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

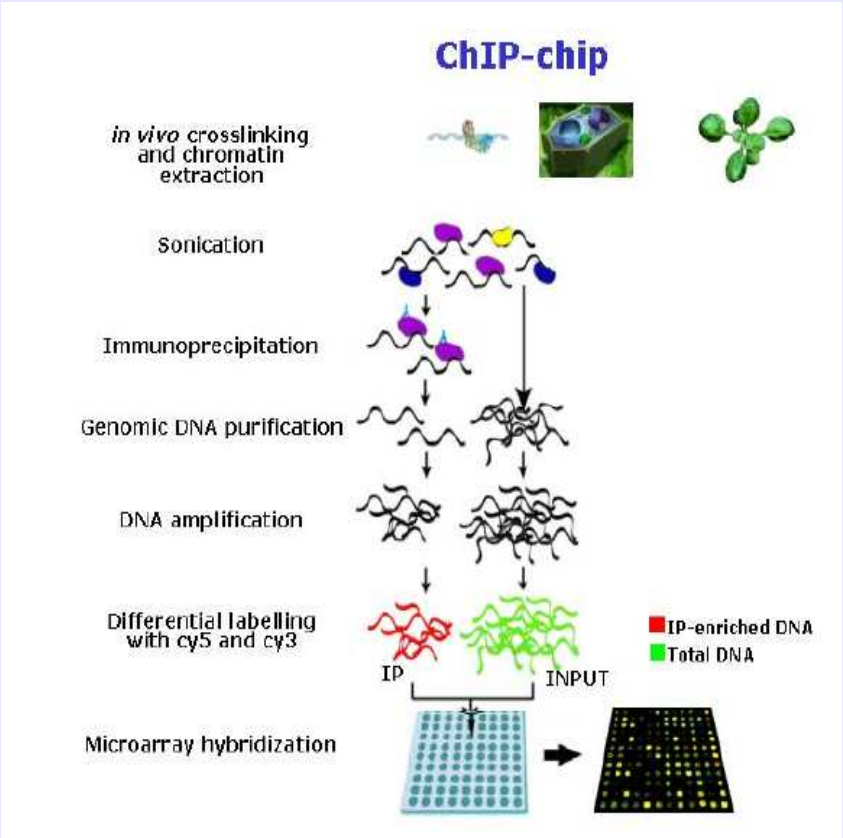
Tristan Mary-Huard, Marie-Laure Martin-Magniette, Caroline Berard, Stephane Robin

ChIPmix: Mixture model of regressions for ChIP-chip experiment analysis

T. Mary-Huard, M.-L. Martin-Magniette, C. Bérard, S. Robin

UMR AgroParisTech/INRA 518
UMR INRA/CNRS/UEVE
This work was supported by the TAG ANR/Genoplante project.

ChIP-Chip experiment



IP sample : immuno-precipited DNA
INPUT sample : total DNA

Both samples are co-hybridized on a same array

Aim to determine which probes have an IP signal significantly higher than the INPUT signal.

Aim to investigate protein-DNA interactions known to be involved in control mecanisms of gene expression.

Unsupervised Classification Problem

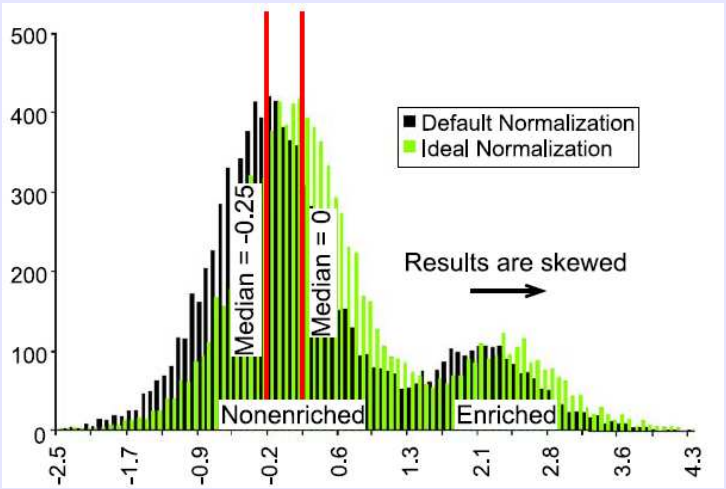
Goal : According to the IP and Input signals, we have to

