Transcriptomics data analysis

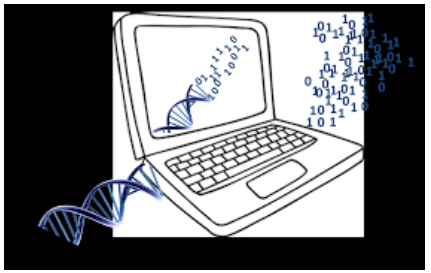
Genomic Networks Team

POPS

Concevoir en amont pour analyser en aval

Organellar Gene Expression Team

a standardized and automated workflow for RNA sequencing data analysis



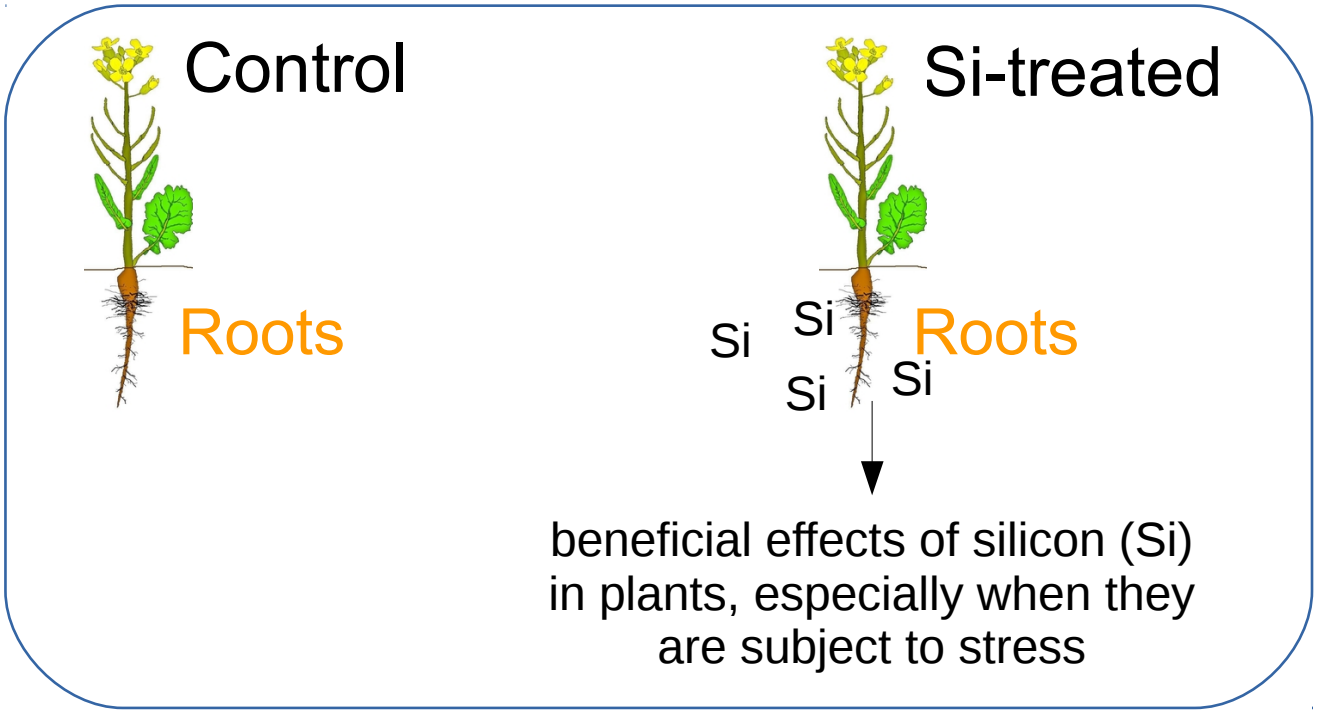
Christine Paysant-Le Roux

IPS2
Institute of Plant Sciences
Paris - Saclay

Transcriptomics data analysis

- An introduction to RNA-seq
- Steps to Illumina sequencing
- BCL to Fastq conversion
- RNA-seq data analysis:
  - ➔ Bioinformatic analysis
  - ➔ Statistical analysis : differential expression analysis

