SOLADIS

EXPERIMENTER • ANALYSER • VALORISER

ANALYSIS OF GENOMIC DATA IN THE CONTEXT OF MACROARRAYS

presented by

Julie POUGET

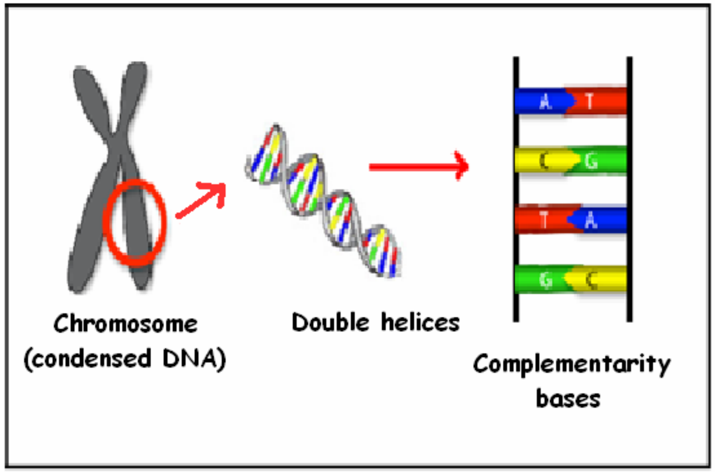
Contents at a glance

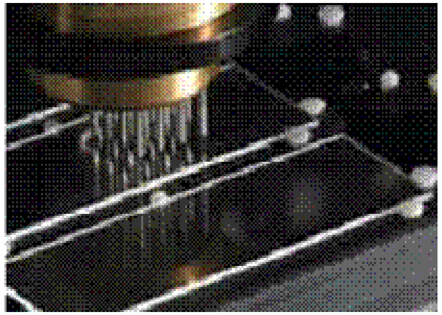
- DNA-arrays technology
- Data to be considered
- Statistical analysis method used
- Discussion

DNA arrays principles

DNA-Arrays

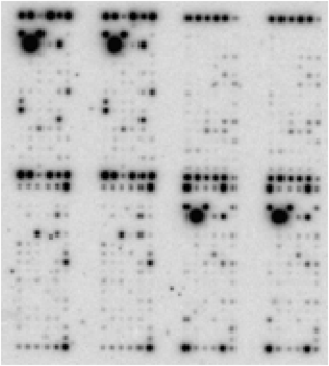
- Based on the principle of complementary bases
- It is a solid surface on which are fixed, in a orderly way, spots of DNA oligonucleotides (probes)





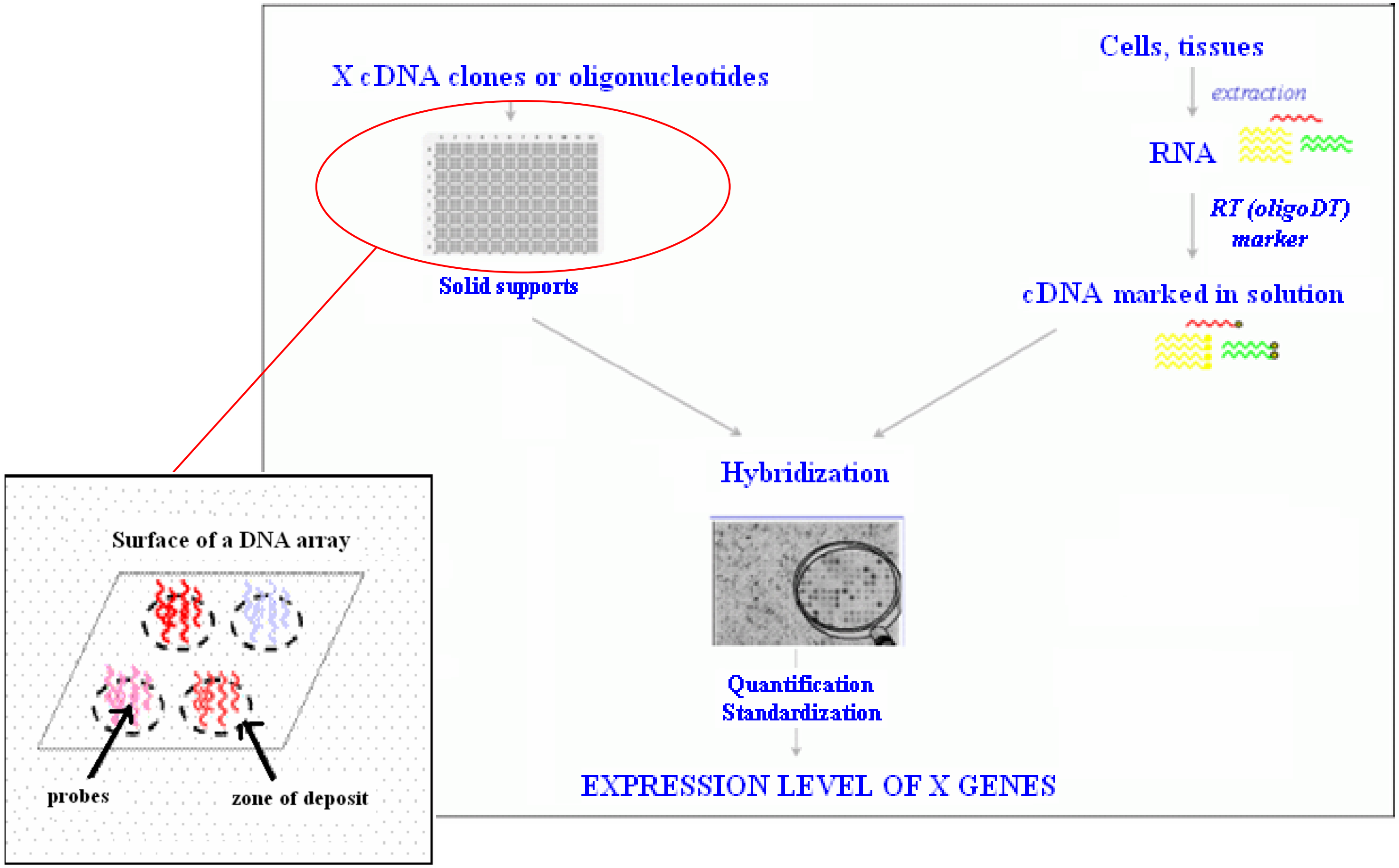
Labeled nucleic acids (targets) are hybridized with the probes on the support

