The Atlas Project: A large-scale microarray project to identify steroid-responsive genes in mice.

Oncology and Molecular Endocrinology Research Center,
Laval University Medical Center (CHUL) and Laval University,

Project Atlas
General Objectives

1. Tissue- and cell-specific profiles of steroid-regulated gene expression

2. Define candidate genes as key regulators of steroid action in physiology and disease: therapeutic targets

List of tissues

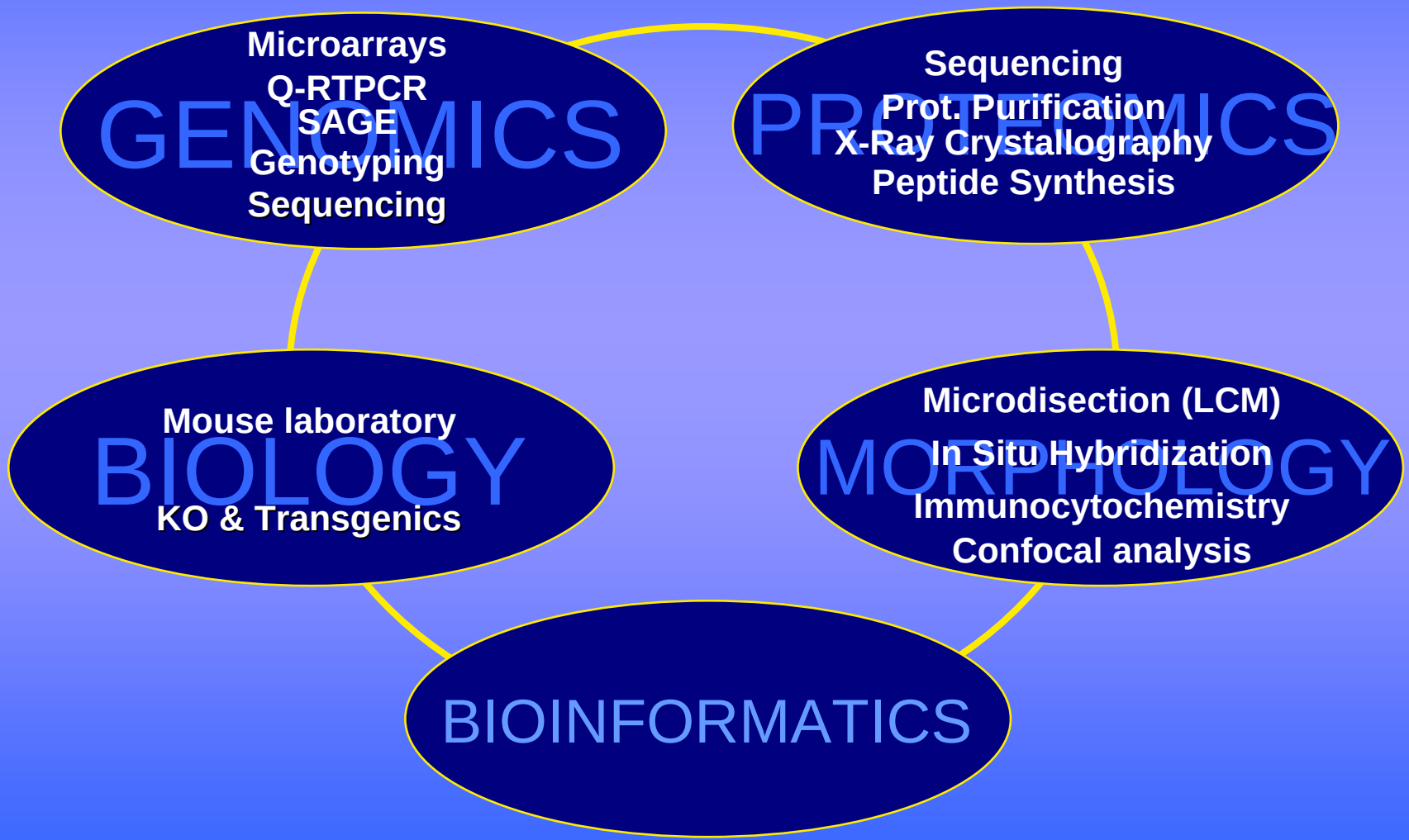
1. Prostate
2. Seminal vesicles
3. Testes
4. Epididymides
5. Mammary glands (inguinal)
6. Uterus
7. Oviducts
8. Vagina
9. Ovaries
10. Pituitary gland
11. Thyroid-parathyroids
12. Adrenals
13. Liver
14. Gallbladder
15. Heart
16. Pancreas
17. Kidneys
18. Spleen
19. Thymus
20. Fat – retroperitoneal
21. Fat – subcutaneous
22. Skin – ventral
23. Skin - dorsal
24. Footpad
25. Eyes
26. Lungs - Bronchi
27. Trachea
28. Brain (cerebral cortex)
29. " (hypothalamus)
30. " (basal ganglia)
31. " (thalamus)
32. " (amygdala)
33. " (tectum)
34. " (tegmentum)
35. " (cerebellum)
36. " (pons)
37. " (medulla)
38. " (hippocampus)
39. Spinal cord (thoracic)
40. Sciatic nerve
41. Urinary bladder
42. Aorta (thoracic)
43. Vena cava
44. Femoral biceps (skeletal muscle)
45. Smooth muscle
46. Parotid gland
47. Mandibular gland
48. Macrophages
49. Lymphocytes
50. Lymph nodes (mesenteric)
51. Bone (femur)
52. Bone (ribs)
53. Bone (sternum)
54. Tongue
55. Oesophagus
56. Stomach
57. Duodenum
58. Ileum
59. Jejunum
60. Caecum
61. Colon
62. Rectum

