Design and exploitation of a versatile Arabidopsis whole-Genome Tiling Array

Tiling-array data: analysis and visualization

Caroline Bérard[1], Sandra Dèrozier[2], Sandrine Balzergue[2], Tristan Mary-Huard[1], Francois Roudier[3], Stéphane Robin[1], Alain Lecharny[2], Vincent Colot[3], Michel Caboche[2], Sébastien Aubourg[2] and Marie-Laure Martin-Magniette[1,2]

[1] *UMR AgroParisTech/INRA MIA 518, Equipe statistique et génome, Paris.*
[2] *Unité de Recherche en Génomique Végétale (URGV), Evry.*
[3] *Institut de Biologie de l'ENS, Equipe Epigénétique et Epigénomique Végétale.*

ANR/Genoplante *Tiling Array* Genome (TAG) project

Goal of the ANR Genoplante TAG Project

- Design of a *tiling-array* covering the *Arabidopsis thaliana* whole genome.

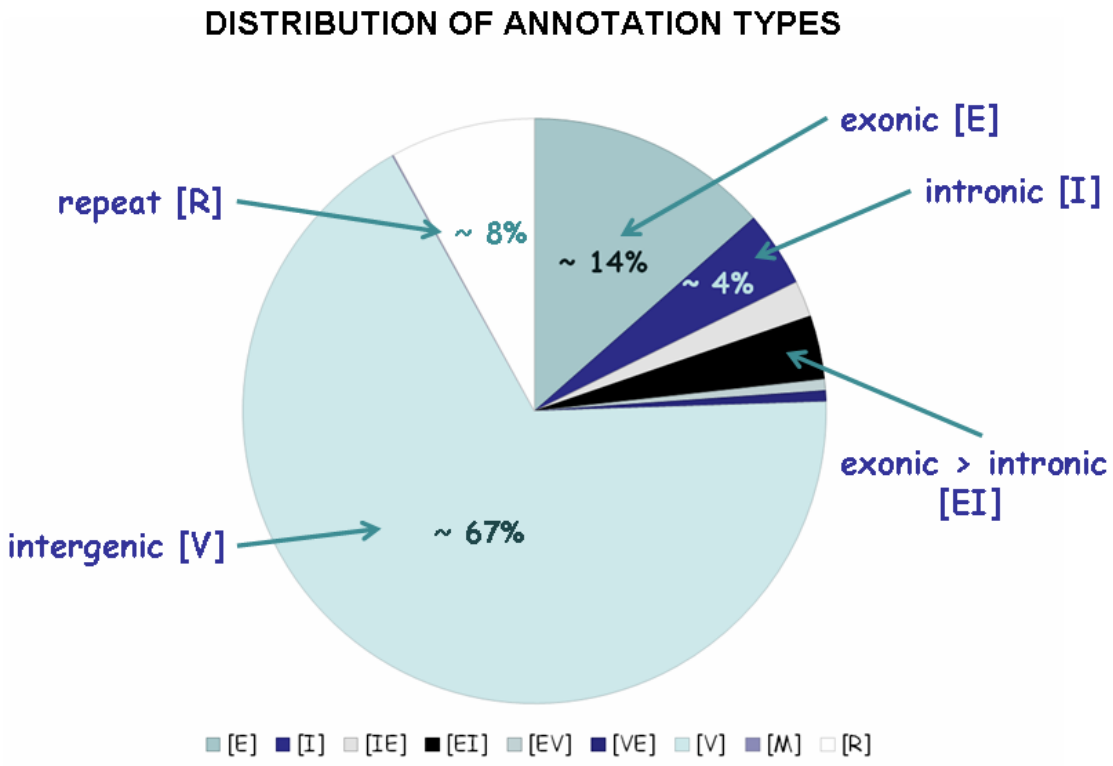
- Different types of application
  - **Transcriptome**: detection of transcripts, conditions of gene expression
  - **ChIP-chip**: study control mechanism of gene expression (DNA methylation, histone modifications, transcription factor)

$$\longrightarrow \text{Development of adapted statistical methods}$$

- Visualization of probe features and integration of the statistical results in the FLAGdb++ environment.

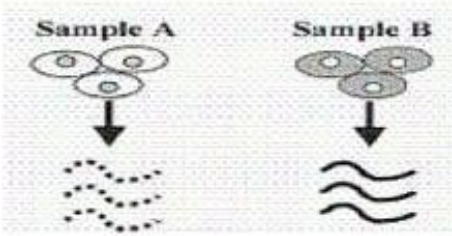
Tiling-array features

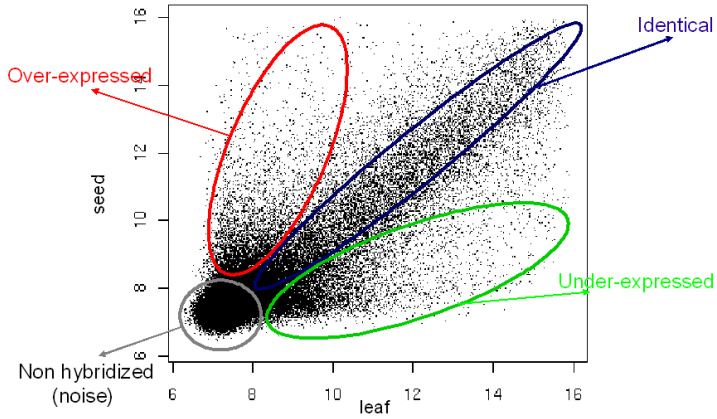
- 717.246 probes per strand, Resolution of 160 bp.
- NimbleGen probes
  - Variable length (between 50 and 75 nt)
  - Constant TM $\sim 76°$