- Probes-targets hybridization is detected and quantified to determine relative abundance of the target
- Quantification of the gene expression



Aims of DNA-arrays

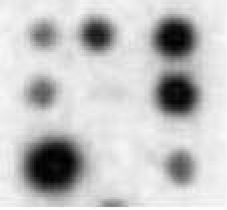
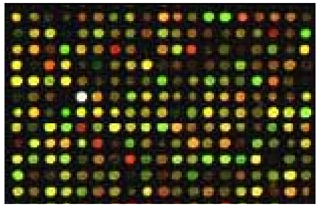
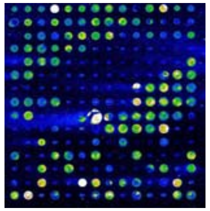
To measure and to evaluate gene expression differences between genes, on a large scale in a specific cell context



Analysis

The data to be considered

Macroarrays

High Density Array (macroarrays)	**(microarrays)**	**Oligonucleotides** array
		
Support: Nylon membrane	*Support*: Glass slides	*Support*: Glass slides
Size of spots : 0.5 – 1 mm	*Size of spots* : ~ 100µm	*Size of spots* : ~20µm
Density: some hundreds of spots per cm²	*Density*: 1000 to 10000 spots per cm²	*Density*: Until 250000 spots per cm²
Marker : Radioactive marker	*Marker* : Fluorescent marker Cy3 et Cy5	*Marker* : Fluorescent marker
1 experimental condition	2 experimental conditions	1 experimental condition

Microarrays and macroarrays may be used to differentiate the spot density on the support

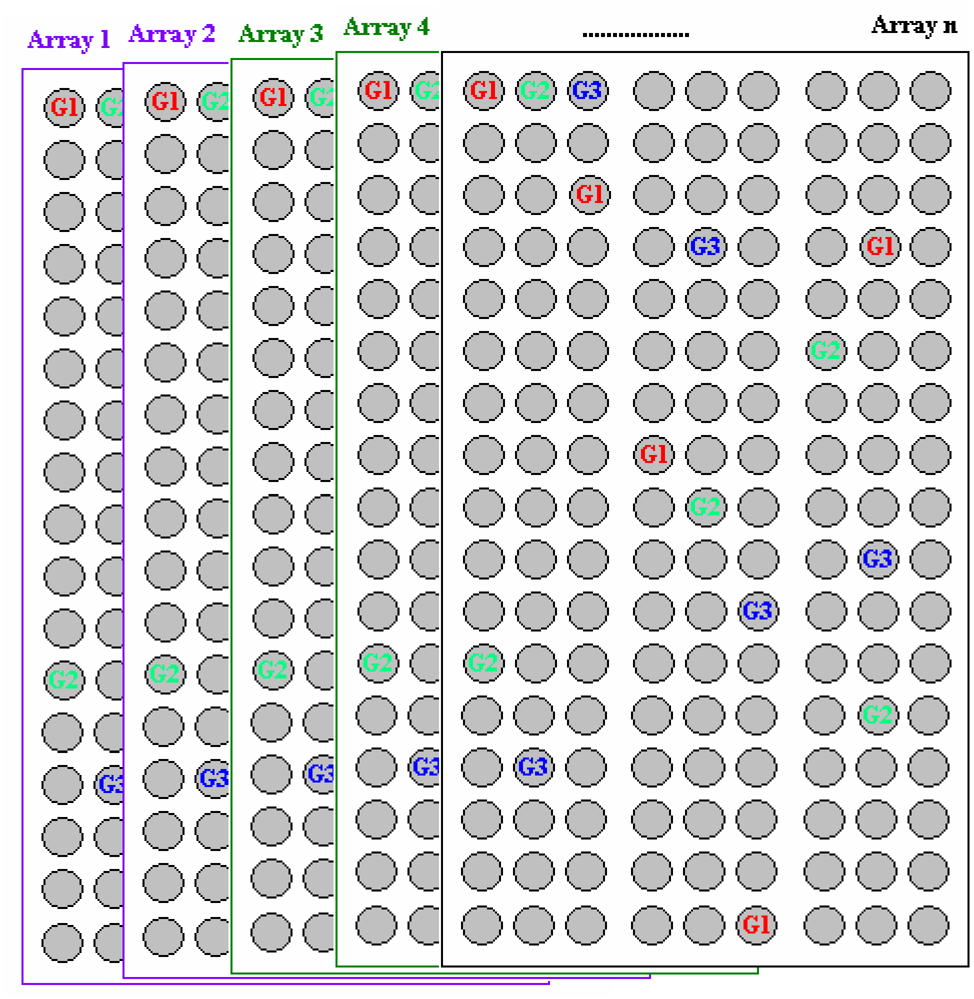
Macroarray term is usually used for the larger support and relatively low spot density (<200 spots/cm²).