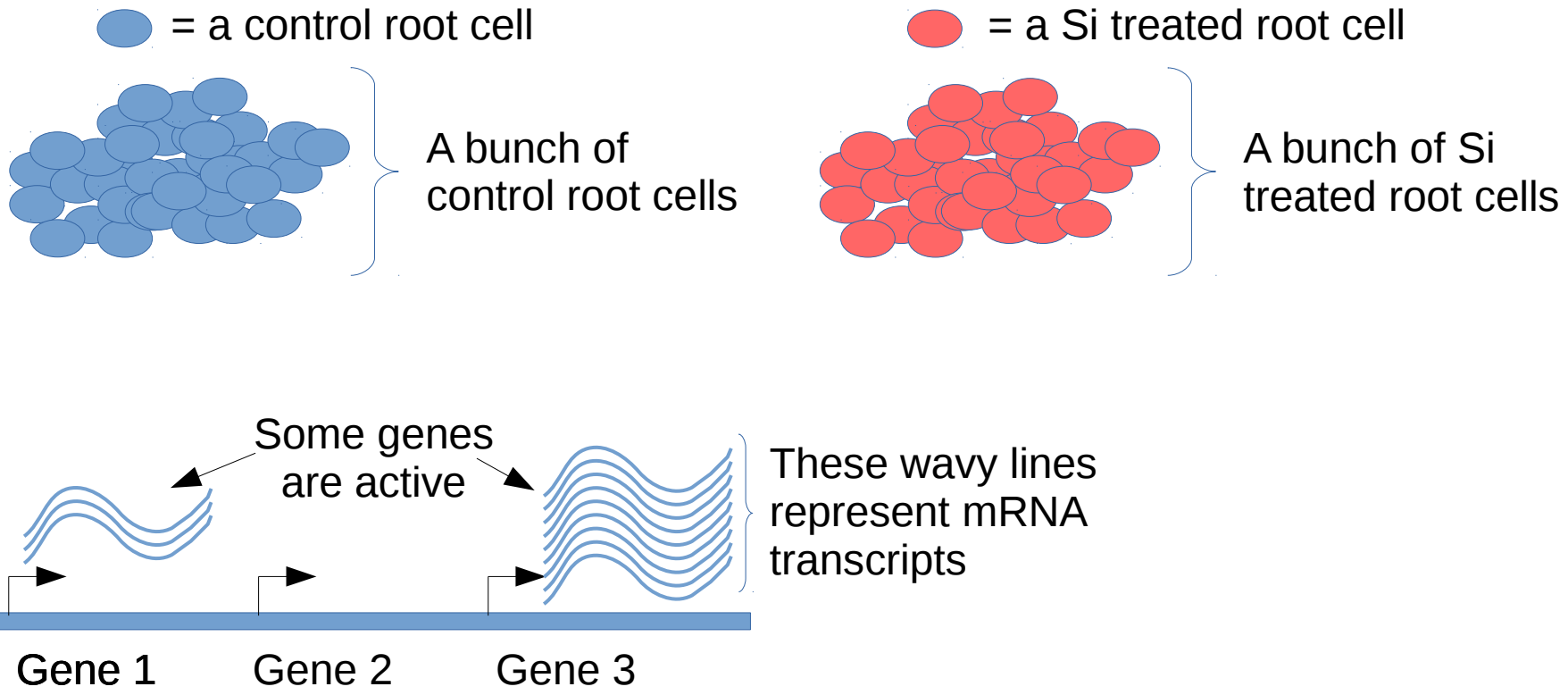
An introduction to RNA-seq



What genetic mechanism is causing the difference ?

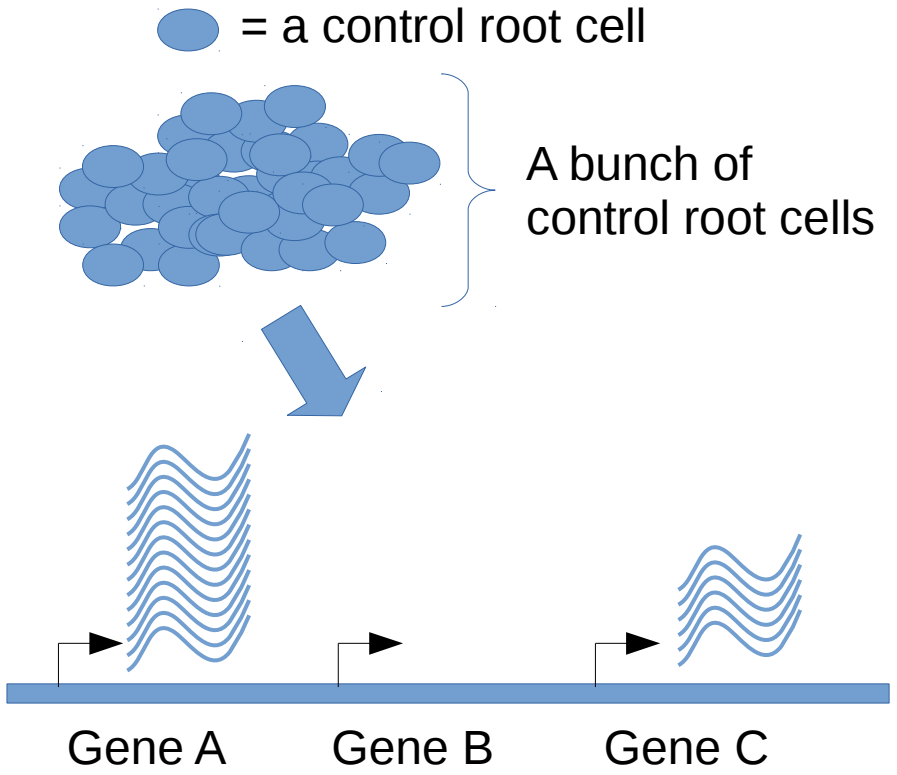
This means whe want to look at differences in gene expression

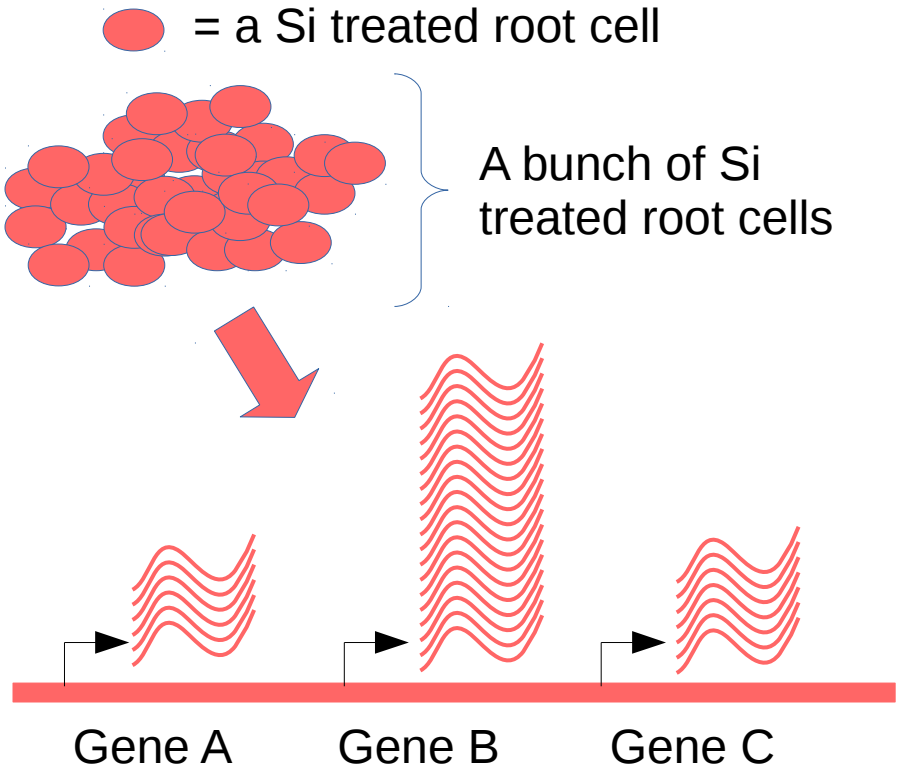
An introduction to RNA-seq



Principe du RNA-seq: abondance des ARNm reflète l'expression des gènes.