An unsupervised Classification Problem

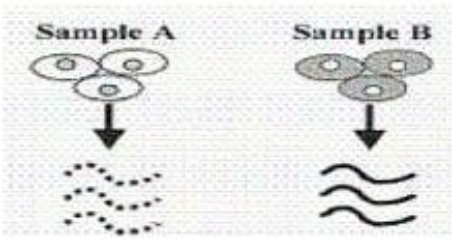
- **Transcriptome**: noise, identical, over or under expressed (4 groups)

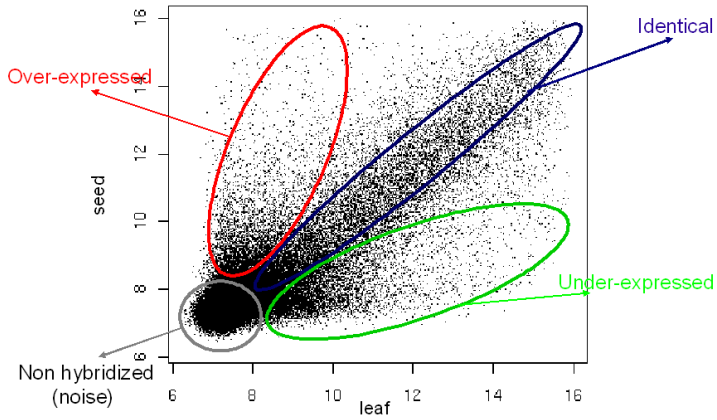




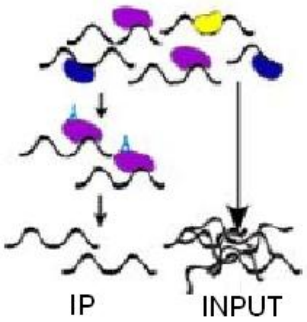
An unsupervised Classification Problem

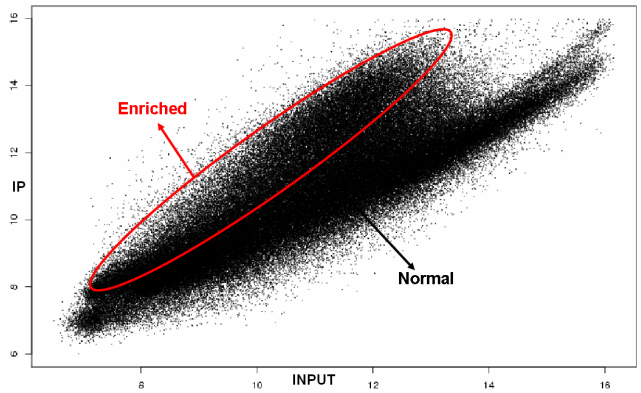
- **Transcriptome**: noise, identical, over or under expressed (4 groups)





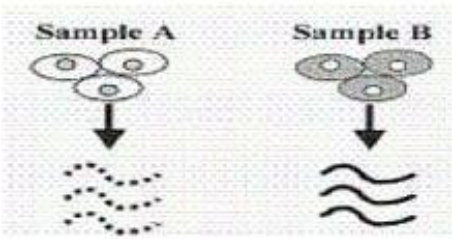
- **ChIP-chip**: enriched, normal (2 groups)