Macroarrays : data to be treated

CONTEXT :

- One array = one experimental condition
- +/- UVA and +/- Se
- About 850 zones of spot deposition by macroarray including :
  - More than 300 oligos (probe) with replicates
  - Additional oligos : TOM et TOM-as (used as control of hybridization and as quality control)
  - Some blanks to measure the background level

Macroarrays : data to be treated



- Genes repeated on the position

- One-color array

- Several arrays for the same experimental condition

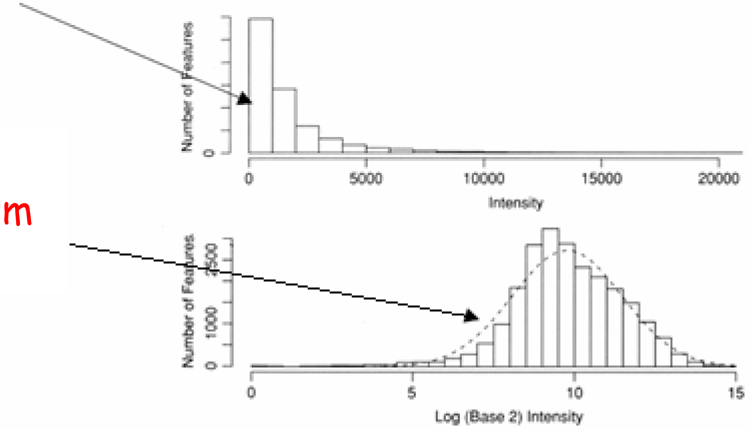
Macroarrays
Statistical analysis

Statistical analysis: Data processing (1)

Logarithmic transformation in base 2 from the raw data

Advantages of log2:

- ¤ *Treating differential up-regulation and down-regulation*
- ¤ *The extreme values have a lesser contribution (= robustness)*
- ¤ *The distribution is pseudo-normal*



Most of the intensities measured are low

Continuous spectrum of values.

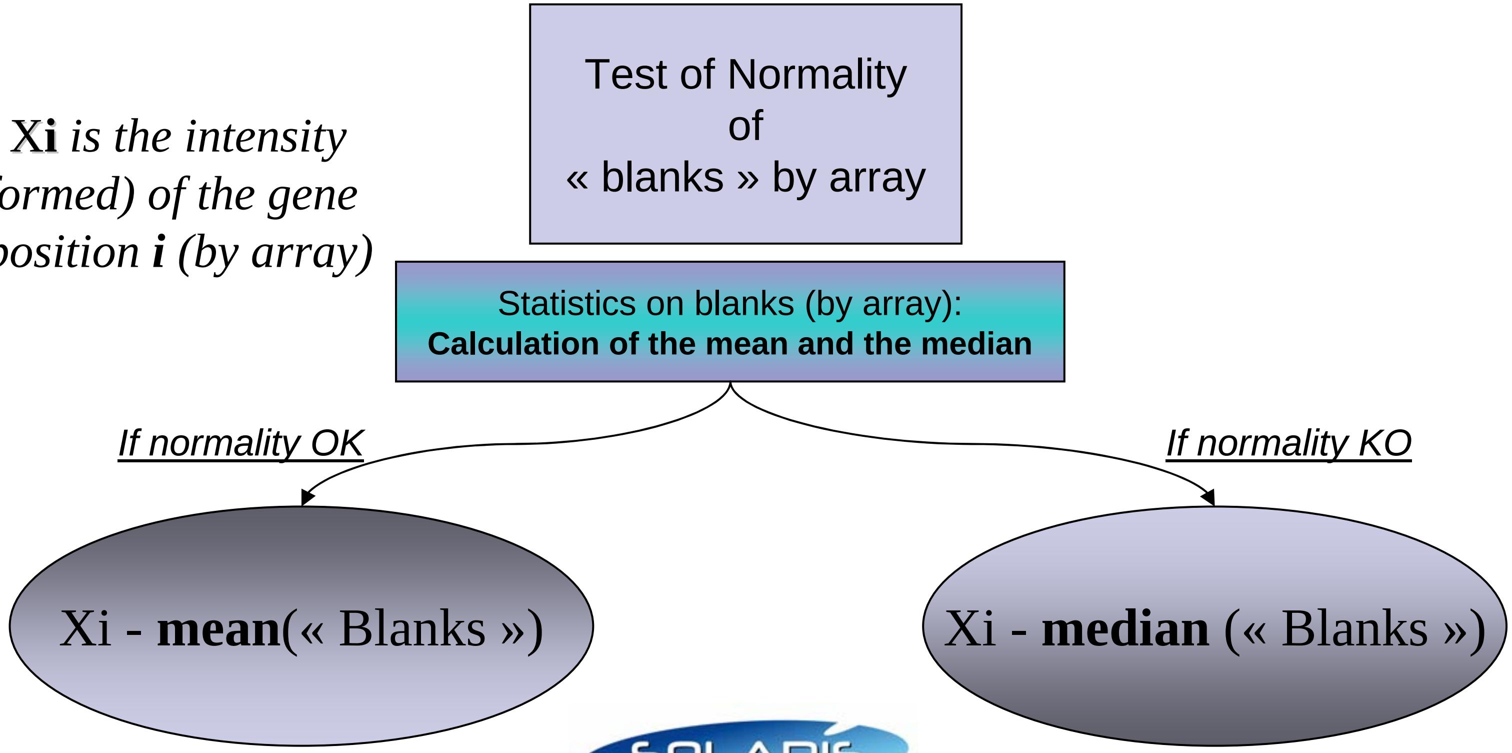
Statistical analysis: Data processing (2)