An introduction to RNA-seq



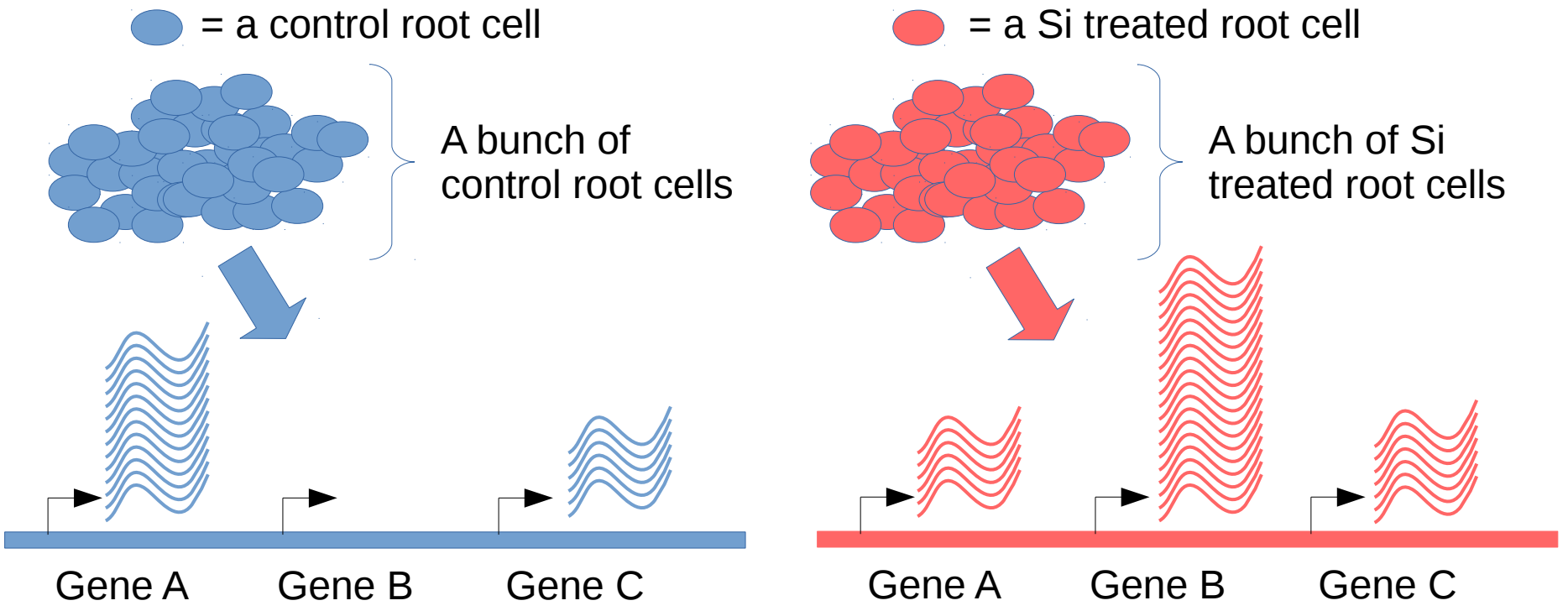


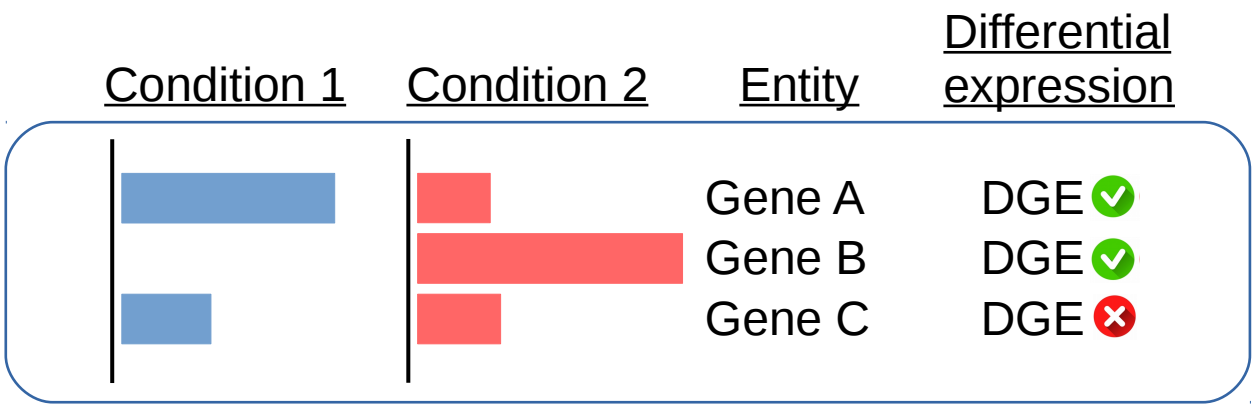
We can use RNA-seq to measure gene expression in control root cells …

… then use it to measure gene expression in Si-treated root cells …

RNA-seq focused primarily on quantifying gene expression between samples in different groups, treatments, time-points, ...

An introduction to RNA-seq





The goal is to identify genes whose expression level changes between conditions

Steps to Illumina sequencing

Step 1: isolate the mRNA

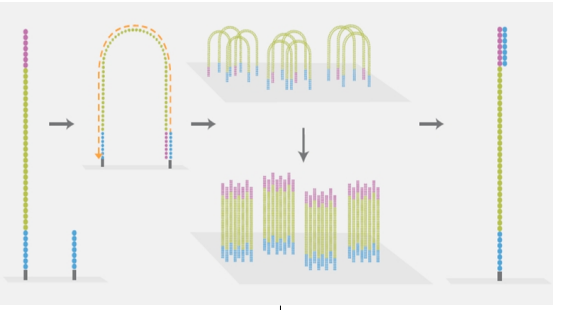
Step 2: break the RNA and convert the RNA fragments into double stranded DNA

Step 3: add adapters and PCR amplify

Library construction involves generating a collection of DNA fragments for sequencing

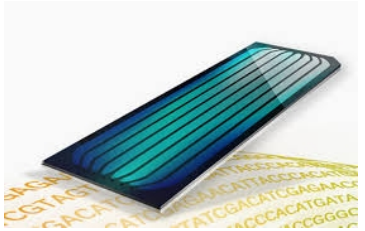
Illumina sequencing

About 400,000,000 fragments laids out vertically in a flow cell



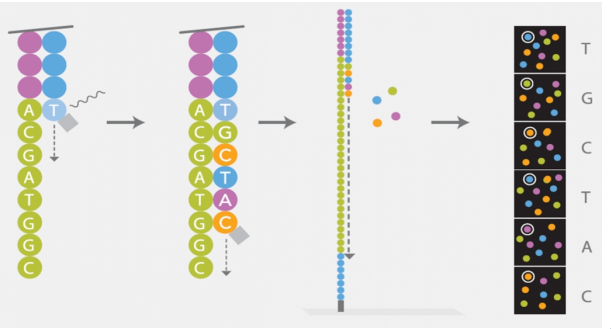
Step 1: cluster generation

- Add to flow cell
- Bridge amplification



Step 2: sequencing

- Single base at a time
- image capture between each nucleotide addition



This cycle is repeated n times to create a read length of n bases

BCL to Fastq conversion

BCL file

Here is an example of a single entry in a FASTQ file :

```
@NB501040:2:HGVWFBGXX:1:11101:9426:1040 1:N:0:GCCAAT
TGAGANTGGGCGTCGTATGGAGATTGTTGAACTACCAAGTCATCCATACTTTGTGGGTGCTCAGTTCCATCCTGAA
+
AAAAA#AEEEEEEEEEEEEEEEEAAEEEE6EEEAEEEEEEE6EAEEEEEAEEEAEEEEEEEEEAEEE//EEE<EEE
```

1. @read name with information about the sequencing run and the cluster.