Endocrine conditions

male and female

1- Intact
2- Adrenalectomy (ADX)
3- ADX + glucocorticoid
4- ADX + mineralocorticoid
5- Gonadectomy (GDX)
6- GDX + DHT
7- GDX + estradiol (E_2)
8- GDX + progesterone (P)

Quebec Genome Center Platforms

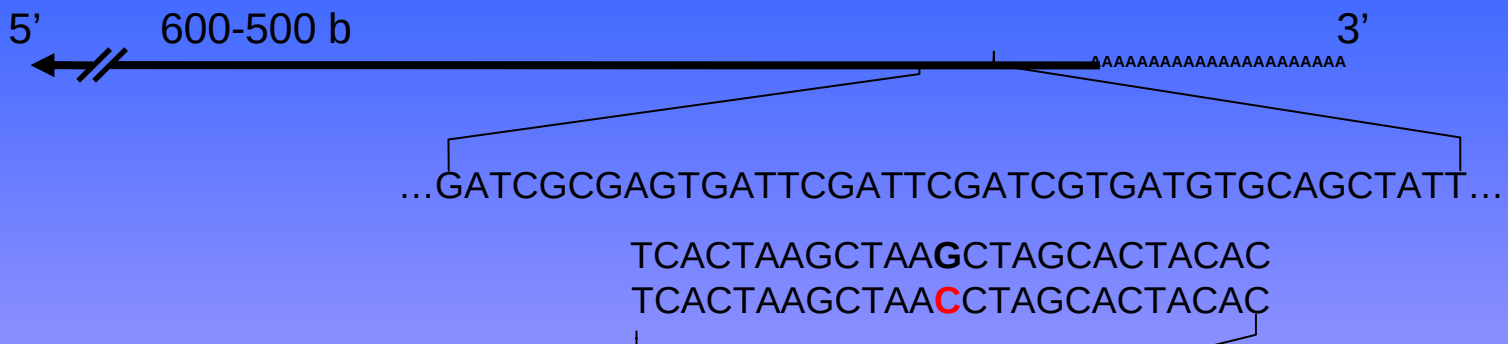


Murin Database

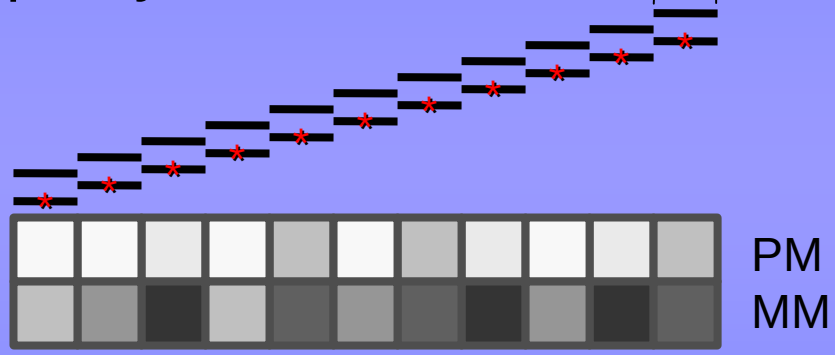
- In vivo
  - Specimen tracking
- Microarrays
  - Data storage
  - Data mining
- SAGE
  - Data storage
  - SAGE map set generated from Unigene
  - Data mining
- Proteomics
  - Platform integration (2D-gel, maldi-TOF, LCQ)
  - Protein identification (Mascot)
- In situ Hybridization
  - Probe description and image storage
- Sequence
  - Sequence tracking

Microarray Platform

Oncology and Molecular Endocrinology Research Center,
Laval University Medical Center (CHUL) and Laval University,



Probe pair multiplicity:



-Assess the expression level of a specific transcript

-Improves the signal to noise ratio (efficiencies of hybridisation are averaged over multiple probes)

-Reduces the rate of false positives.