Background correction

It consists in “subtracting” the background level from each measured spot intensity.

From the data “blanks" ***(transformed by log2)*** :

where **Xi** *is the intensity (transformed) of the gene at the position* ***i*** *(by array)*



Test of Normality of « blanks » by array

Statistics on blanks (by array):
Calculation of the mean and the median

If normality OK

If normality KO

Xi - **mean**(« Blanks »)

Xi - **median** (« Blanks »)

Statistical analysis: Data processing (3)

Normalization : *Why normalizing?*

Normalization : a step to eliminate bias factors

¤ To be confident about the data qualities coming from an array

¤ To be able to compare several macroarrays using the same set of genes coming from the same experimental condition (duplicate or triplicate)

¤ To be able to exploit any array possessing a gene or a group of genes of interest as:

- get back public data
- include the results of several experiments

Statistical analysis: Data processing (4)

Normalization : *How ?*

Several methods exist such as:

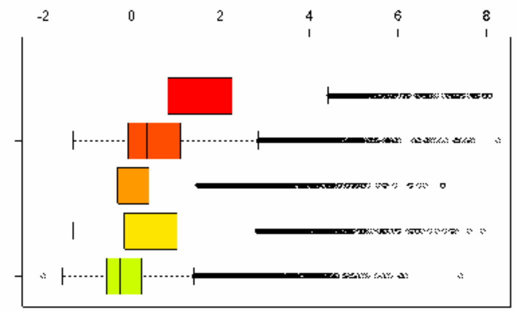
- LOWESS normalization (two channels, so depends on the labeling of the target)
- Quantiles normalization
- Normalization by standard scores on arrays
- Global normalization methods
- Etc…

Statistical analysis: Data processing (5)

Normalization : *A question of point of view !*

To center :

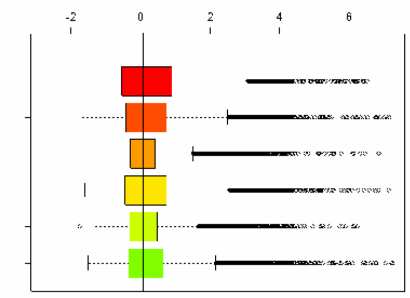
- Mean (TOM)
- Mean (Array)
- Median (Array)
- Global

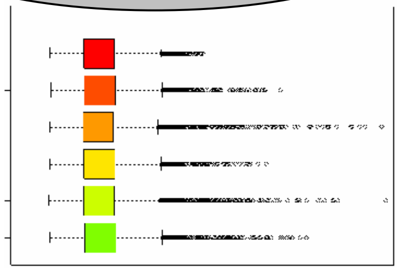


To reduce :

- Standard deviation (Array)
- MAD (Array)

MAD *(Median Absolute Deviation)*
Médian(|Xi-Médian|)