2. The sequence (the base calls; A, C, T, G and N).

3. A separator, which is simply a plus (+) sign.

4. The base call quality scores encoded as a single byte ASCII characters to represent the numerical quality scores

BCL to Fastq conversion

BCL file

Here is an example of a single entry in a FASTQ file :

```
@NB501040:2:HGVWFBGXX:1:11101:9426:1040 1:N:0:GCCAAT
TGAGANTGGGCGTCGTATGGAGATTGTTGAACTACCAAGTCATCCATACTTTGTGGGTGCTCAGTTCCATCCTGAA
+
AAAAA#AEEEEEEEEEEEEEEEEAAEEEE6EEEAEEEEEEE6EAEEEEEAEEEAEEEEEEEEEAEEE//EEE<EEE
```

1. @read name with information about the sequencing run and the cluster.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores encoded as a single byte ASCII characters to represent the numerical quality scores

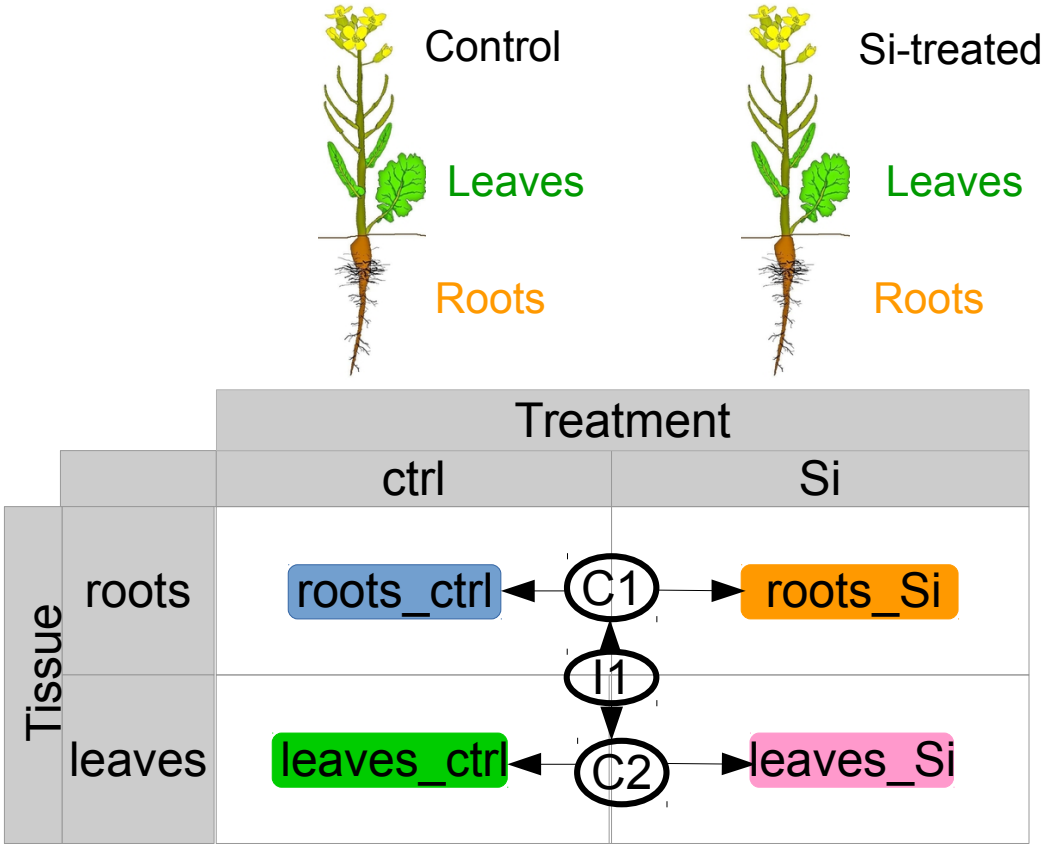
```
                ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I
                |                                 |       |                   |
Code ASCII      33                                59      64                  73
Qscore           0 . . . . . . . . . . . . . . . . . . . . 26 . . . 31 . . . . . . . . 41
```