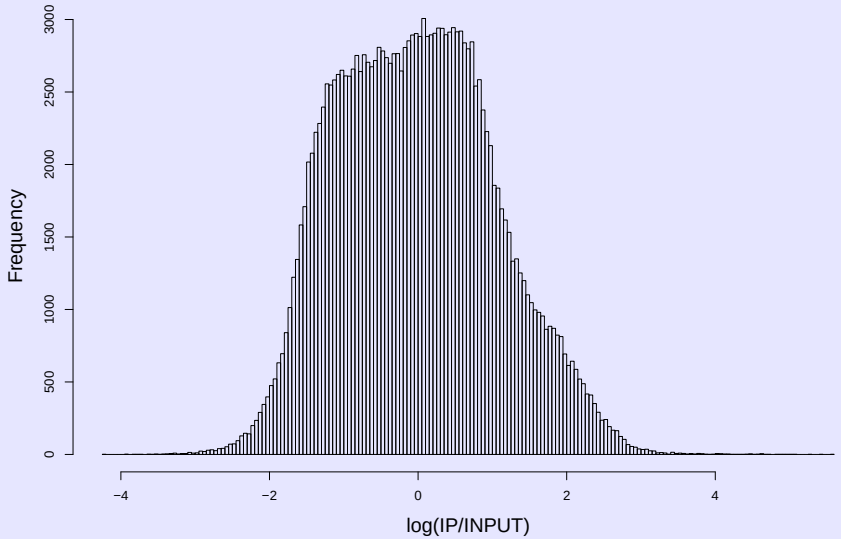
- classify each probe into the enriched group or into the normal group,
- avoid numerous false detections (false enriched probes).

Bibliography :

- methods based on the spatial structure of the data (sliding windows, peak detection, HMM),
- methods considering that the whole population can be divided into two groups (mixture models).

About logratios...





Good case (Buck & Lieb, 2004)

Bad case

All the methods work on the logratio $\log(IP/Input)$

$\Rightarrow$ suppose that the logratio distribution is informative about the probe status...

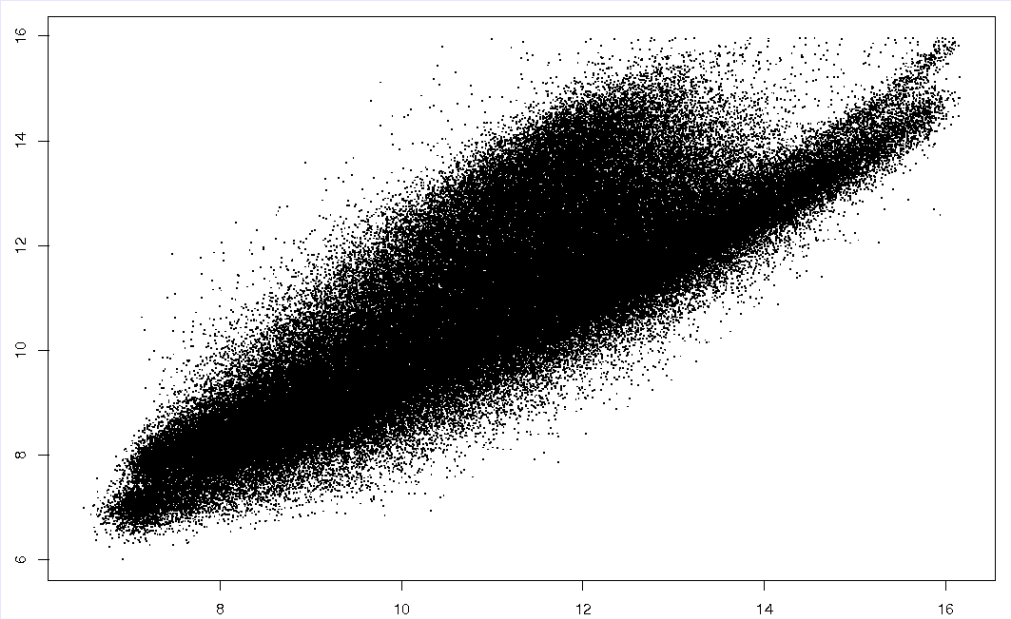
About logratios...

