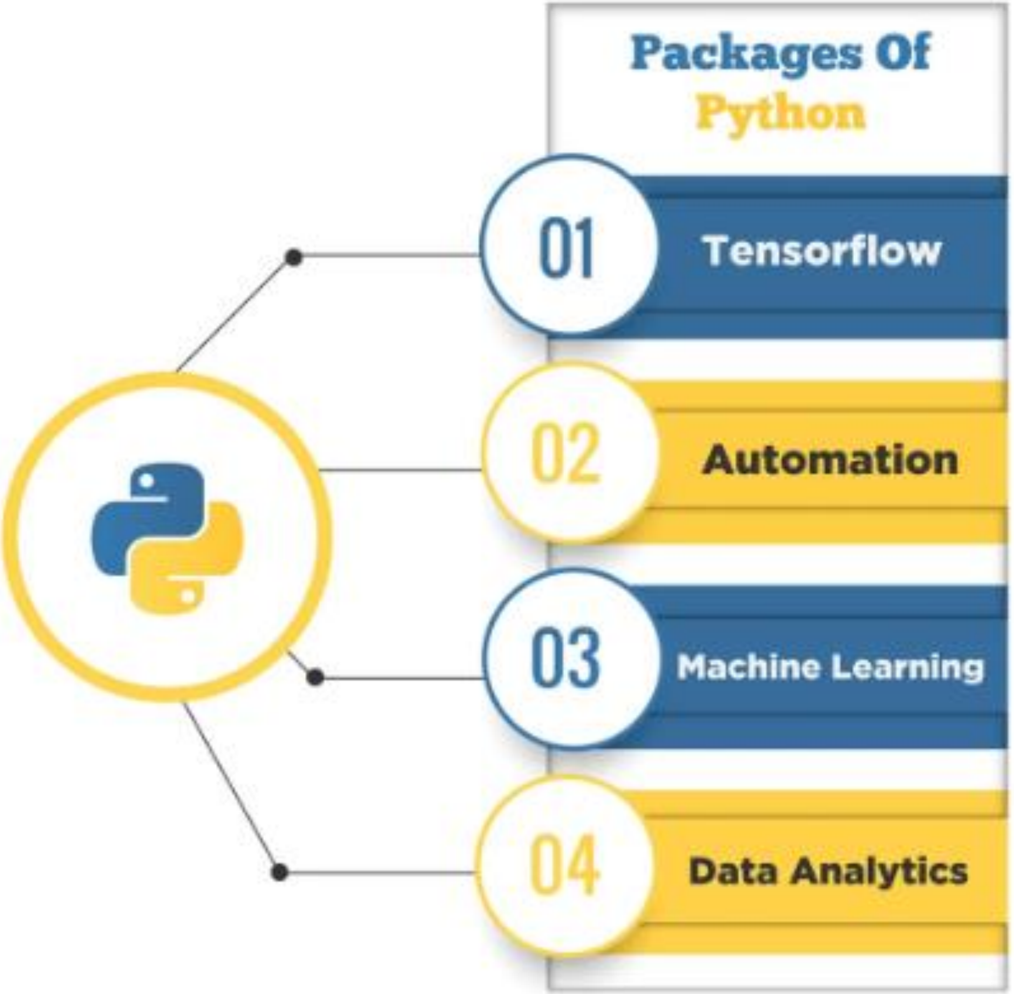
5- Pathway analysis



- Is there other language we should learn and why?

Python

- Python consider the jocker of programming language
- It can be used to replace any other language
- Normally used
  - Web Development
  - Game and apps Development
  - Machine Learning and Artificial Intelligence
  - Data Science and Data Visualization
  - Business Applications
  - Audio and Video Applications
  - CAD Applications