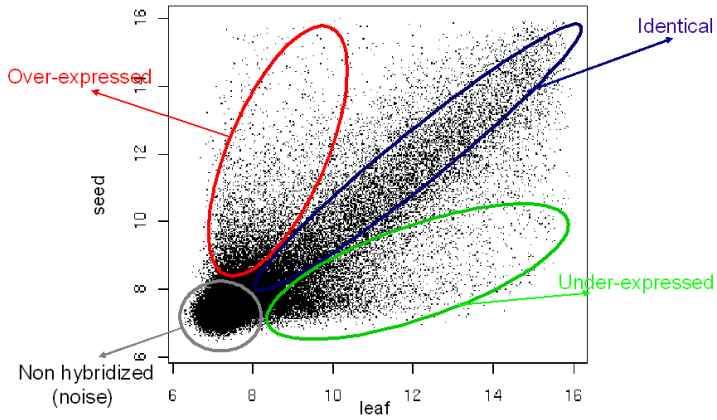




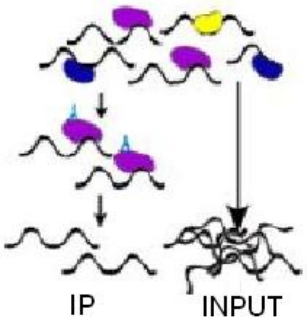
An unsupervised Classification Problem

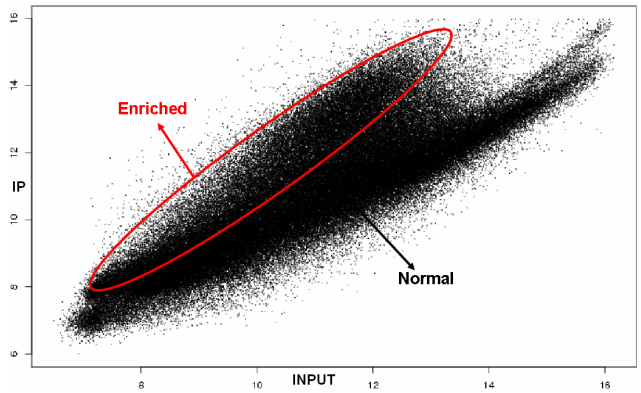
- **Transcriptome**: noise, identical, over or under expressed (4 groups)





- **ChIP-chip**: enriched, normal (2 groups)





→ **Find the status of the probe**

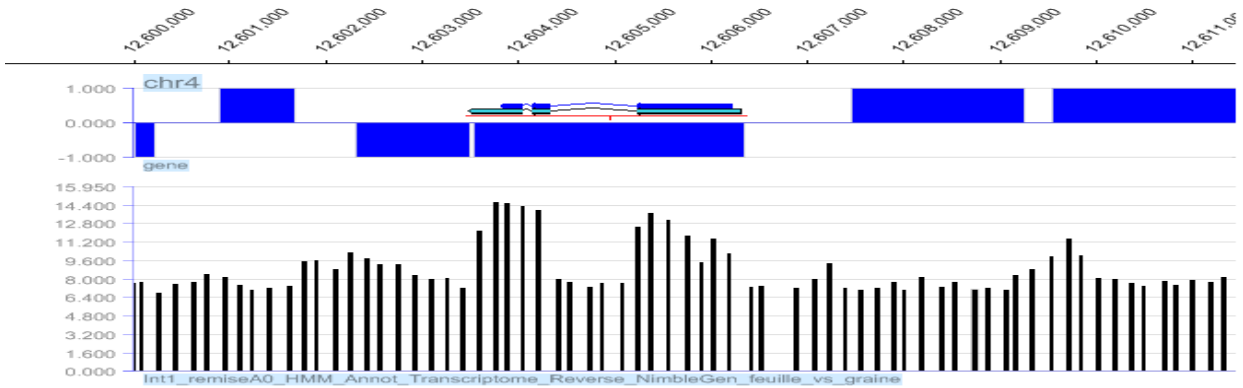
Available informations

- Position of the probes along the genome $\rightsquigarrow t$
- Structural annotation $\rightsquigarrow C_t$

Available informations

- Position of the probes along the genome $\rightsquigarrow t$
- Structural annotation $\rightsquigarrow C_t$

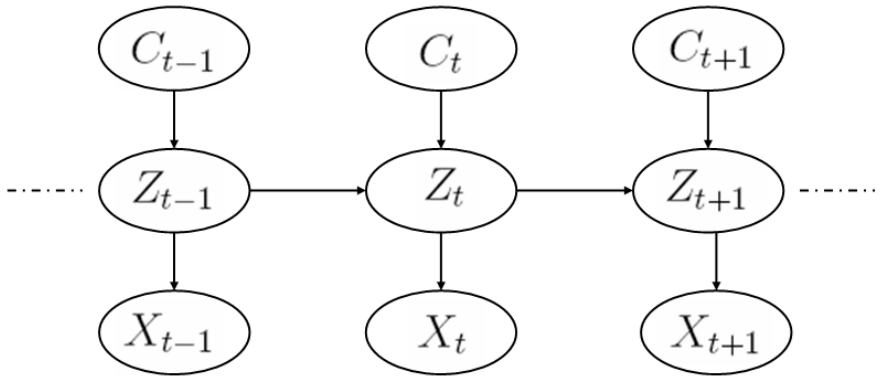
- Visualization of the signal intensity



$\rightarrow$ **Dependence between neighboring probes**

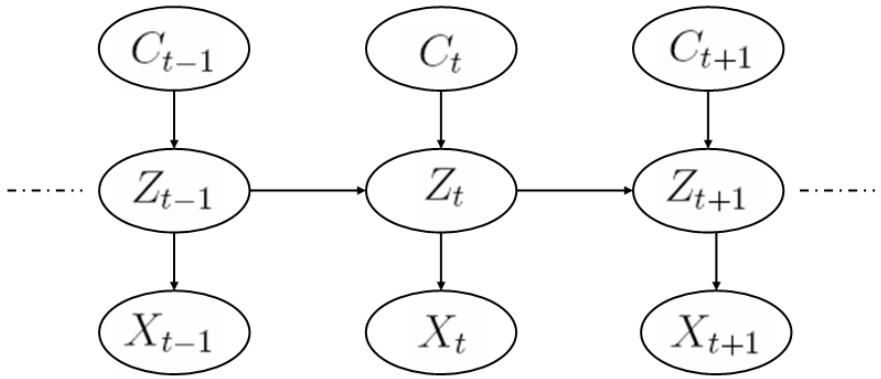
General Modeling

X = Data, Z = unknown status, C = annotation, t = position



General Modeling

X = Data, Z = unknown status, C = annotation, t = position



- **Transcriptome**: $X_t = (\text{Int}_1, \text{Int}_2)$
  - Symmetrical: Bidimensionnal Gaussian mixture
- **ChIP-chip**: $X_t = (IP, Input)$
  - Non symmetrical: Mixture of regressions