Why a linear relationship ?

⋆ Technical difficulties to obtain the IP sample

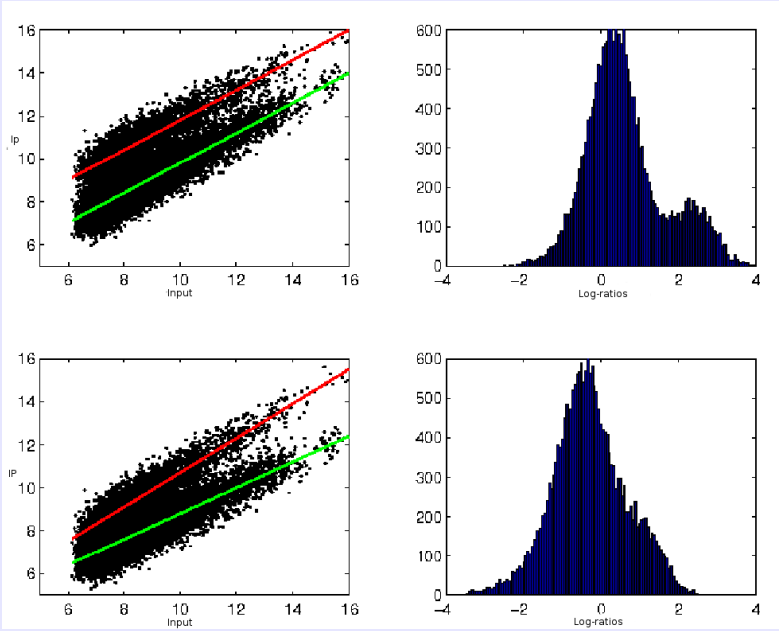
⋆ Possible cross-hybridization phenomena

A closer look shows that the relation between IP and Input may differ between the 'normal' group and 'enriched' group.



IP versus Input for the *Arabidopsis* dataset

About logratios on synthetic data



Top : Two populations with linear relationship and equal slopes.
The corresponding logratio histogram is bimodal.
Bottom : Two populations with linear relationship but different slopes.
The corresponding logratio histogram is unimodal.

It is worth working directly with the 2 signals rather than the logratio

Mixture model of regressions

We assume that each probe i has probability π to be enriched :

$$\Pr\{\text{Probe } i \text{ enriched}\} = \pi, \qquad \Pr\{\text{Probe } i \text{ normal}\} = 1 - \pi,$$

and that the relation between log-IP (Y_i) and log-Input (X_i) depends on the status of the probe :

$$Y_i = \begin{cases} a_0 + b_0 X_i + E_i & \text{if } i \text{ is normal} \\ \\ a_1 + b_1 X_i + E_i & \text{if } i \text{ is enriched} \end{cases}$$

The marginal distribution of Y_i for a given x_i is

$$(1-\pi)\phi_0(Y_i|x_i) + \pi\phi_1(Y_i|x_i),$$

where $\phi_j(\cdot|x)$ stands for the probability density function of a Gaussian distribution with mean $a_j + b_j x$ and variance σ^2.

Parameter Estimation and Probe Classification

Task. We have to estimate

- the proportion π of enriched probes,
- the regression parameters : intercepts (a_0, a_1), slopes (b_0, b_1) and the variance σ^2.

Algorithm. This can be done using the E-M algorithm which alternates

E-step : prediction of the probe status given the parameters,

M-step : estimation of the parameters given the (predicted) probe status.