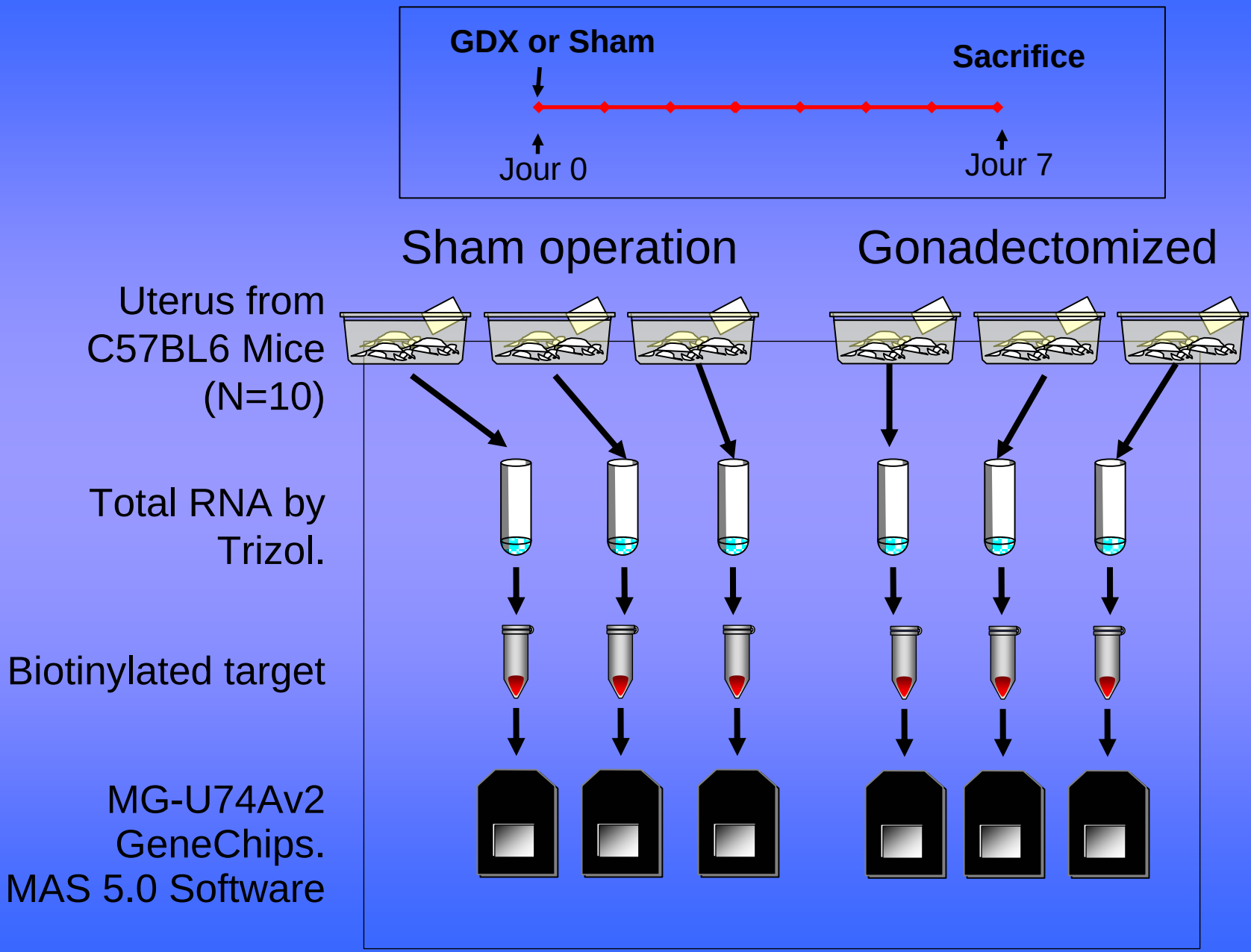
The intensity information from probes pair from each probe set can be combined in many ways to get an overall intensity measurement for each gene, **but there is currently no consensus as to which approach yields more reliable results.**

What is the best approach to extract probe set information?

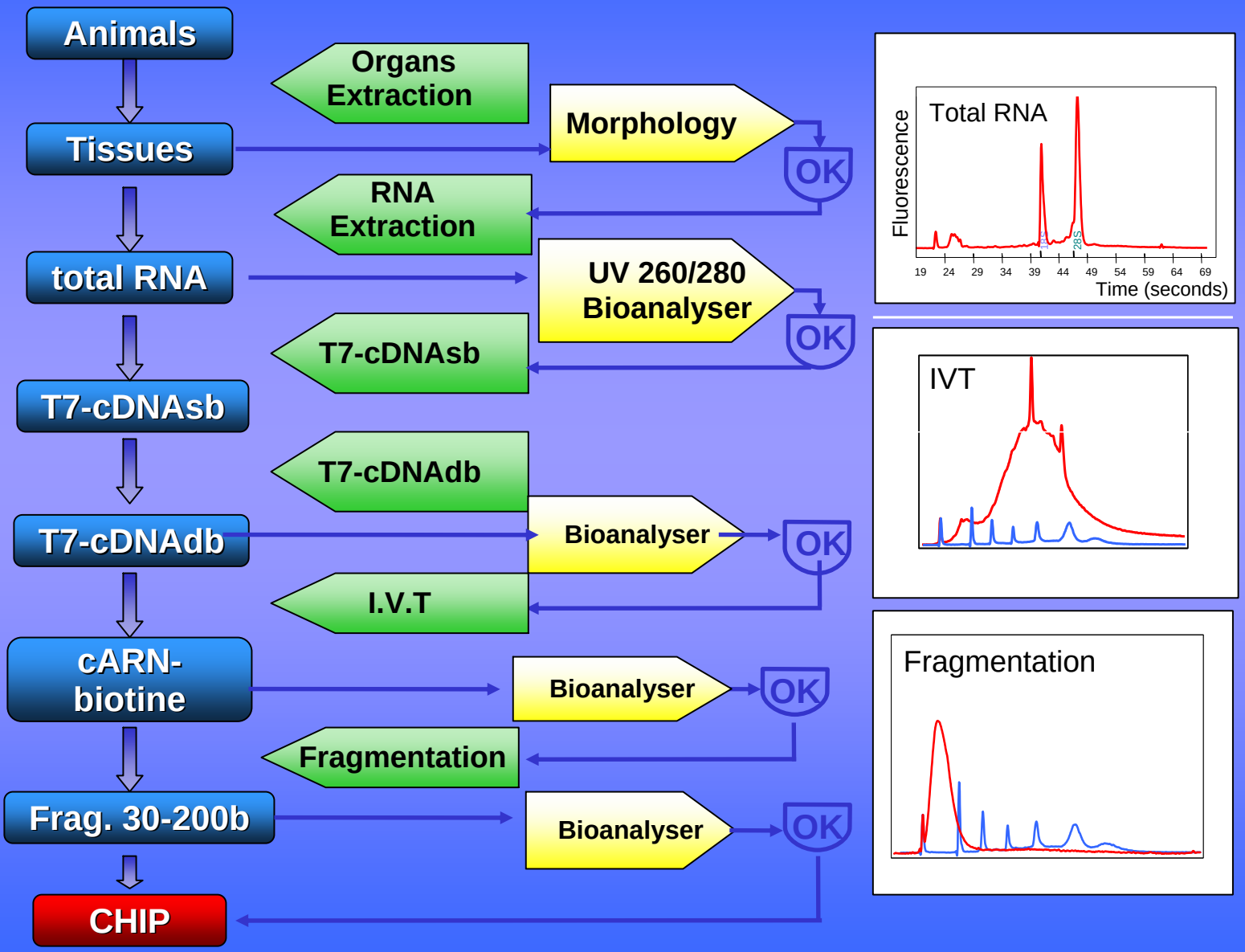
Objective:

To compare four probe level data extraction algorithms availables as open source* by validating each one by Q-RT-PCR.

*Bioconductor Project (www.bioconductor.org), and official software releases.


GDX or Sham
Sacrifice
Jour 0
Jour 7
Sham operation
Gonadectomized
Uterus from C57BL6 Mice (N=10)
Total RNA by Trizol.
Biotinylated target
MG-U74Av2 GeneChips. MAS 5.0 Software

Standard protocols for Affymetrix chips



Validation by Q-RT-PCR

-Total RNA by QiaZOL+ Rneasy colon kit purification (Qiagen)
- UV Quantification.
- Reverse transcription of 5 ug total RNA by SSIIRT and poly(T)20. (Invitrogen)
- RNAseA digestion (Invitrogen)
- QIAquick colons purification (Qiagen)
- Oligoprimers by GeneTools (Biotools) (~200 bp amplicons)
- LightCycler Realtime PCR apparatus and SYBR®Green I kit (Roche)

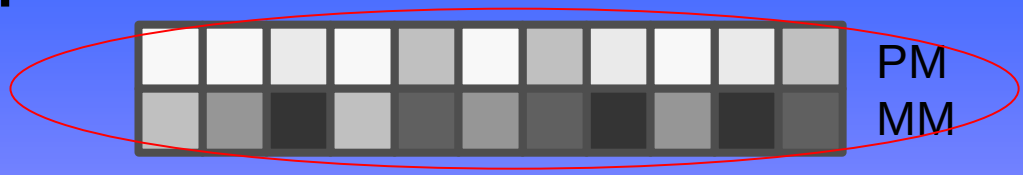
Conditions: Total RNA = 20ng

Denaturation	94oC 15 sec.
Annealing	Primer specific oC 10 sec.
Elongation	72oC 20 sec.

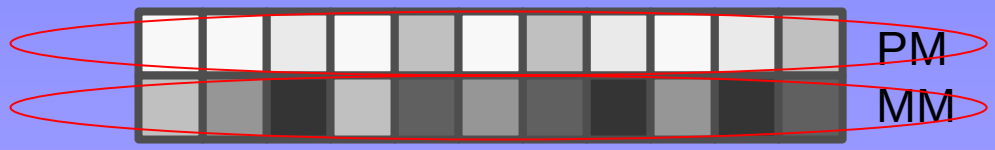
- LightCycler v1.1 software.
-Quantification and normalization by the housekeeping gene subunit O of ATPase (ATP5O).

The four methods:

- Microarray Suite 5.0
  (MAS 5.0, Affy).

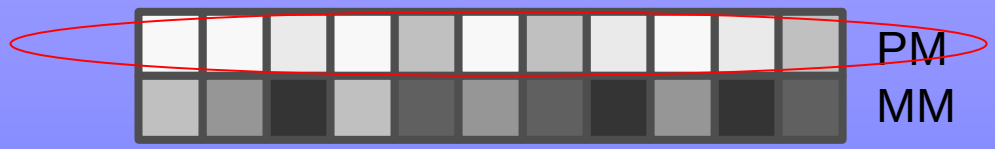


- Model Based Expression Index
  (dChip, Li & Wong)



- Multiplicative Noise Model
  (Sasik et al.)

- Robust Multi-array Average
  (RMA, Irrizary et al.)



Data extracted by each algorithm was filtered by LFC method* to identify significant differentially expressed genes.

*David M. Mutch et al. BMC Bioinformatics 2002, 3:17, with modifications

Identification of differentially expressed genes

Because the variability in the measurement of gene expression is greater at low expression, we want a method that takes into account both expression levels and fold changes.

The LFCM equation

A variable fold change limit (LFC) decreasing with the gene expression value was used to select differentially expressed genes.

The LFC equation is $Y = a + b / X$,where X is the minimum intensity of gene expression from two conditions and Y is the fold change limit.

The parameters a and b were estimated based on the distribution of ratios calculated from replicated chips.

The resulting cut-off point, $Y = a+b / X$, gives an approximately constant rate of false positive modulated genes of 0.1%.