Transcriptome data: Model with HMM and Annotation

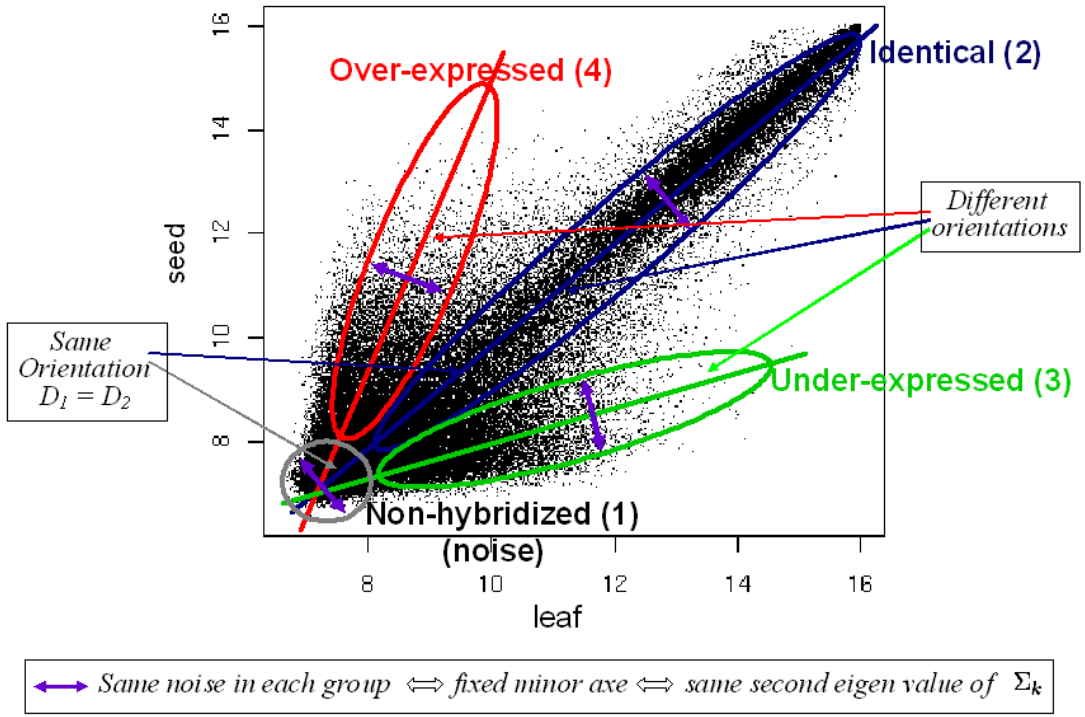
- Model
  - C_t = annotation of the probe t (intron, exon, intergenic, ...)
  - Z_t (status of the probe) $\sim$ Markov chain
  - $P(Z_t = l | Z_{t-1} = k, C_t = p) = \pi_{kl}^{C_t}$ $\rightarrow$ one transition matrix for each annotation category

Transcriptome data: Model with HMM and Annotation

- Model
  - C_t = annotation of the probe t (intron, exon, intergenic, ...)
  - Z_t (status of the probe) $\sim$ Markov chain
  - $P(Z_t = l | Z_{t-1} = k, C_t = p) = \pi_{kl}^{C_t}$ $\rightarrow$ one transition matrix for each annotation category
  - X : observed signal (Int_1, Int_2)
  - $(X_t | Z_t = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$
    μ = mean vector $\rightsquigarrow$ center of the ellipse
    Σ = variance matrix $\rightsquigarrow$ shape, orientation, volume of the ellipse
  - **Bidimensional Gaussian Mixture:**

$$\begin{aligned} f(x, \psi) &= p_0 \phi(x | \mu_0, \Sigma_0) + p_1 \phi(x | \mu_1, \Sigma_1) \\ &\quad + p_+ \phi(x | \mu_+, \Sigma_+) + p_- \phi(x | \mu_-, \Sigma_-) \end{aligned}$$