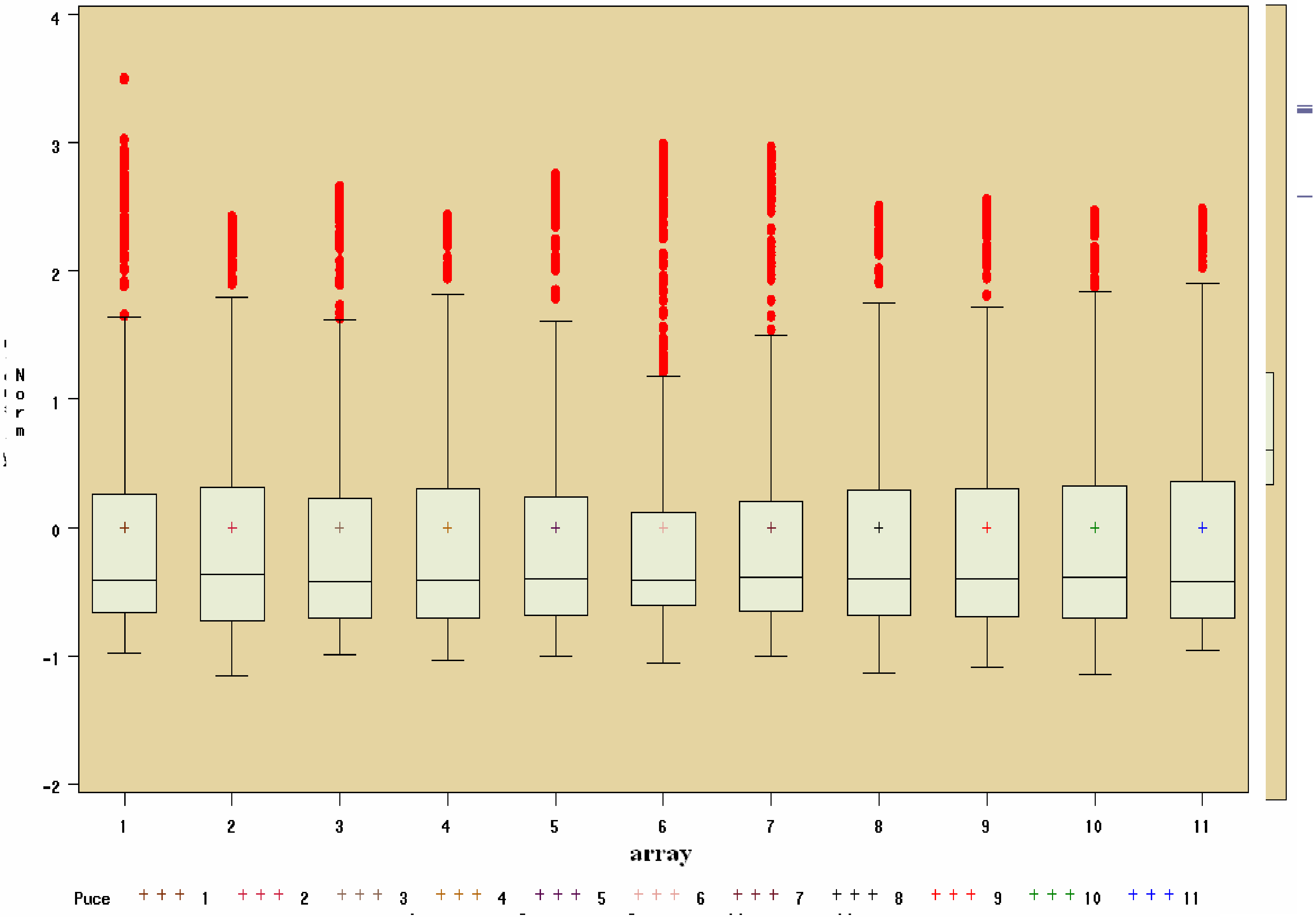
CHOICE
by BOXPLOT

Need to construct an indicator

Mean (TOM) / St.dev (Array)	Mean (TOM) / MAD (Array)	Mean (Array) / St.dev (Array)	Mean (Array) / MAD (Array)	Median (Array) / St.dev (Array)	Median (Array) / MAD (Array)	Global Normalization *(Mean of Means)*

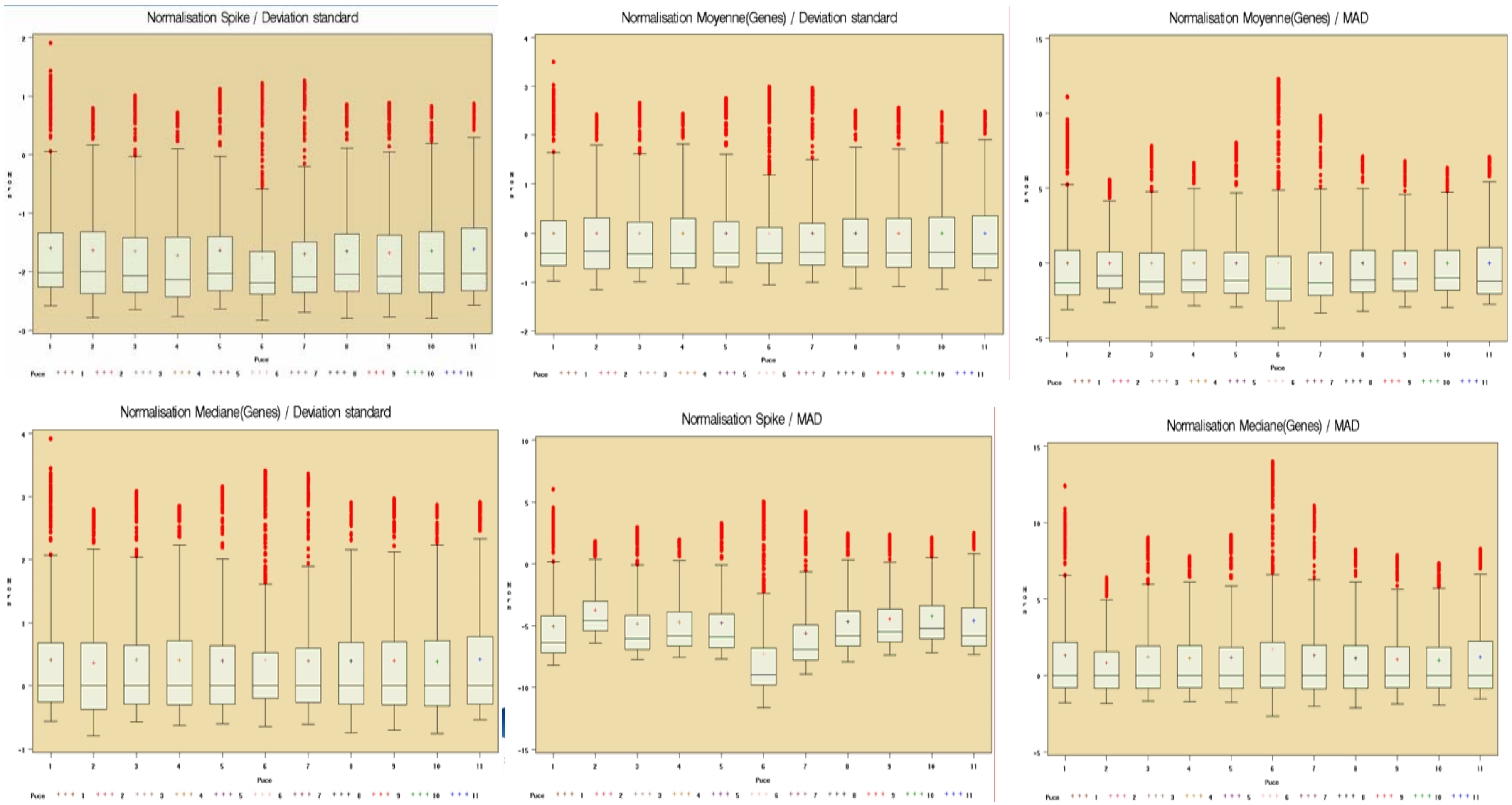
SOLADIS
EXPERIMENTER • ANALYSER • VALORISER