$Qscore = -10 \log_{10} (Pe)$

Qscore = 20: Pe = 1/100 = 0.01
Qscore = 30: Pe = 1/1000 = 0.01

Images transformed into reads and base call quality scores => fastq file

A example of RNA-seq experiment

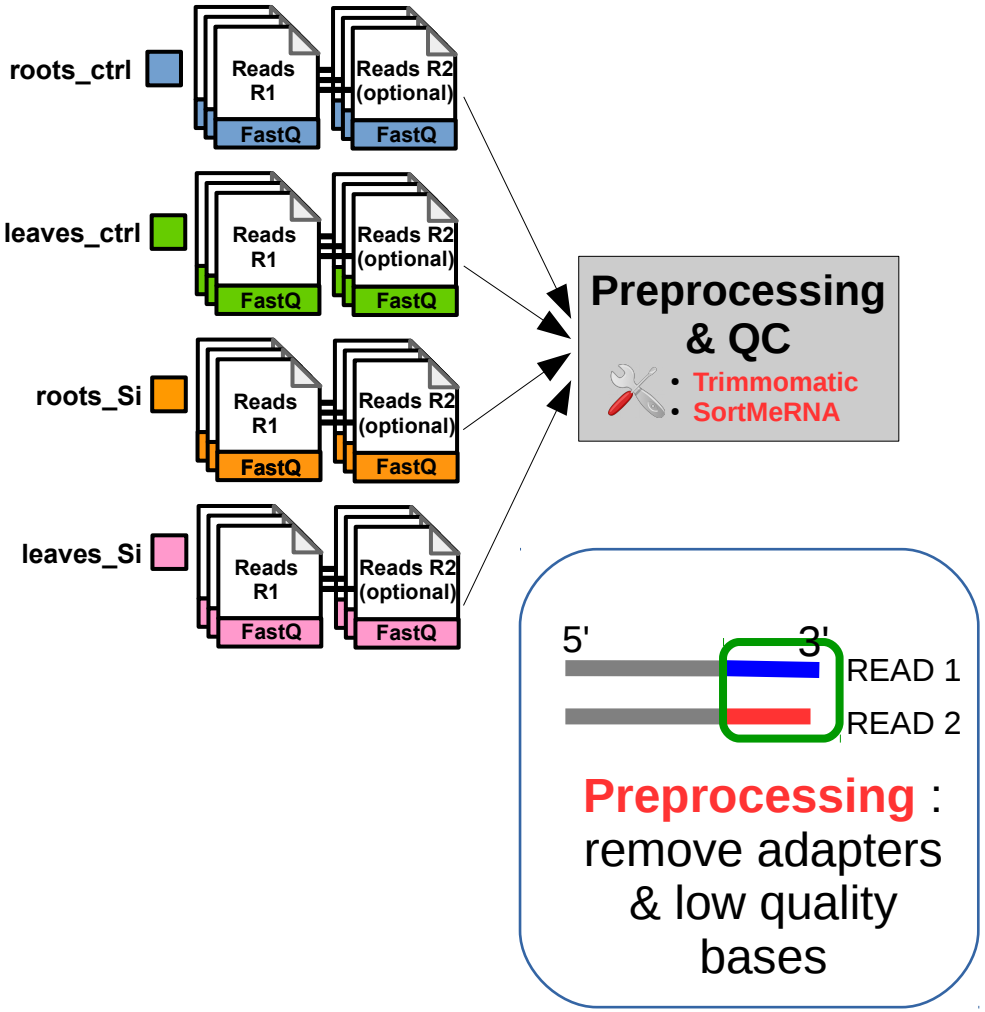


Several contrasts can be tested:

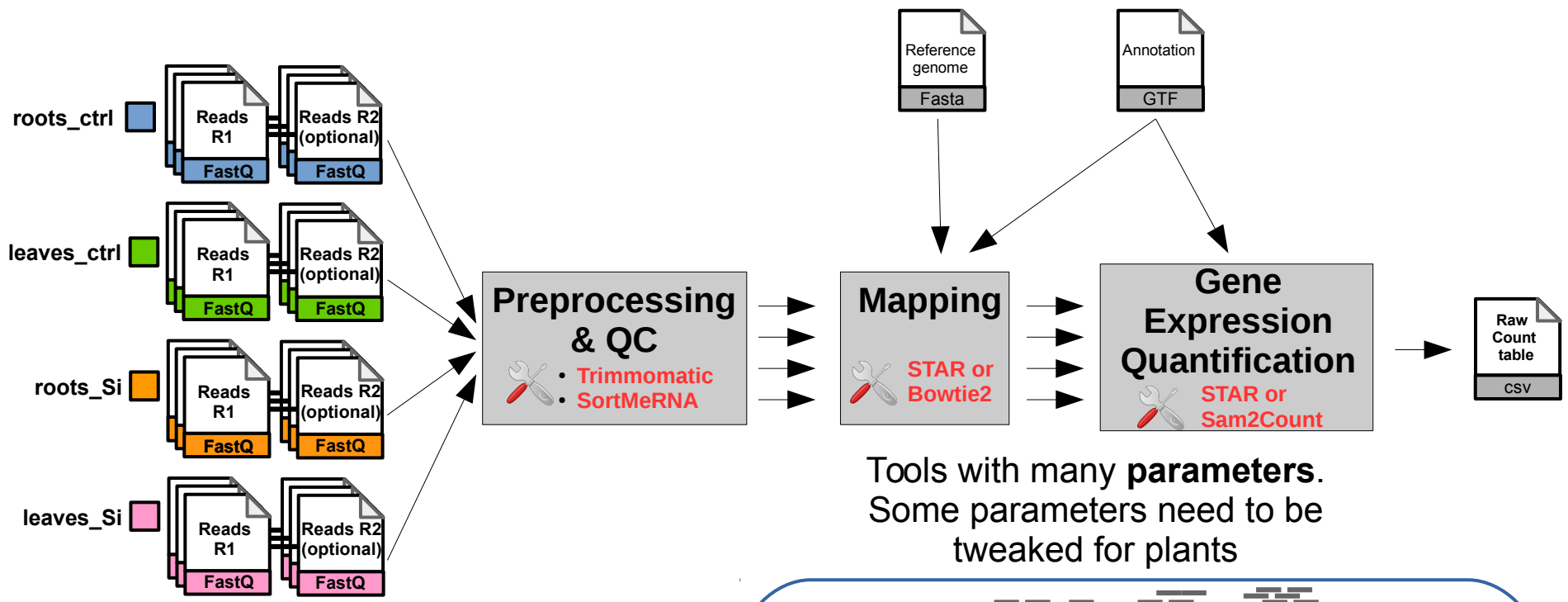
- In roots, is there a silicium treatment effect (C1) ?
- In leaves, is there a silicium treatment effect (C2) ?
- Is silicium effect similar for roots (C1) and leaves (C2) ?

Specific DE model: batch + tissue + treatment + tissue : treatment

Bioinformatic data analysis



Bioinformatic data analysis



Tools with many **parameters**.
Some parameters need to be tweaked for plants

Mapping : finding to which transcript the reads belong

Counting reads : counting aligned reads overlapping genes (Estimate the relative abundances of transcripts)