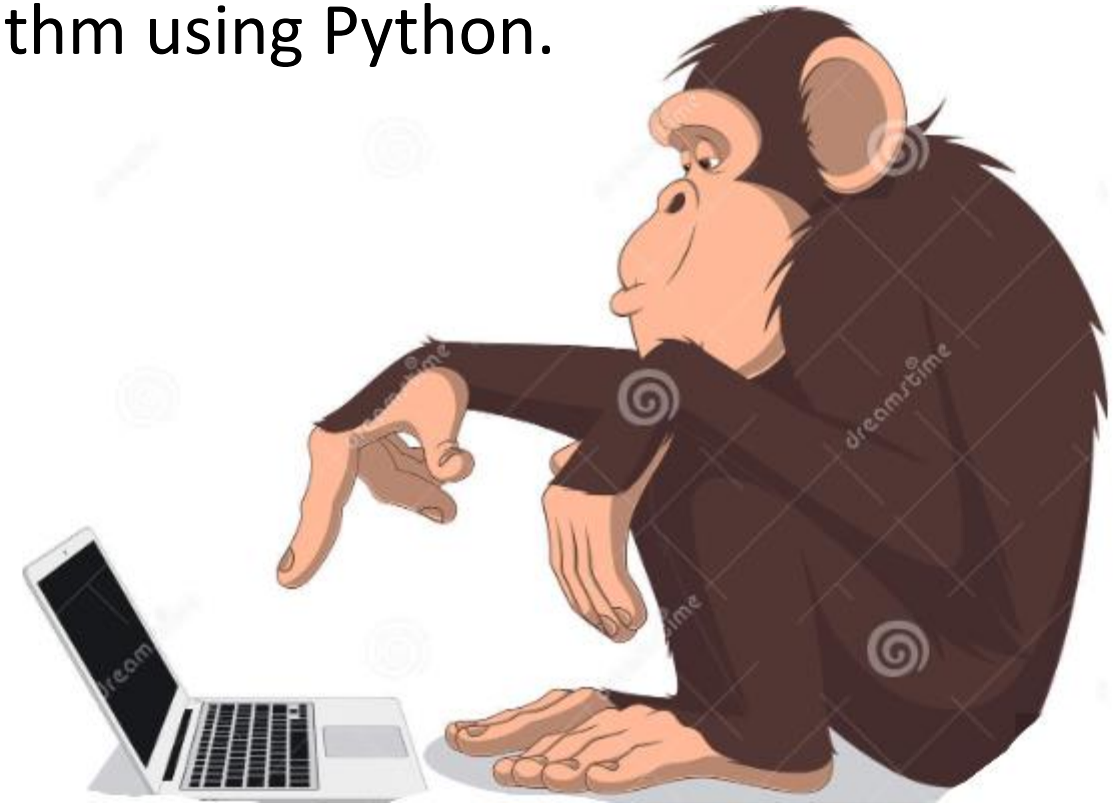
Python Language

- In bioinformatics, python applied in both machine learning and data analysis.

- Python usually installed on Linux environment because most of the used algorithms were written in Python.

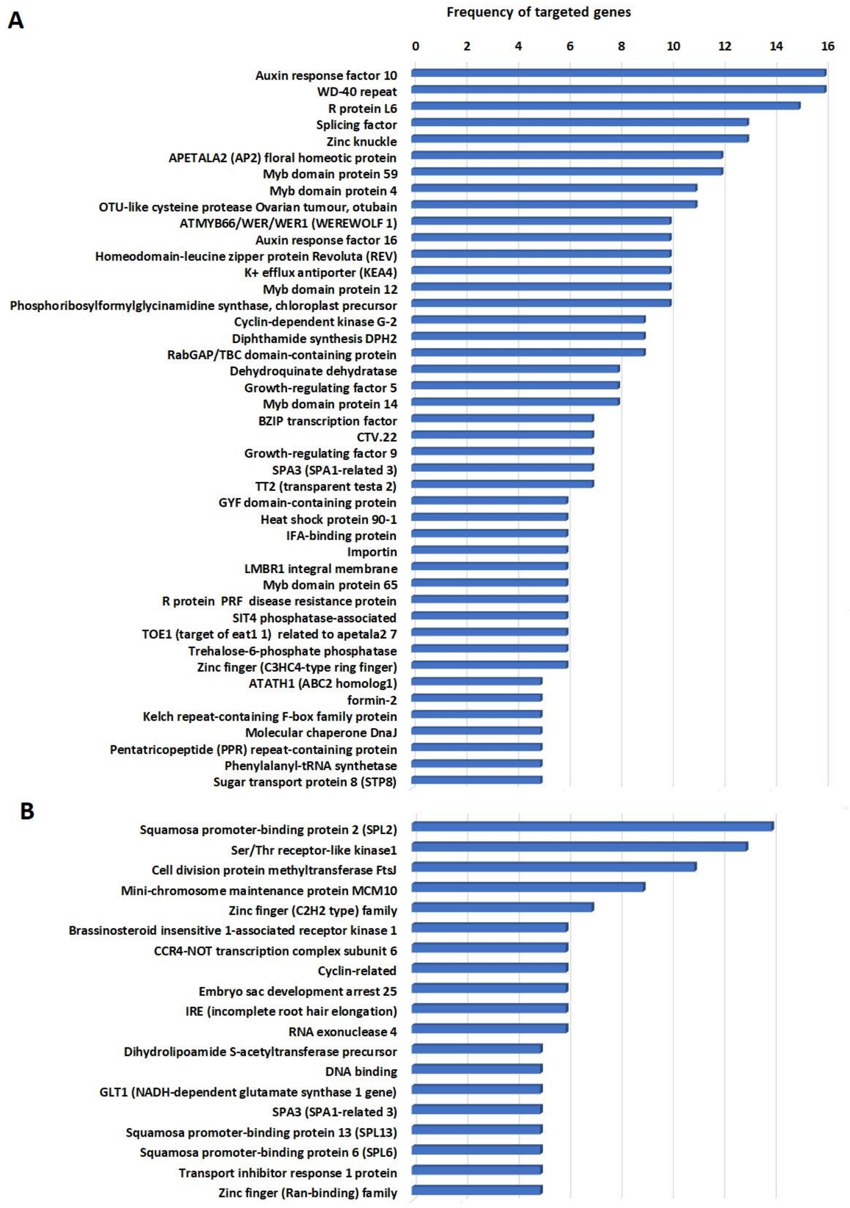
Python

In addition sometime we can’t find appropriated algorithm for a specific analysis such as those used with the small RNA data. Therefore Python used to create our own algorithm using Python.



Python Language

In this figure an algorithm has been developed to count the frequency of targeted genes by the microRNA using Python



Thanks for your time