Normalization by Mean(array) / MAD(array)



Statistical analysis: Data processing (6’)

Comparison

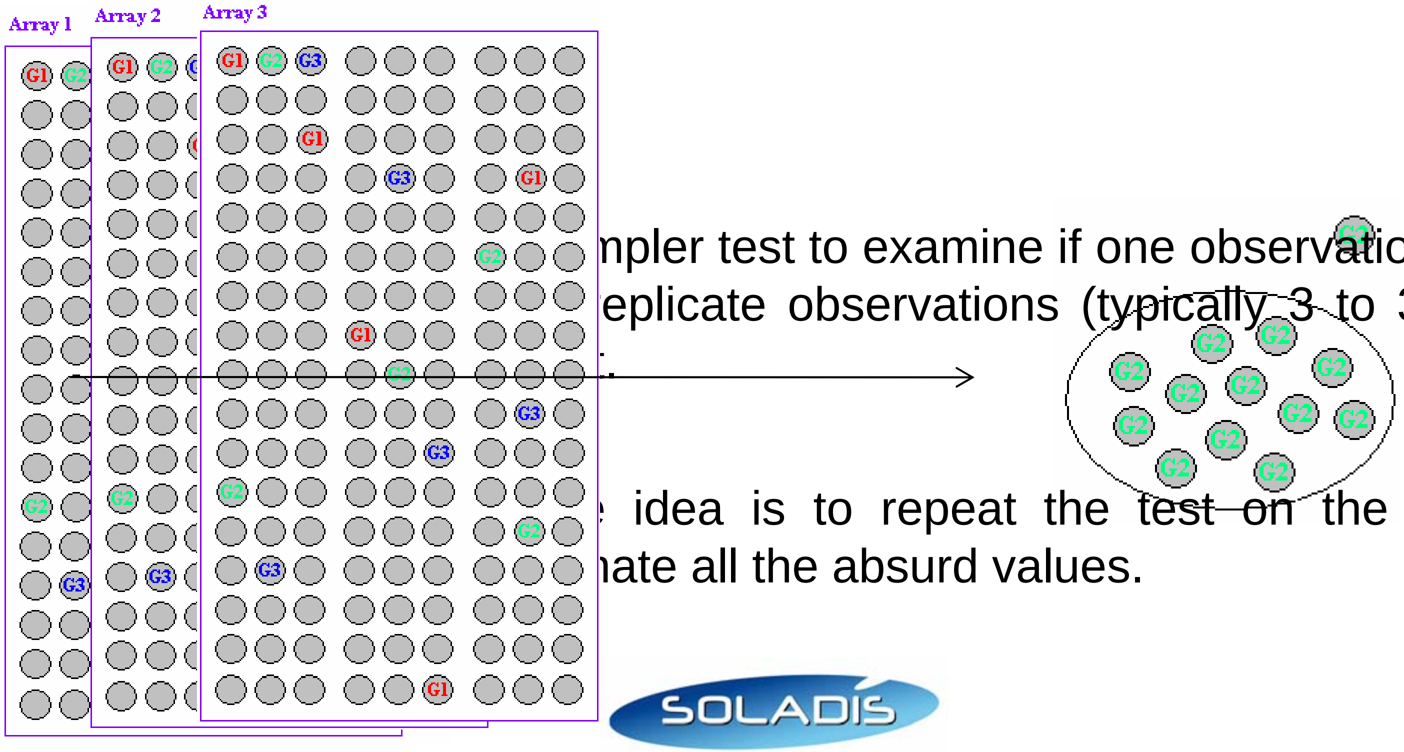


Statistical analysis: Data processing (7)

Last step of data processing

What is an outlier ?