Raw count data

12 samples

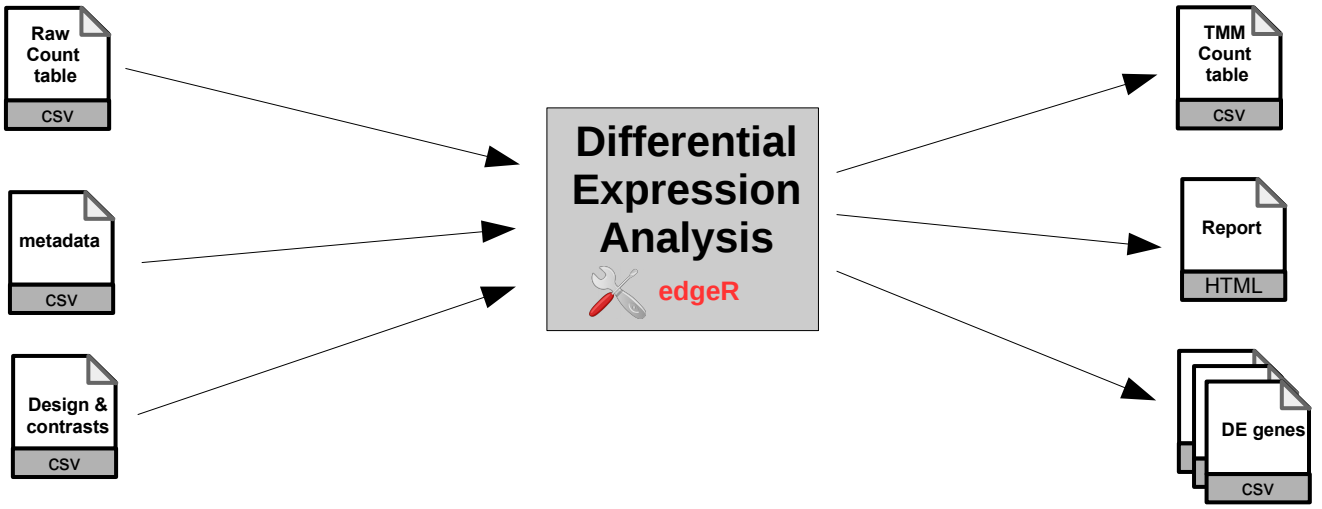
Gene	Sample #1	Sample #2	
id	Leaf_ctrl_1	Leaf_ctrl_2	
BnaA01g00010D	136	163	
BnaA01g00020D	209	184	
BnaA01g00030D	414	306	
BnaA01g00050D	1103	1054	
BnaA01g00060D	27	30	
BnaA01g00080D	52	67	
BnaA01g00210D	3923	3573	
BnaA01g00240D	0	0	
BnaA01g00270D	118	43	
BnaA01g00280D	1446	1214	

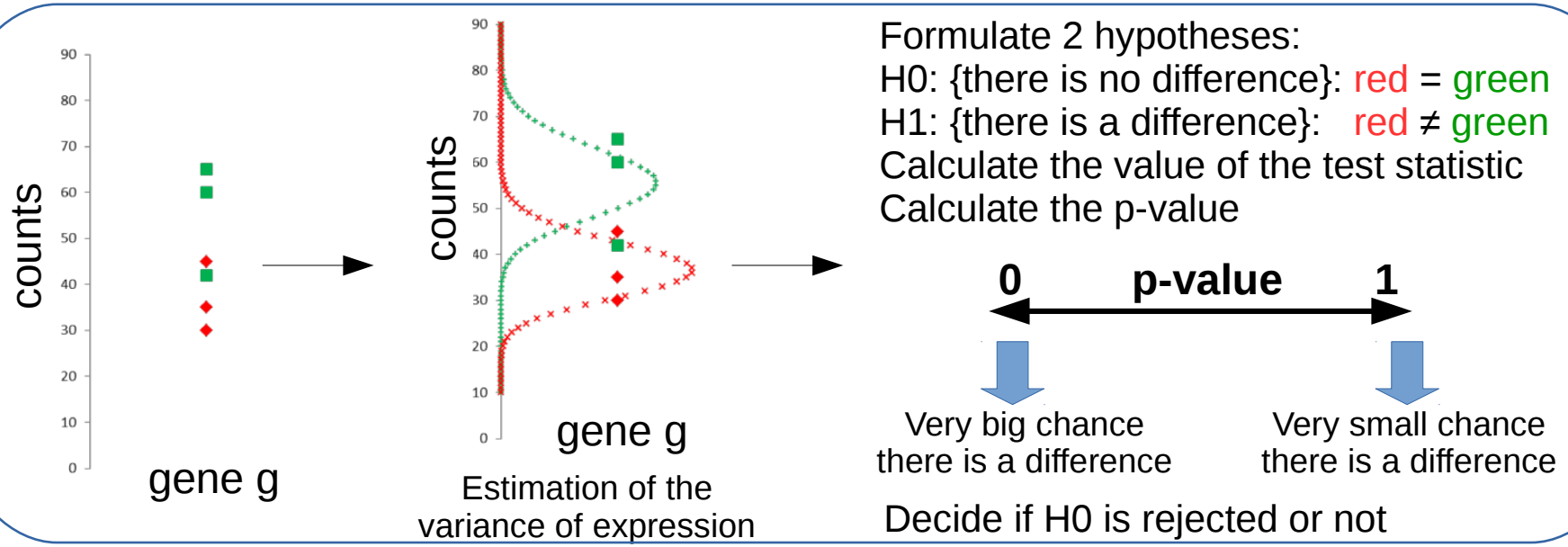
Sample #1 has 136 reads assigned to BnaA01g00010D gene

52,962 genes

library size = the total number of mapped reads from a sample

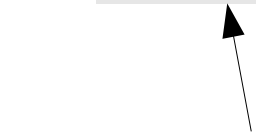
Statistical data analysis





Metadata

label	tissue	treatment	replicate
Leaf_ctrl_1	Leaf	ctrl	repbio1
Leaf_ctrl_2	Leaf	ctrl	repbio2
Leaf_ctrl_3	Leaf	ctrl	repbio3
Leaf_Si_1	Leaf	Si	repbio1
Leaf_Si_2	Leaf	Si	repbio2
Leaf_Si_3	Leaf	Si	repbio3
Root_ctrl_1	Root	ctrl	repbio1
Root_ctrl_2	Root	ctrl	repbio2
Root_ctrl_3	Root	ctrl	repbio3
Root_Si_1	Root	Si	repbio1
Root_Si_2	Root	Si	repbio2
Root_Si_3	Root	Si	repbio3



Normalized sample names

TMM normalization of count data

https://support.bioconductor.org/p/73844/

TMM (Trimmed Mean of M-values)

id	dyw2_HO_rep1	dyw2_HO_rep2	dyw2_WT_rep1
AT1G01010.1	328	656	1312
AT1G01020.1	124	248	129
AT1G01030.1	90	180	93
AT1G01040.2	463	926	480
AT1G01046.1	4	8	4
Taille librairie	1009	2018	2018

Profondeur de séquençage X 2

Biais de composition X 0,5

L'expression d'un gène ne dépend pas seulement de la profondeur de séquençage. Elle dépend aussi du niveau d'expression des autres transcripts. Même si les tailles de librairies sont identiques entres échantillons, des gènes peuvent masquer l'expression d'autres gènes.