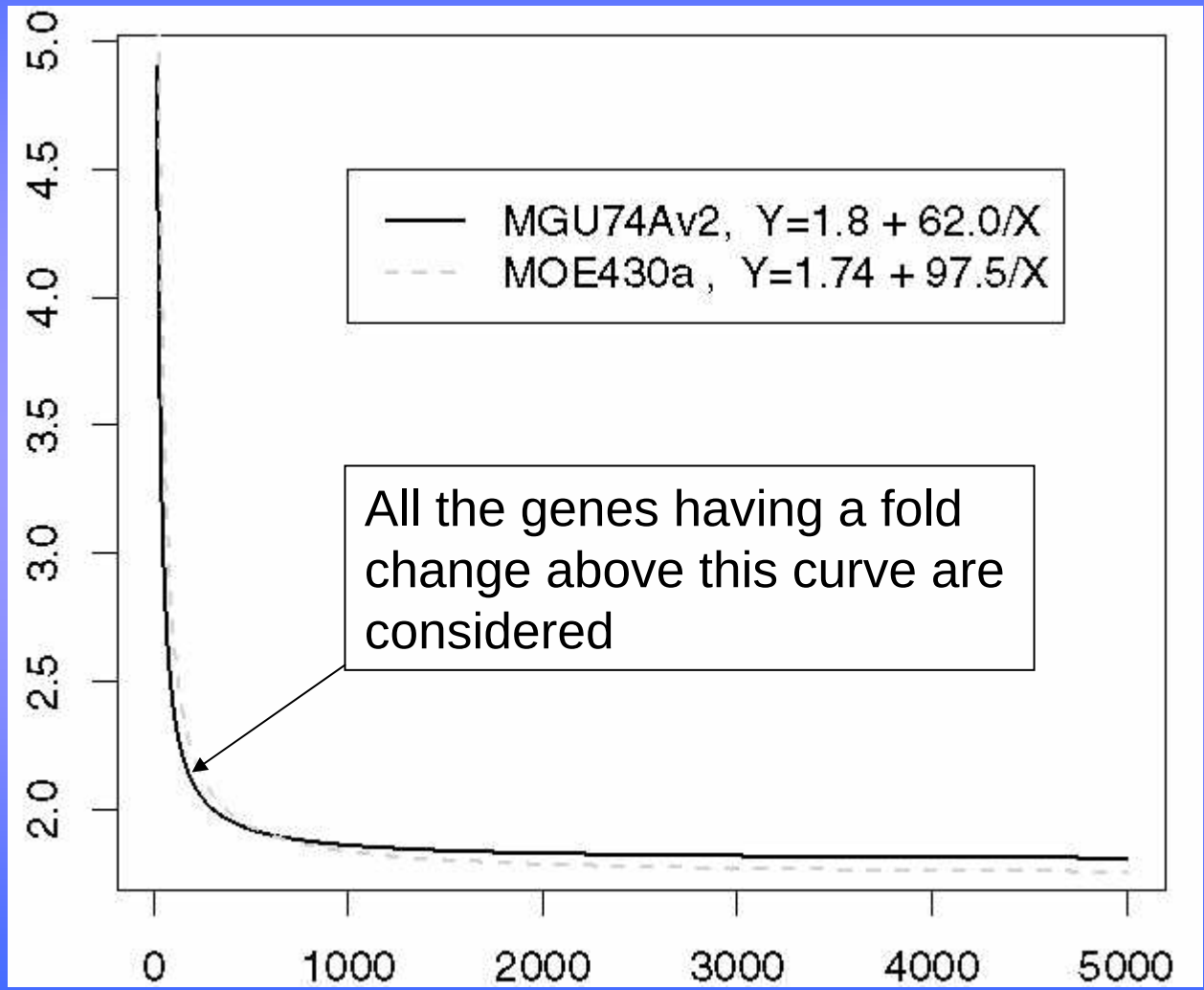
All the genes having a fold change above this curve are considered

LFC equation
Y = a+b/X

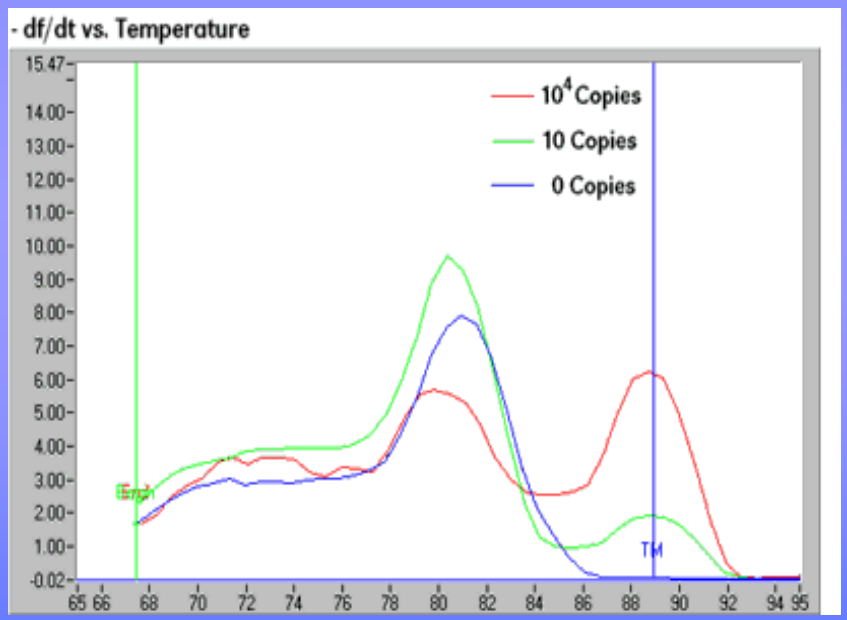
MAS 5.0	LFC= 1.8+62.0/X (MG-U74Av2)
	LFC= 1.7 + 97.5/X (MOE430A)
dChip	LFC= 1.58/X
RMA	LFC= 1.62/X
MNM	LFC= 1.69/X