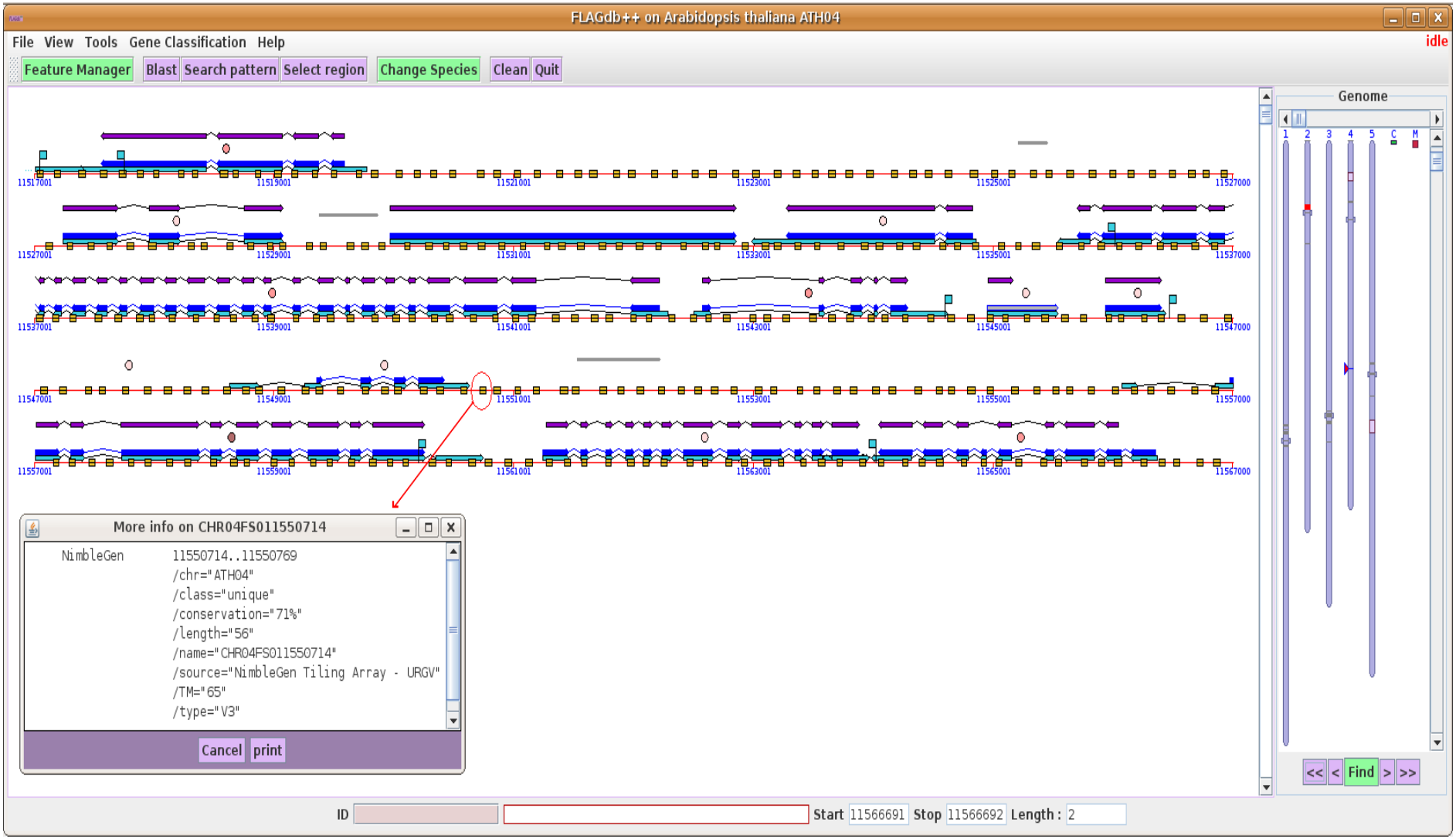
- Assumptions on the ellipse form
  - Noise and Identical groups have same orientation
  - Identical variance in each group



  - Maximum likelihood estimation with constraints on the variance matrices using the EM algorithm

- Posterior probability: $\tau_t = \Pr\{Z_t = 1|X\}$

View of NimbleGen probes in FLAGdb++



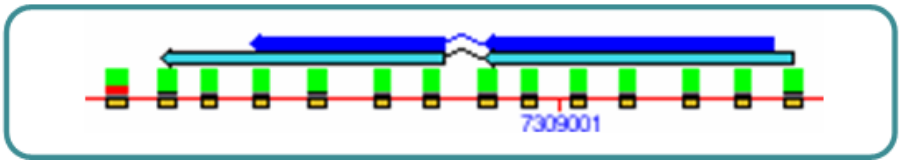
Integration of statistical results in FLAGdb++

Statistic file
(tabulate format)

"ID"	"p1"	"p2"	"p3"	"p4"	"Int1"	"Int2"
"CHR01FS018836436"	1.86099126178244e-37				0.999851378593742	1.11913350032989e-10
0.000148621294344768		10.3134558599527			10.9593445849487	
"CHR01FS019035569"	0.775876513612473		0.0189380091820311		0.166126154489924	
0.0390593227155723	7.99795121957173		7.58800446360571			
"CHR01FS002137107"	1.23449350953109e-11				0.00354325348988523	6.66069694294767e-07
0.996456080428076	8.44007661970305		9.88017752977436			