Posterior probability. The status prediction is based on the posterior probability τ_i :

$$\tau_i = \Pr\{i \text{ enriched} \mid X_i, Y_i\} = \frac{\pi\phi_1(Y_i|x_i)}{(1-\pi)\phi_0(Y_i|x_i) + \pi\phi_1(Y_i|x_i)},$$

This probability provides a *probe classification rule*.

Limiting False Detections

Maximum A Posteriori (MAP) rule. Probes are classified into their most probable class, using the 50% threshold :

$$\begin{aligned} \tau_i &\geq 50\% \quad \Rightarrow \quad i \text{ classified as 'enriched'}, \\ \tau_i &< 50\% \quad \Rightarrow \quad i \text{ classified as 'normal'}. \end{aligned}$$

Misclassifications into 'enriched' or 'normal' groups have the same cost.

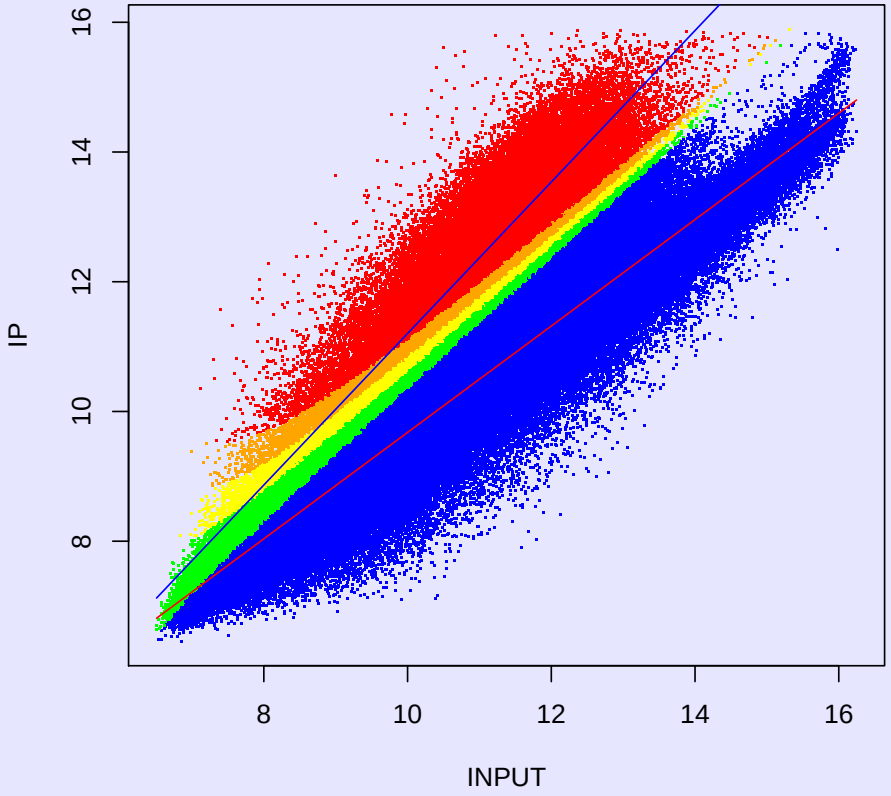
Controlling false detections. We want to control the probability for the τ_i of a normal probe to fall above the classification threshold.