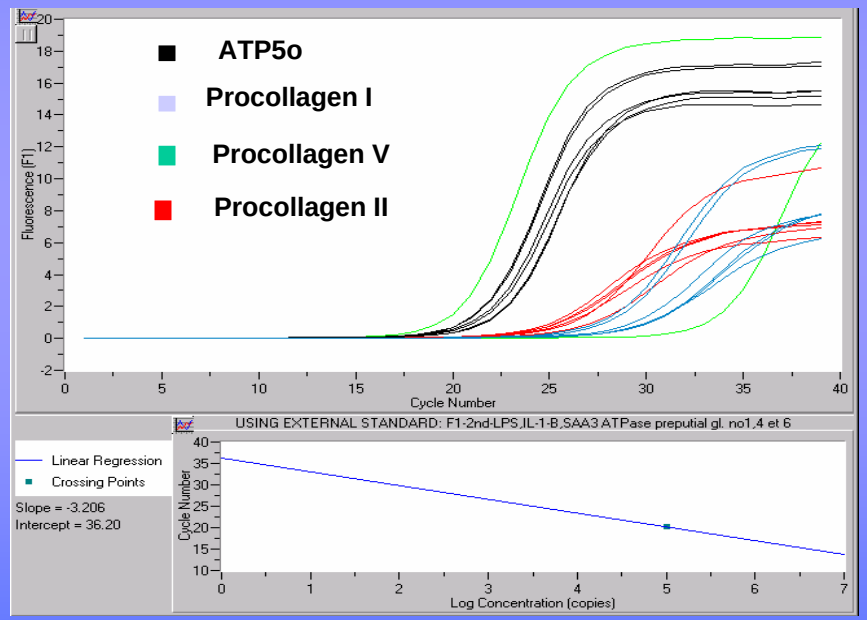
For three of the four methods (i.e. dChip, RMA and MNM) the parameter b is equal to 0.

LFC equation for MAS 5.0
Y = 1.8 + 62.0/X



Example of Q-RT-PCR LightCycler data output.



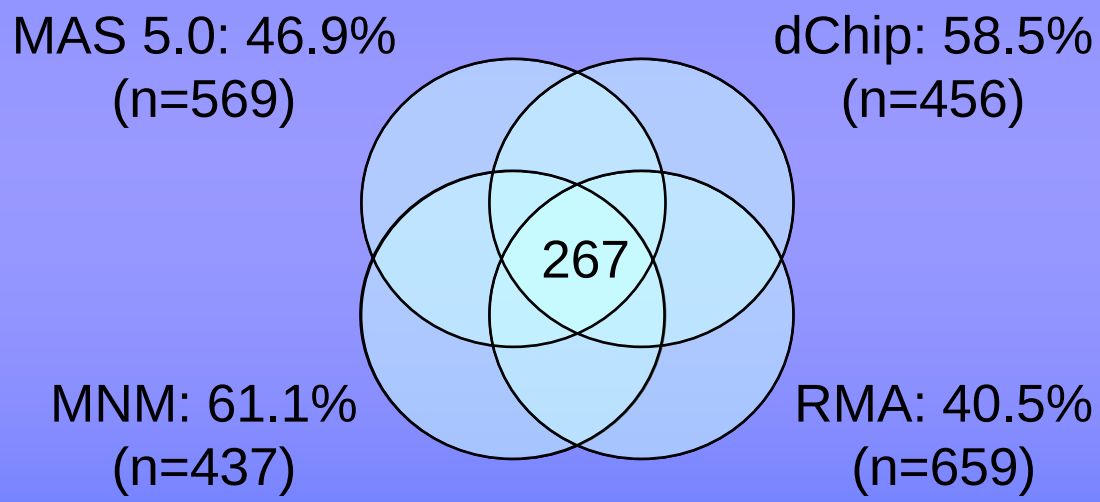


Significative fold change in RT-PCR expression:

For triplicates = 1.21 (0.82)

For simplicates = 1.39 (0.72)

Intersection of the gene lists for each method:
Modulated genes

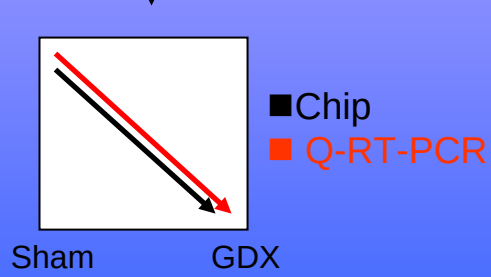


Validation by Q-RT-PCR of the gene list obtained by the four methods.