Usually, an **outlier** is defined as an observation generated from a different distribution (or a different model) from the main set of data.



D

mpler test to examine if one observation (the max eplicate observations (typically 3 to 30) can be .

Ca idea is to repeat the test on the genes by ex ate all the absurd values.

Statistical analysis: Tests (1)

Determination of the model

Proc MIXED

*A mixed model is a statistical model containing both **fixed effects** and **random effect**. It is particularly useful in settings **where repeated measurements are made** on the same statistical units, or where measurements are made on clusters of related statistical units.*

Are there some repeated data?

One model for each biological question…

Are there random effects?

What problem do we want to solve?

Statistical analysis: Tests (2)

Multiple tests

It means making a test **gene by gene**. For each test,

H0 : {the gene is not differentially expressed}

H1 : {the gene is differentially expressed}

What does mean multiple tests ?

- Several thousand tests simultaneously
- Structure of dependence: plans of correlation intern complex between variables

	Decision	
	H0 no rejected	**H0 rejected**
H0 true	**True positives**	**False positives**
H1 true	**False negatives**	**True negatives**

⇨ Two types of error associated to the multiple tests: **the FWER** (Family Wise Error Rate) and **the FDR** error (False Discovery Rate).

Statistical analysis: Tests (3)

Multiple Tests Adjustment

- FWER (*Family Wise Error Rate*) - Bonferroni
- FDR (*False Discovery Rate*) – Benjamini et Hochberg

OBJECTIVE : **Reduce the number of false positives and false negatives**

320 genes	Mixed model	Mixed Model + FDR
Test on normalized values	37	1
Test on normalized values without outlier	41	2

Statistical analysis: Tests (4)

Pairwise comparisons

When an effect is significant, it is possible to look at the pairwise comparisons but it needs multiple comparison adjustments

Dunnett's adjustment:

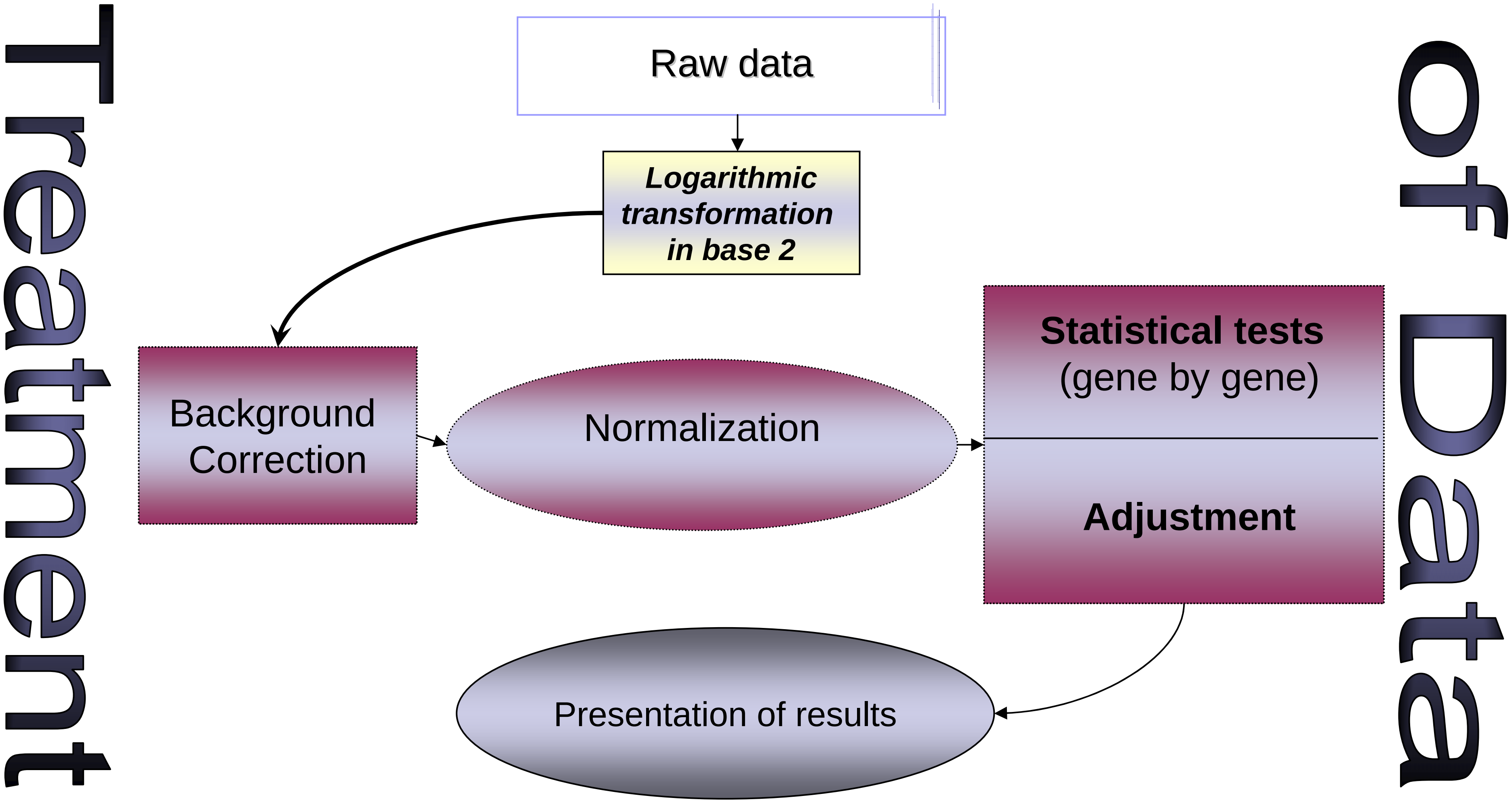
When all differences are analyzed with a control level