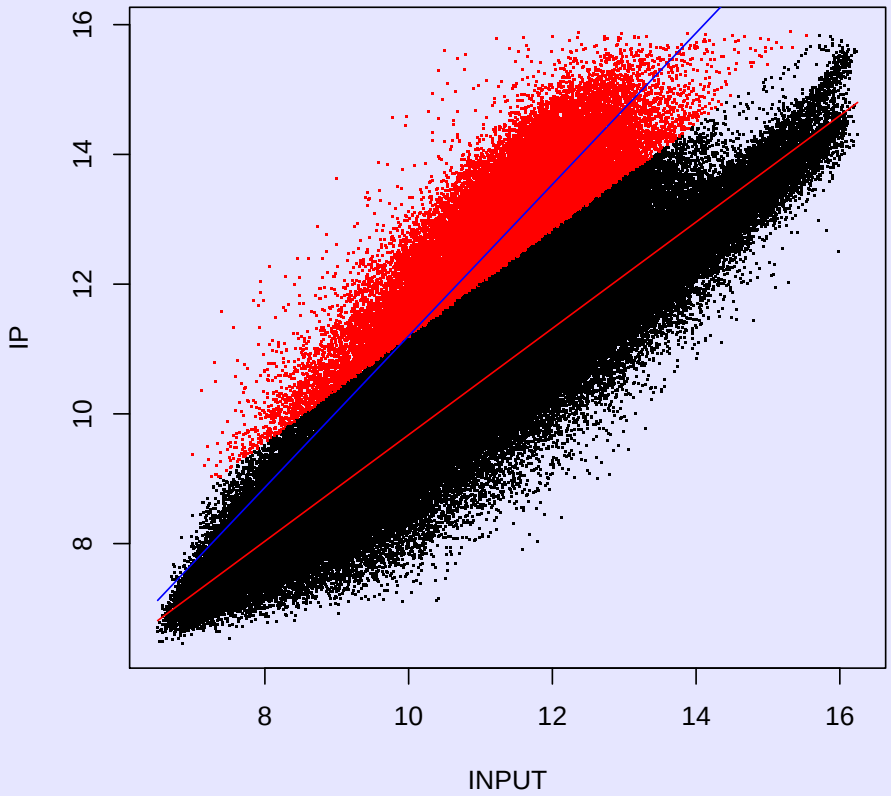
For a fixed risk α we calculate the threshold s such that

$$s: \qquad \Pr\{\tau_i > s \mid i \text{ normal}, \log(\text{Input}) = X_i\} = \alpha$$

and if $\tau_i > s$ then i is classified as 'enriched'.
The threshold s depends on both α and the log-Input X_i.
This control focuses on misclassifications in 'enriched' group.

Probe classification rule





Left : Estimated regression lines and posterior probabilities.

Rigth : Classification of the probes into the enriched (red) or normal (black) group, with a α=0.01% threshold.

Application to real data

Promoter DNA methylation in the human genome

Data (Weber *et al.* 2007)

⋆ NimbleGen tiling array of promoter regions of 15 609 human genes.

⋆ Each promoter region has 15 probes.

⋆ We consider the Intermediate CpG class of promoters (2056).

Results of ChIPmix

⋆ A high proportion of enriched probes ($\widehat{\pi} \approx 80\%$).

⋆ 403 of the 460 promoter regions found by Weber have 5 enriched probes or more.

⋆ 38 new promoter candidates with 9 enriched probes or more.

⋆ 1 promoter region found by Weber has no enriched probe.

Application to real data

Histone modification of *Arabidopsis thaliana*

Data

⋆ NimbleGen tiling array of very high density, more than 1,000,000 probes (≈200,000 per chromosome)

⋆ Two biological replicates with dye-swap, normalized according to Kerr et al., 2002

Results of ChIPmix

⋆ At level $\alpha = 0.01$, $\sim 20\%$ of the probes are enriched on each chromosome.

⋆ Overlap of 2/3 between the two biological replicates.

⋆ Probes are clustered in genomic regions.

⋆ For chromosome 4 : $\widehat{\pi} = 0.28$, $\widehat{b}_0 = 0.75, \widehat{b}_1 = 1.05$