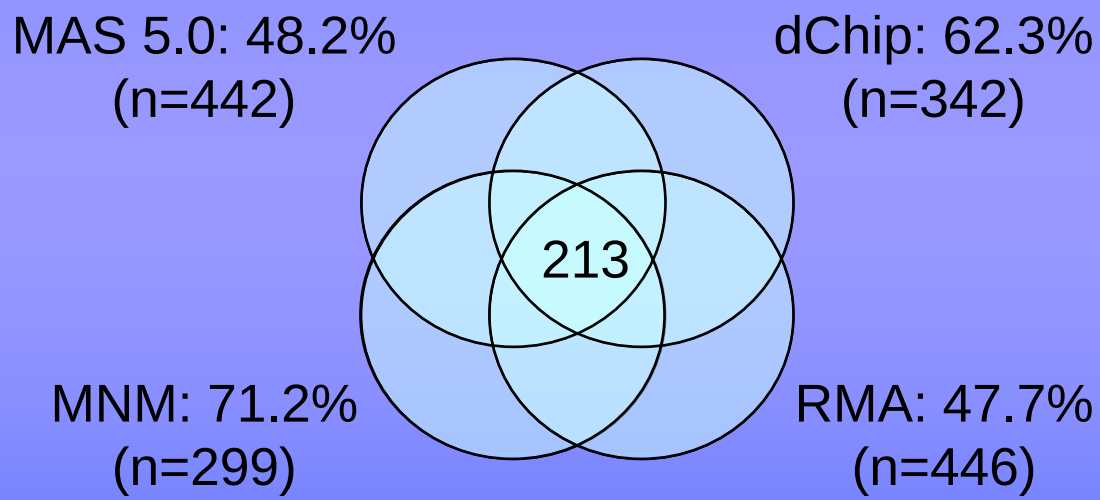
Method	LFCM Equation	Total number of modulated genes (unvalidated)	number of genes verified	Genes validated by Q-RT-PCR (ratio > 1.4)	%	% of TOTAL (n=620)	Validated in the oposite sense	%
MAS 5.0	1.8+62.0	569	560	442	**78.9**	71.0	16	3.0
dChip	1.58	456	451	342	**75.8**	55.0	16	3.5
RMA	1.62	659	645	446	**69.1**	72.0	18	2.9
MNM	1.69	437	426	299	**70.2**	48.0	11	2.8
TOTAL		**855**	839	**620**				

Validation by Q-RT-PCR of the gene list obtained by the four methods.

Method	LFCM Equation	Total number of modulated genes (unvalidated)	number of genes verified	Genes validated by Q-RT-PCR (ratio > 1.4)	%	% of TOTAL (n=620)	Validated in the oposite sense	%
MAS 5.0	1.8+62.0	569	560	442	**78.9**	71.0	16	3.0
dChip	1.58	456	451	342	**75.8**	55.0	16	3.5
RMA	1.62	659	645	446	**69.1**	72.0	18	2.9
MNM	1.69	437	426	299	**70.2**	48.0	11	2.8
TOTAL		**855**	839	**620**				

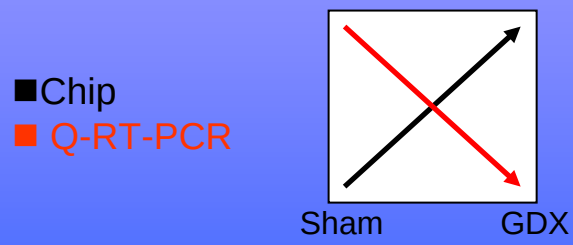


Intersection of the gene lists validated by Q-RT-PCR (ratio>1.4) for each method.



Validation by Q-RT-PCR of the gene list obtained by the four methods.

Method	LFCM Equation	Total number of modulated genes (unvalidated)	number of genes verified	Genes validated by Q-RT-PCR (ratio > 1.4)	%	% of TOTAL (n=620)	Validated in the oposite sense	%
MAS 5.0	1.8+62.0	569	560	442	**78.9**	71.0	16	3.0
dChip	1.58	456	451	342	**75.8**	55.0	16	3.5
RMA	1.62	659	645	446	**69.1**	72.0	18	2.9
MNM	1.69	437	426	299	**70.2**	48.0	11	2.8
TOTAL		**855**	839	**620**				



Each one of the four methods tested is capable of selecting between 69% to 79% of significantly modulated genes.

The intersection value for the four list includes only 31.2% of genes (267 of 855).

The RMA method gives the biggest list of modulated genes (n=659), but shows the lowest percentage of validation by Q-RT-PCR (69%).

The MAS 5.0 together with a modulated fold change method produce a list 16% smaller than the RMA method, but their percentage of validation (78.9%) is the highest of the four.

The four methods include in their lists a similar percentage (~3%) of genes validated in opposite sense of modulation.

A similar percentage of validation is obtained for the common list (intersection of 4 methods) and for the list extracted by MAS 5.0 (80.6 vs 78.9%)

Conclusions

Under the conditions utilized:

Each method could recognize only a percentage (sub-population ?) of the whole modulated genes.

The combination of several methods does not necessarily improve the selection of modulated genes.

MAS 5.0 in combination with the LFCM method seems to be the best combination among the tested models.

A further analysis of data aiming to find clues on the selectiveness of each method is actually in progress.

Microarrays Team:

Nathalie Saulnier
Annick Ouellet
Marie-Eve Marcoux
Josée Parent
Alain St-Pierre
Ézéquiel Calvo

Bioinformatics Team:

Jean Morissette

Pascal Beleau
Héléne Boucher
Erick Chamberlain
Ariel Chernomoretz
Astrid Deschenes
Benoit Hébert
Sonia Jean
René Paradis

Bioinformatics Team:

Jean Morissette

Pascal Beleau
Héléne Boucher
Erick Chamberlain
Ariel Chernomoretz
Astrid Deschenes
Benoit Hébert
Sonia Jean
René Paradis

(an Apple user)

Merci / Thanks

Application of the LFC equation:

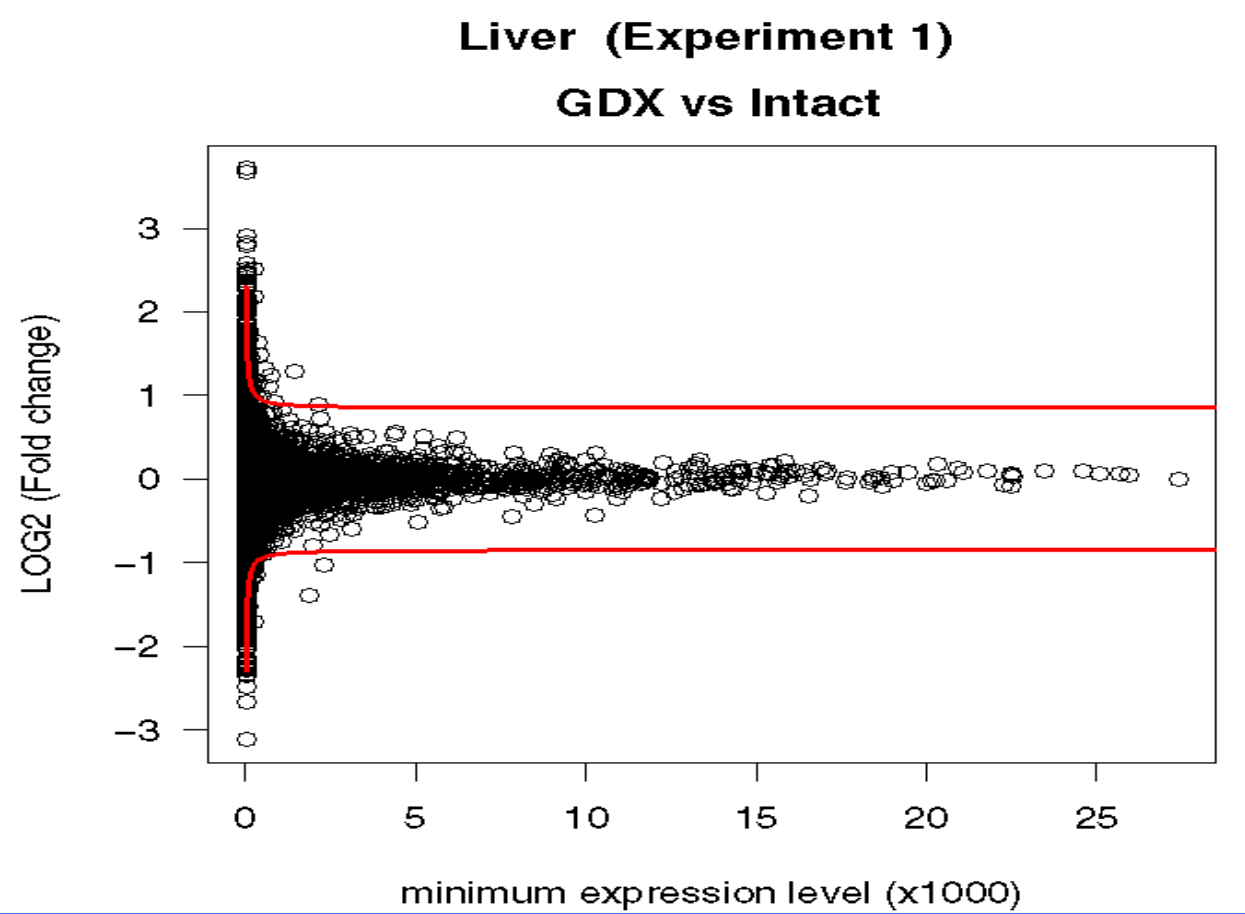
- Two different conditions are compared.
- The minimum expression value X is evaluated as well as the fold change (FC)
- If FC > LFC , the fold change is above the LFC curve, therefore the gene is considered significantly modulated.

How the LFC (limit fold change) equation was obtained:

- Pick 2 replicate chips *within* an experimental condition
- Calculate the expression values
- Let E_{kr} be the expression value of the k-th gene for the r-th replicate
- Let the fold change $FC = \max(E_{k1}, E_{k2}) / \min(E_{k1}, E_{k2})$
- Genes are ordered according to expression levels
- Expression range is divided into bins of 200 genes
- In each bin, the 99.9^{th} percentile fold change is determined
- Relate the mean expression of each bin with the 99.9^{th} percentile fold change using the equation: $Y = a + b / X$
- Parameters a and b are estimated by least squares

The previous steps are repeated for all the possible pairs of replicates, from different conditions and different tissues. The final equation is obtained with the average values of the parameters a and b over all pairs

LFC equation
Y = 1.8 + 62.0/X



Listes de gènes LFCM (Y= a + b / X)

La limite significative du taux d'expression (LFC) diminue en fonction de l'intensité afin de sélectionner les gènes exprimés différentiellement.

Pour ce faire, l'équation: Y= a + b / X a été utilisée,

où X correspond à la valeur minimale de l'expression des gènes présents dans les deux conditions et Y correspond à la limite du taux d'expression. Les paramètres « a » et « b » ont été estimés grâce à la distribution des ratios préalablement calculés à partir de réplicats de biopuces.

La limite résultante, Y=1.8+62.0/X donne un taux de faux positifs approximativement constant de 0.1%.

En pratique, tous les gènes possédant un taux d'expression au-dessus de cette courbe sont considérés comme étant significativement modulés. Cette liste permet une meilleure sélection des gènes modulés car elle tient compte de l'intensité de fluorescence et du taux d'expression d'un gène ce qui n'est pas le cas lorsque la sélection des gènes est effectuée seulement en fonction du taux d'expression.