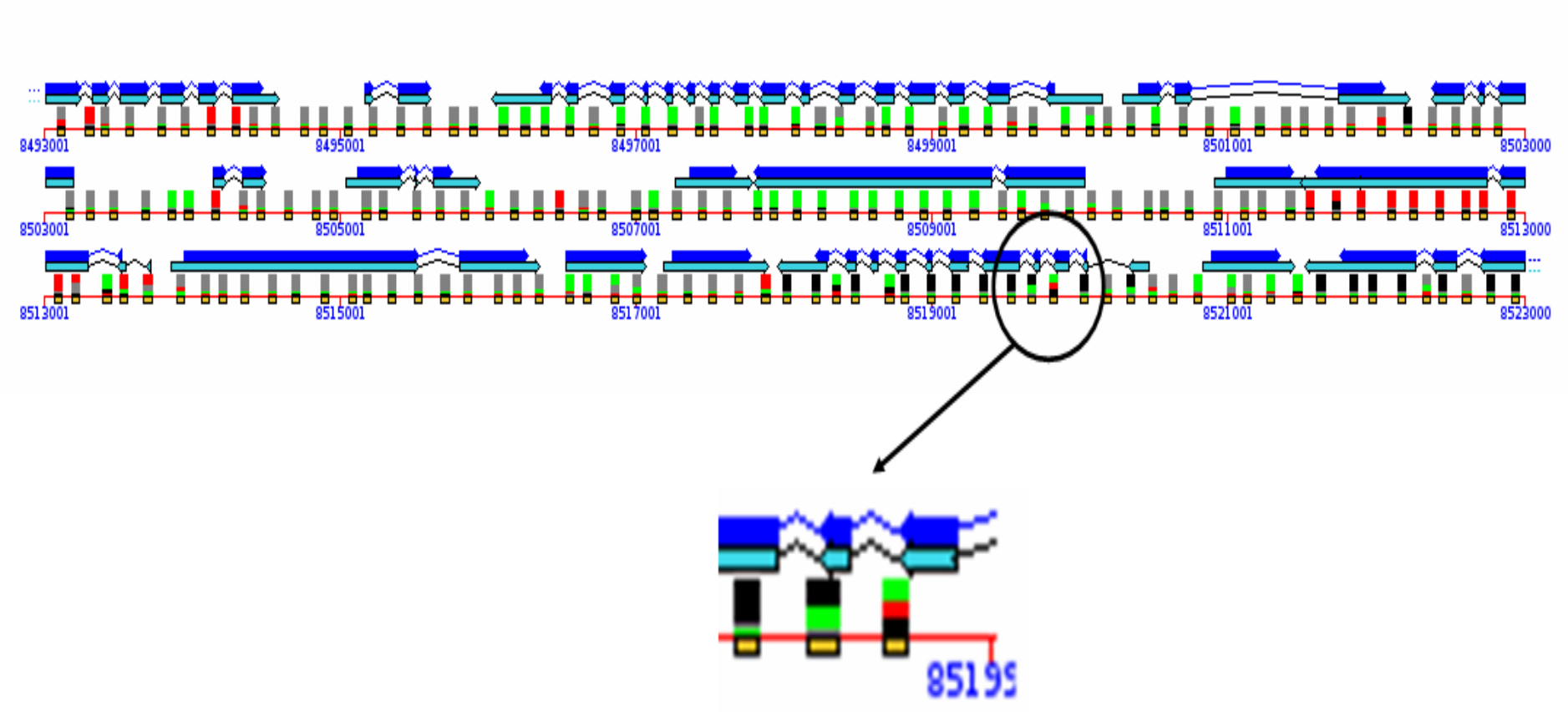
Insertion of statistical data in a local database *Apache Derby* via the interface of *FLAGdb++*

Reading of statistical data on the graphic interface of *FLAGdb++*



Application on Transcriptome Data

Comparison of expression between seed 10 days after pollinisation and leaf of *Arabidopsis thaliana*



The colors of a probe depend on posterior probabilities.

Parameter estimation (in %)

Transition matrix of **intergenic** group:

	Noise	Ident.	Under-exp	Over-exp
Noise	87	1	7	5
Ident.	95	3	1	1
Under-exp	77	1	19	3
Over-exp	75	2	5	18

Proportions
84
1
9
6

Transition matrix of **exonic** group:

	Noise	Ident.	Under-exp	Over-exp
Noise	83	14	3	0
Ident.	2	90	6	2
Under-exp	7	5	87	1
Over-exp	8	6	1	85