Comparison with two other methods

Annotation

NimbleGen

ChIPmix

ChIPOTle

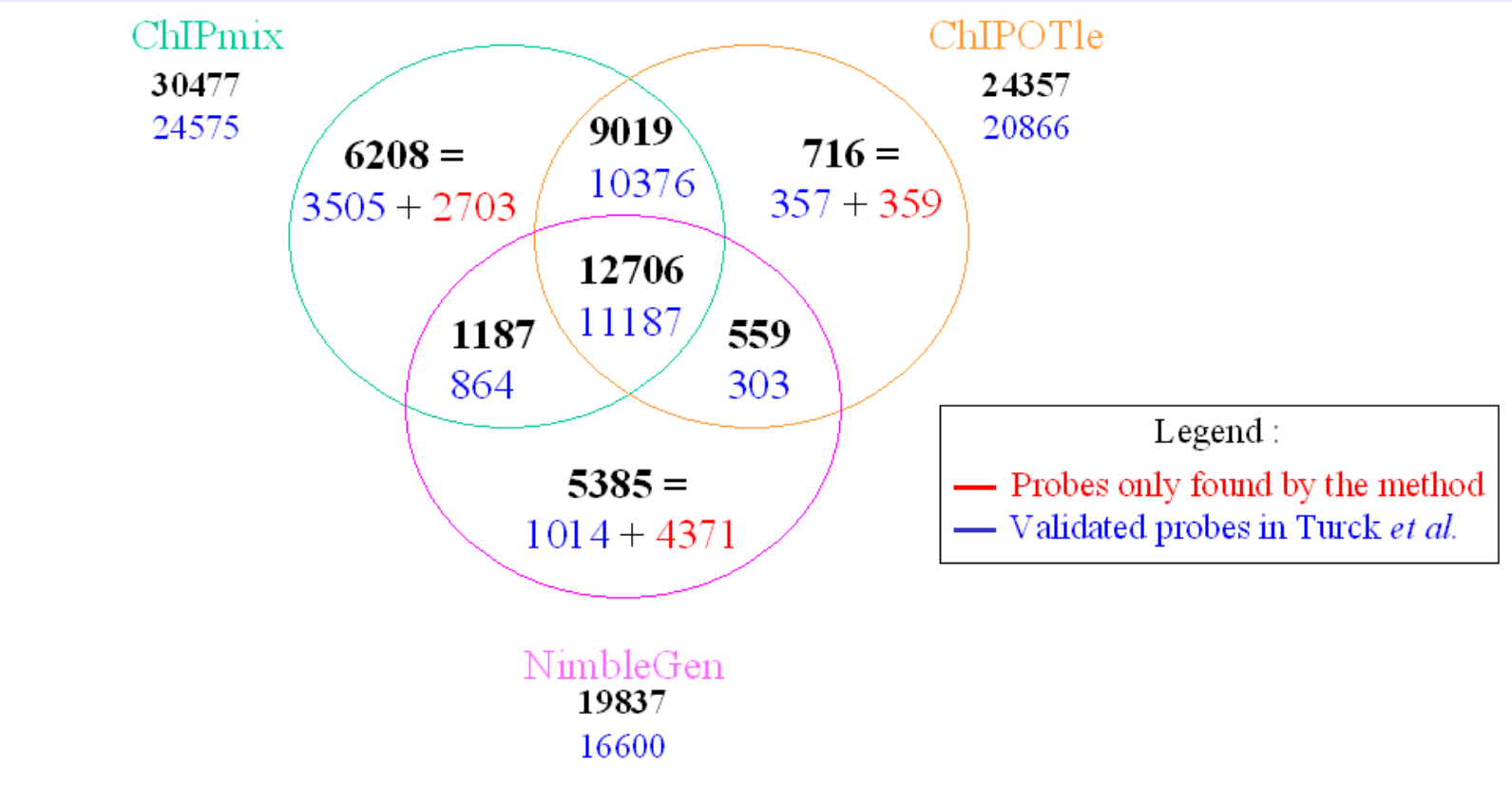
NimbleGen : finds smaller genomic regions, and fail to identify some targets.

ChIPOTle : agrees with ChIPmix, not with NimbleGen.

Comparison with results of Turck *et al.* (2007)

Home-made tiling array of chr 4 of *Arabidopsis thaliana* with the same sample

More than 75 % of the probes declared enriched with the NimbleGen's array cover genomic regions already found in Turck *et al.* (2007).



For NimbleGen, less than 20% of the specific probes are validated targets. In contrast, for ChIPmix, more than 55% of the specific probes

Extension to the analysis of biological replicates

Whereas most methods have to be run on each dye-swap separately, we can generalize Chipmix for the simultaneous analysis of biological replicates.