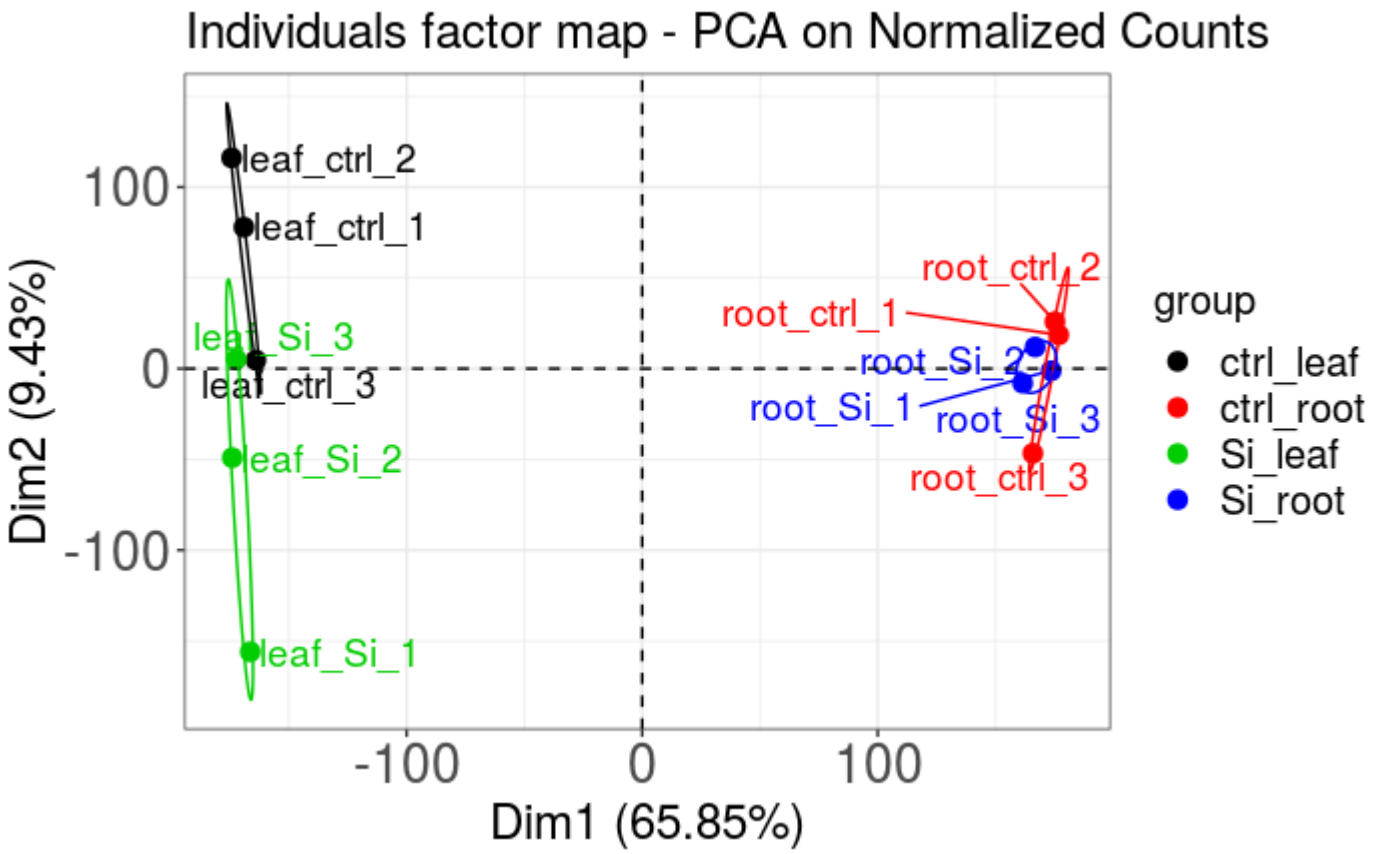
EdgeR use an elaborate normalization method called Trimmed Mean of M-values (TMM).

Normalized count table

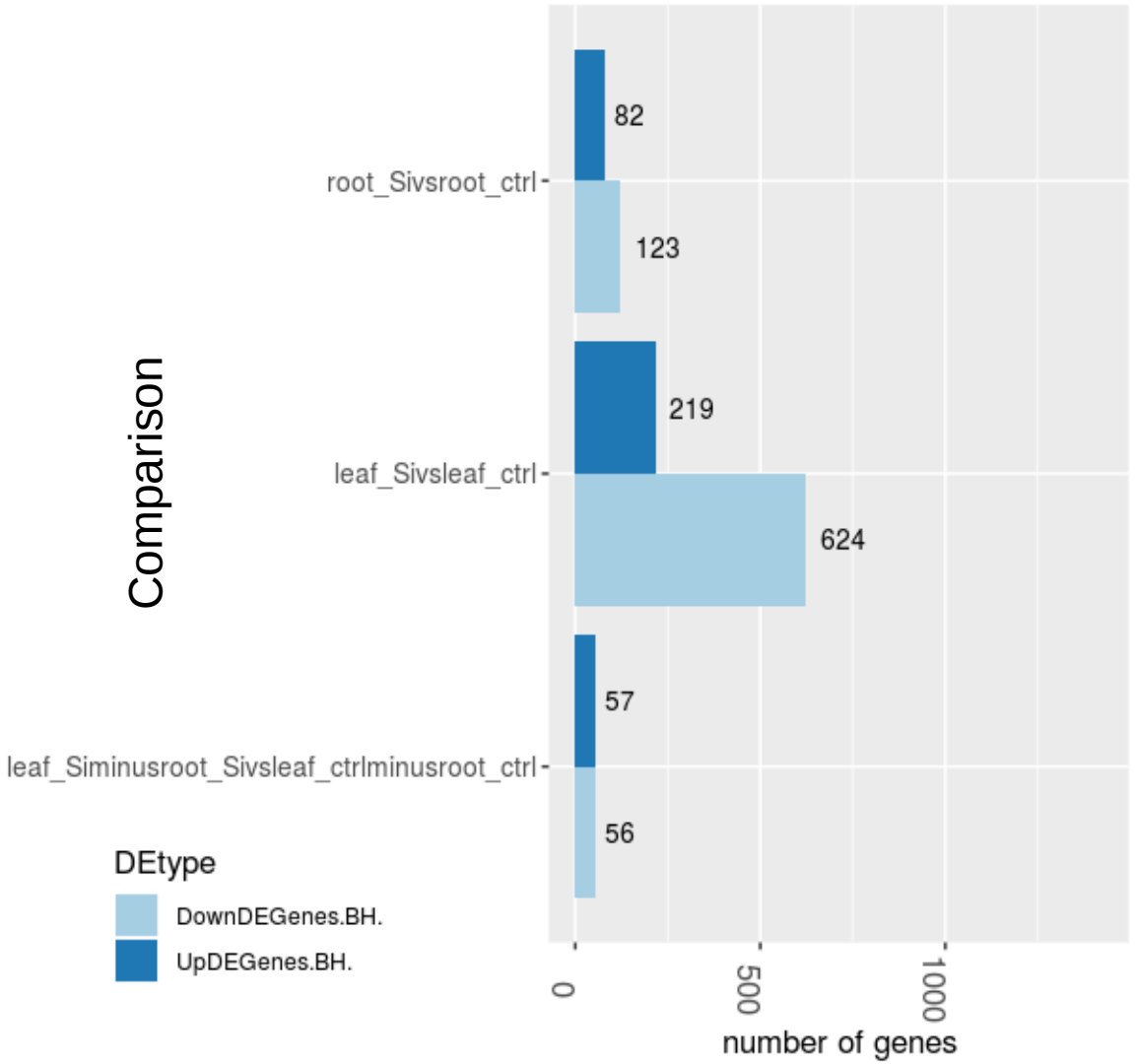
Id	leaf_ctrl_1	leaf_ctrl_2
BnaA01g00010D	197.198152800527	289.047591800266
BnaA01g00020D	303.047161289045	326.286852093551
BnaA01g00030D	600.294376907486	542.62922141645
BnaA01g00050D	1599.33501866898	1869.0562071011
BnaA01g00060D	39.1496332765752	53.1989432761225
BnaA01g00080D	75.3992937178485	118.810973316674