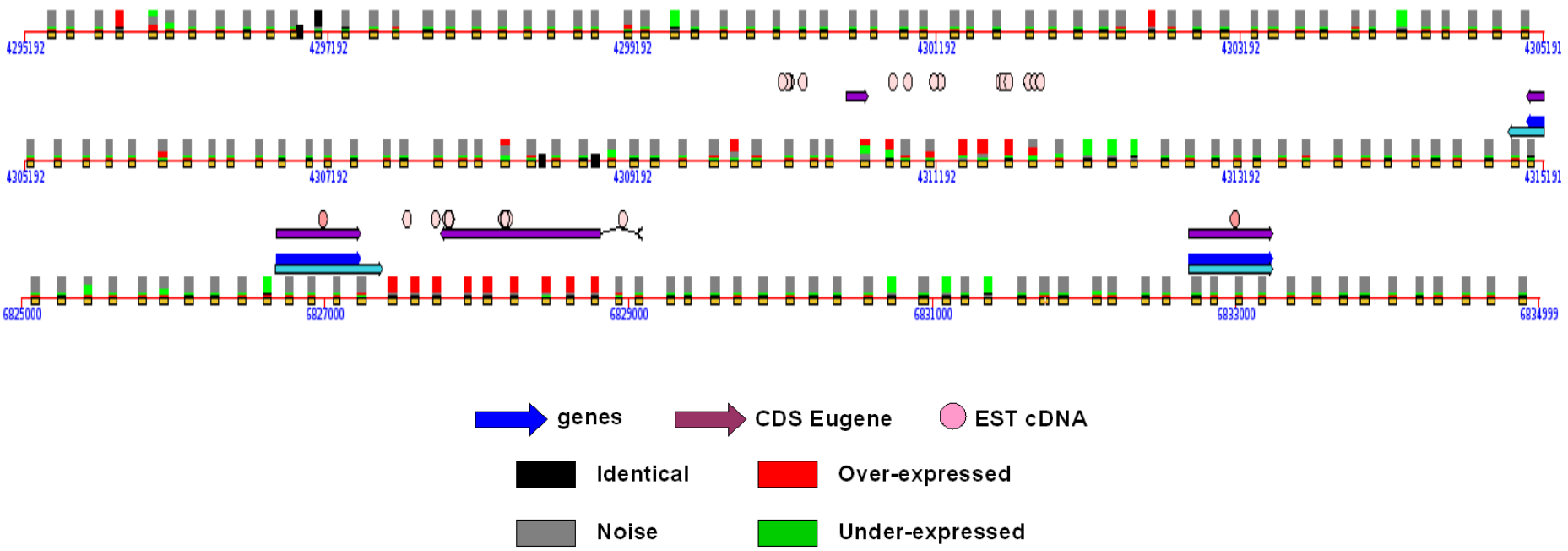
Proportions
22
41
23
14

Transition matrix of **intronic** group:

	Noise	Ident.	Under-exp	Over-exp
Noise	87	2	8	3
Ident.	89	0	1	10
Under-exp	55	2	43	0
Over-exp	96	1	0	3

Proportions
60
7
24
9

Genome exploration to improve structural annotation



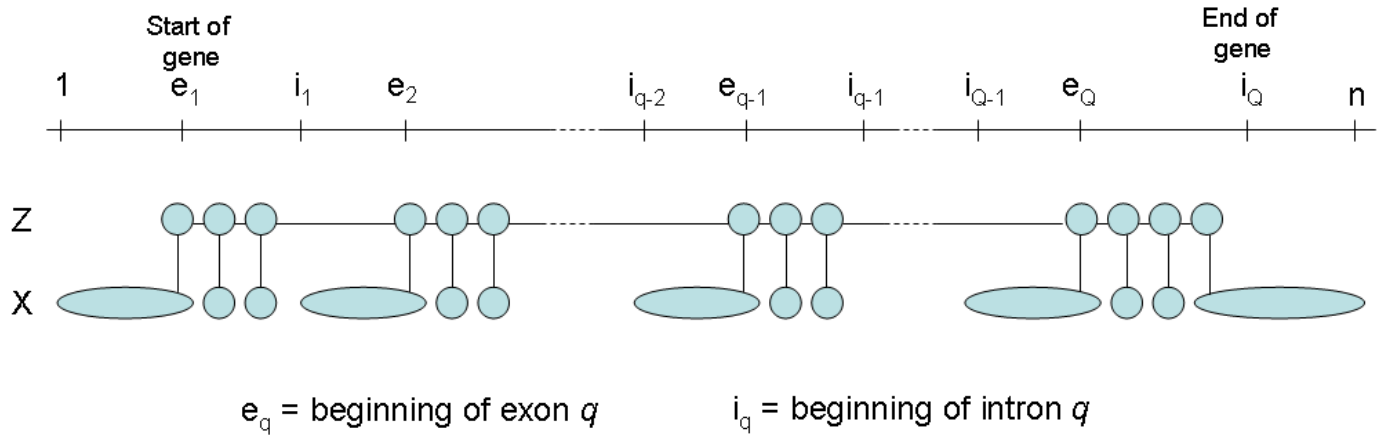
- Number of regions with expressed probes in intergenic