Tukey's adjustment:

To adjust all pairwise differences

Summary and conclusion

Statistical process



Treatment

Raw data

Logarithmic transformation in base 2

Background Correction

Normalization

Statistical tests
(gene by gene)

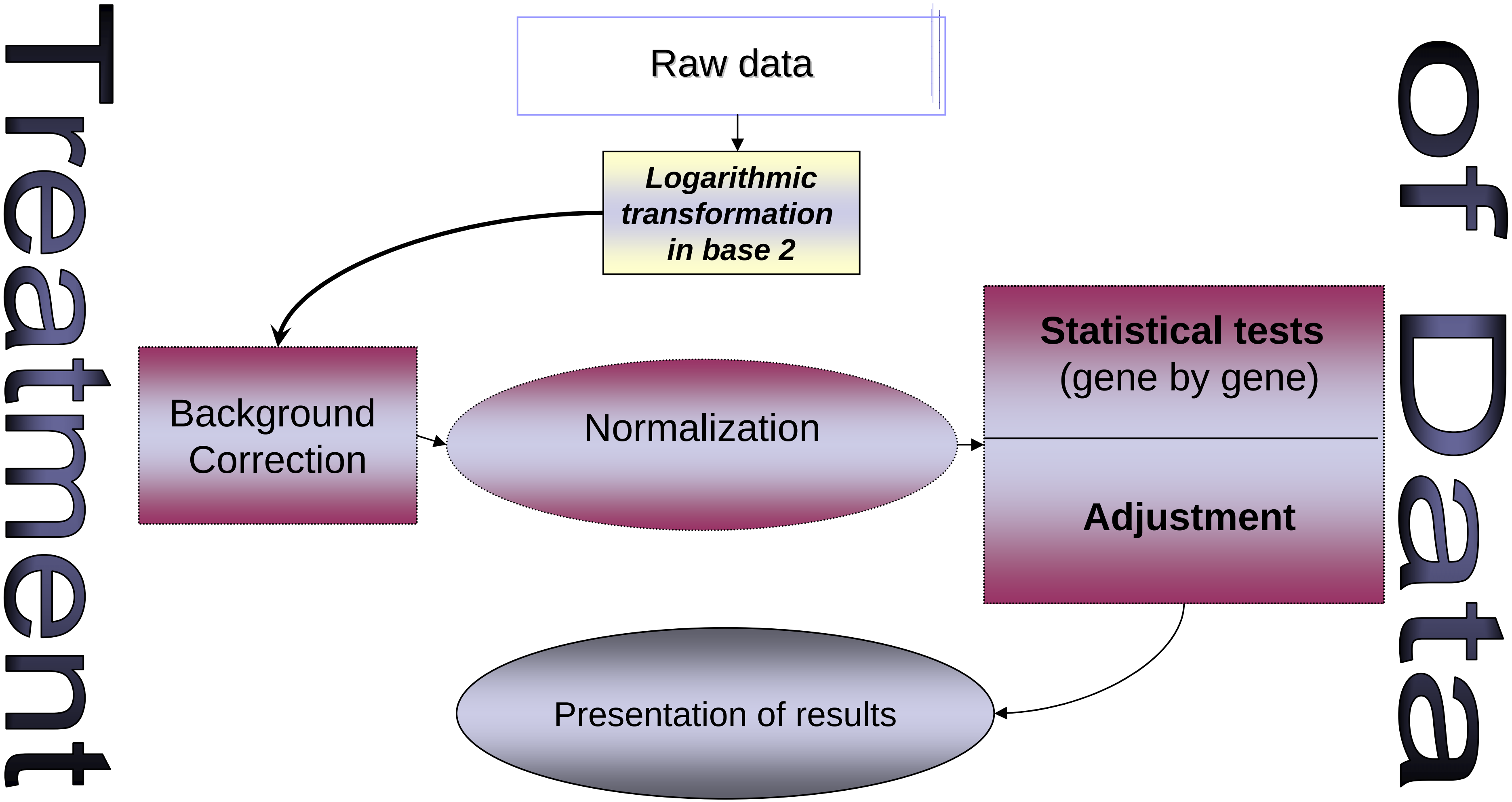
Adjustment

Presentation of results

of Data

Summary and conclusion

Statistical process



Treatment of Data

SOLADIS

EXPERIMENTER • ANALYSER • VALORISER

… Thank you