The observed counts of the features cannot be directly compared across samples, since there are differences in sequencing depth across libraries. The simplest normalization would involve rescaling counts by the library size (i.e. the total number of mapped reads from a sample). This normalization technique, however, is not always effective since few, very highly expressed genes can consume a substantial proportion of the total library size, causing the remaining genes to be under-sampled in that sample. EdgeR use a more elaborate normalization method called Trimmed Mean of M-values (TMM). It relies on the hypothesis that most features are not differentially expressed.

PCA on normalized counts



DE gene lists



Id	adjusted.PValue
BnaA09g13190D	4.82783933809054e-07
BnaA06g13250D	1.84961647978623e-06
BnaA05g08950D	1.36237608815255e-05
BnaA06g33720D	1.36237608815255e-05
BnaA09g25300D	1.36237608815255e-05
BnaA09g00120D	1.3874394551399e-05
BnaC02g08620D	1.3874394551399e-05
BnaC03g45470D	2.18598188783134e-05
BnaA08g22180D	3.16686328983066e-05

Conclusion

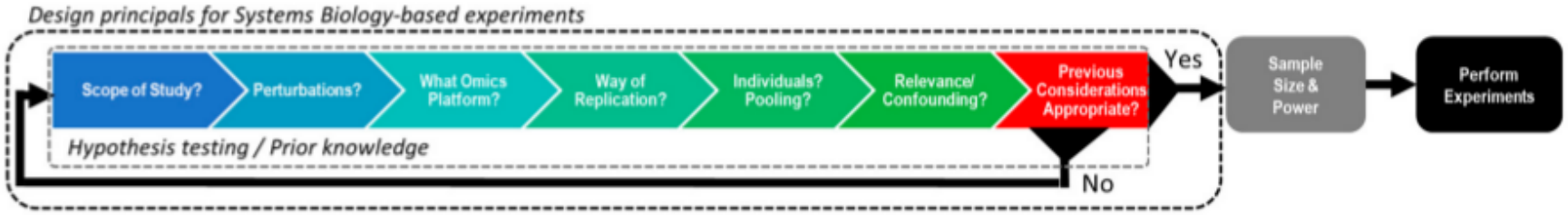
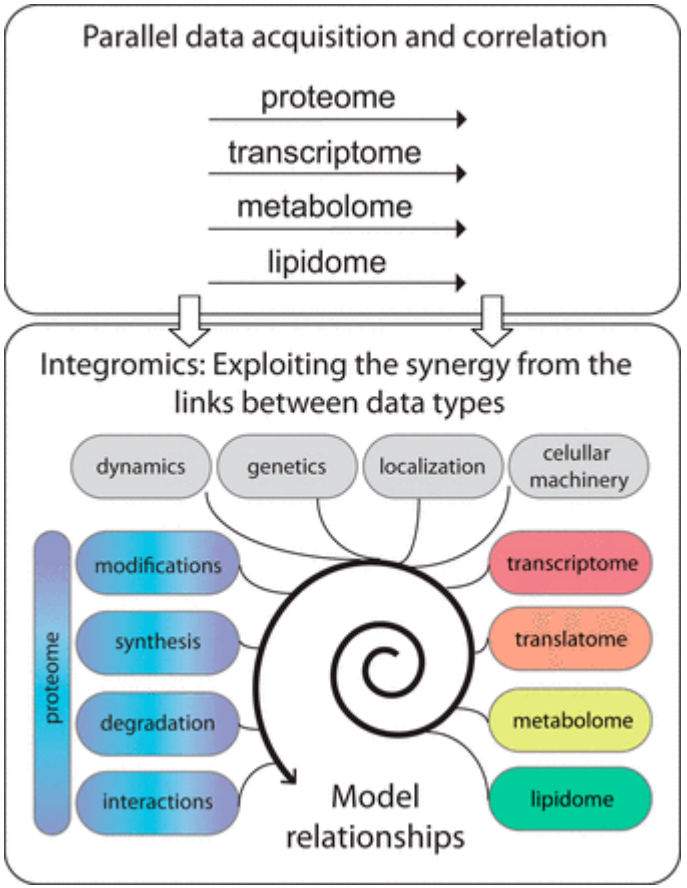


Figure 1. A conceptual model for designing a systems biology experiment.

A high quality experimental design is the key to success for any multi-omics study



Exploiting Interdata Relationships in Next-generation Proteomics Analysis

https://doi.org/10.1074/mcp.MR118.001246

Thank you for your attention