Nb probes	2	3	4	>5
Forward	798	167	70	88
Reverse	786	194	86	90

Probe classification $\longrightarrow$ Gene classification

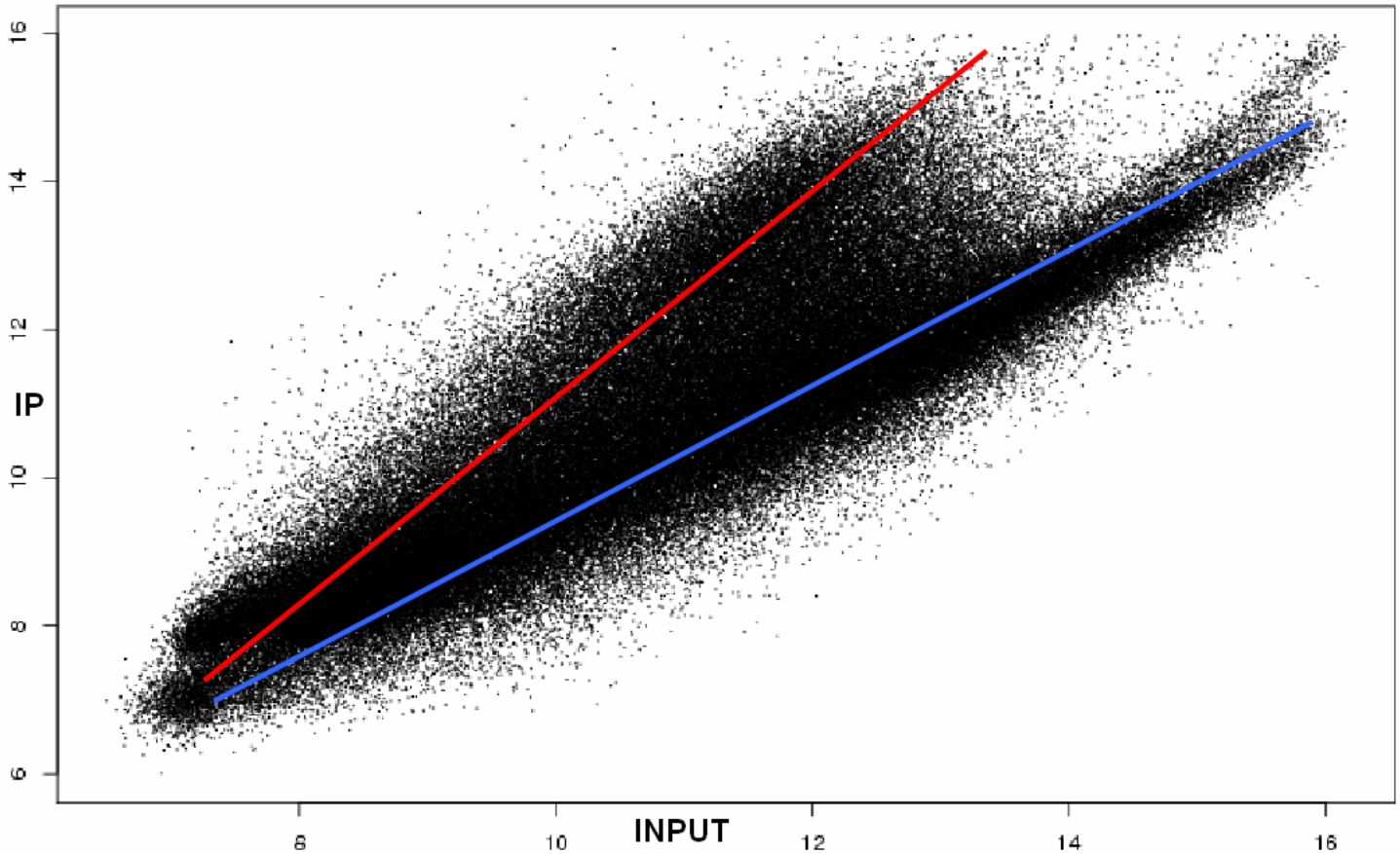
Gene = region covered by several probes $\rightsquigarrow$ interesting biological entity

$$P(\text{for all exon probes } s \text{ for gene } g, Z_s = k|X)$$



$\rightsquigarrow$ Calculations use the Forward step of the Forward/Backward algorithm.

ChIP-chip data



ChIP-chip data: Mixture of regressions

- Model

$$\Pr\{Z_t = 1\} = \pi, \qquad \Pr\{Z_t = 0\} = 1 - \pi.$$

The relation between IP and Input depends on the status of the probe:

$$IP_t = \begin{cases} a_0 + b_0 Input_t + E_t & \text{if } Z_t = 0 \text{ (normal)} \\ a_1 + b_1 Input_t + E_i & \text{if } Z_t = 1 \text{ (enriched)} \end{cases} \qquad V(IP_t) = \sigma^2$$

- Maximum likelihood estimation using the EM algorithm
- Posterior probability: $\tau_t = \Pr\{Z_t = 1|X\}$

ChIPmix: mixture model of regressions for two-color ChIP-chip analysis.

M-L. Martin-Magniette, T. Mary-Huard, C. Bérard and S. Robin. *Bioinformatics* (2008)

Classification: control of false positives

- ChIP-chip: 2 groups

  Controlling false detections: We want to control the probability for the τ_t of a normal probe to fall above the classification threshold.
  For a fixed risk α we calculate the threshold s such that

$$s: \qquad \Pr\{\tau_t > s \mid t \text{ normal}, \log(\text{Input}) = X_t\} = \alpha$$

  and if $\tau_t > s$ then the probe t is classified as 'enriched'.

  The threshold s depends on both α and the log-Input X_t.
  This control focuses on misclassifications in 'enriched' group.

Conclusions

- Versatile Arabidopsis whole-Genome Tiling Array.
- Protocols developed at URGV for Transcriptome, ChIP-chip and CGH.

$\rightsquigarrow$ balzerg@evry.inra.fr

- General modeling of the hybridized signal.
  - Use the whole available information of the probes.
  - Comprises independent mixture (no spatial dependence) and homogeneous HMM (no annotation).
  - Can be also used with one-color *tiling arrays* to compare 2 samples.
  - Can be also used for ChIP-chip data to compare two IP samples.
  - Package R available upon request.

$\rightsquigarrow$ caroline.berard@agroparistech.fr

- Visualization in FLAGdb++.

$\rightsquigarrow$ http://urgv.evry.inra.fr/FLAGdb