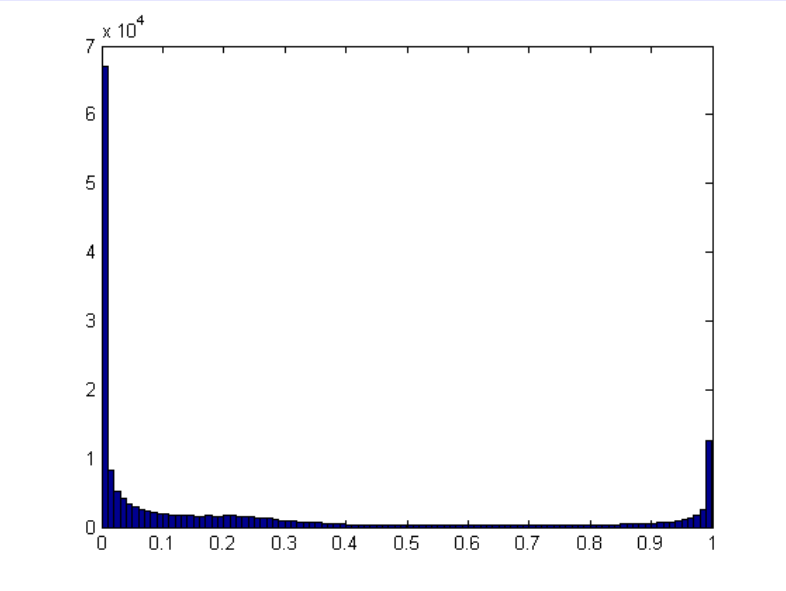
The model for the status of probe i and swap k is now :

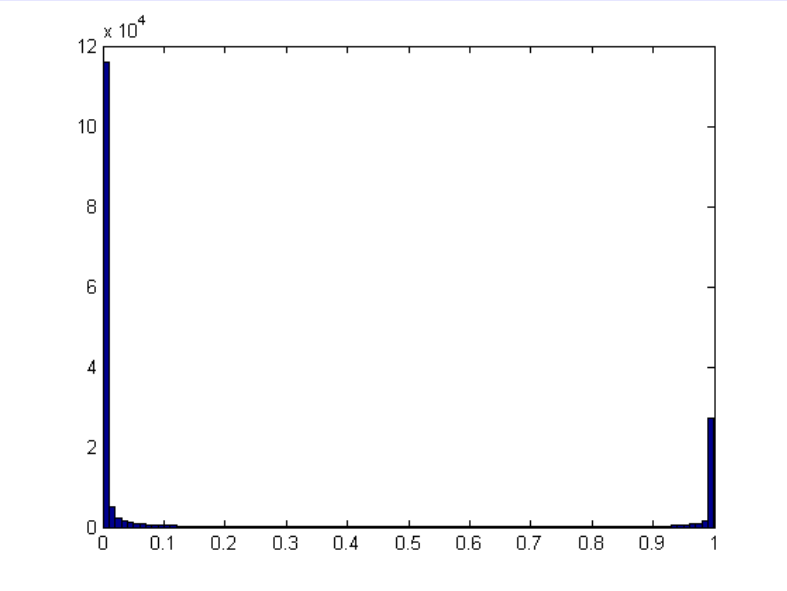
$$Y_{ik} = \begin{cases} a_0^k + b_0^k X_{ik} + E_{ik} & \text{if } i \text{ is normal,} \\ a_1^k + b_1^k X_{ik} + E_{ik} & \text{if } i \text{ is enriched.} \end{cases}$$

The posterior probability is given by :

$$\tau_i = \frac{\pi \phi_1(Y_{i1}, ..., Y_{iK} | x_{i1}, ..., x_{iK})}{(1-\pi)\phi_0(Y_{i1}, ..., Y_{iK} | x_{i1}, ..., x_{iK}) + \pi \phi_1(Y_{i1}, ..., Y_{iK} | x_{i1}, ..., x_{iK})}$$

Comparaison between single and multi-replicate analyses

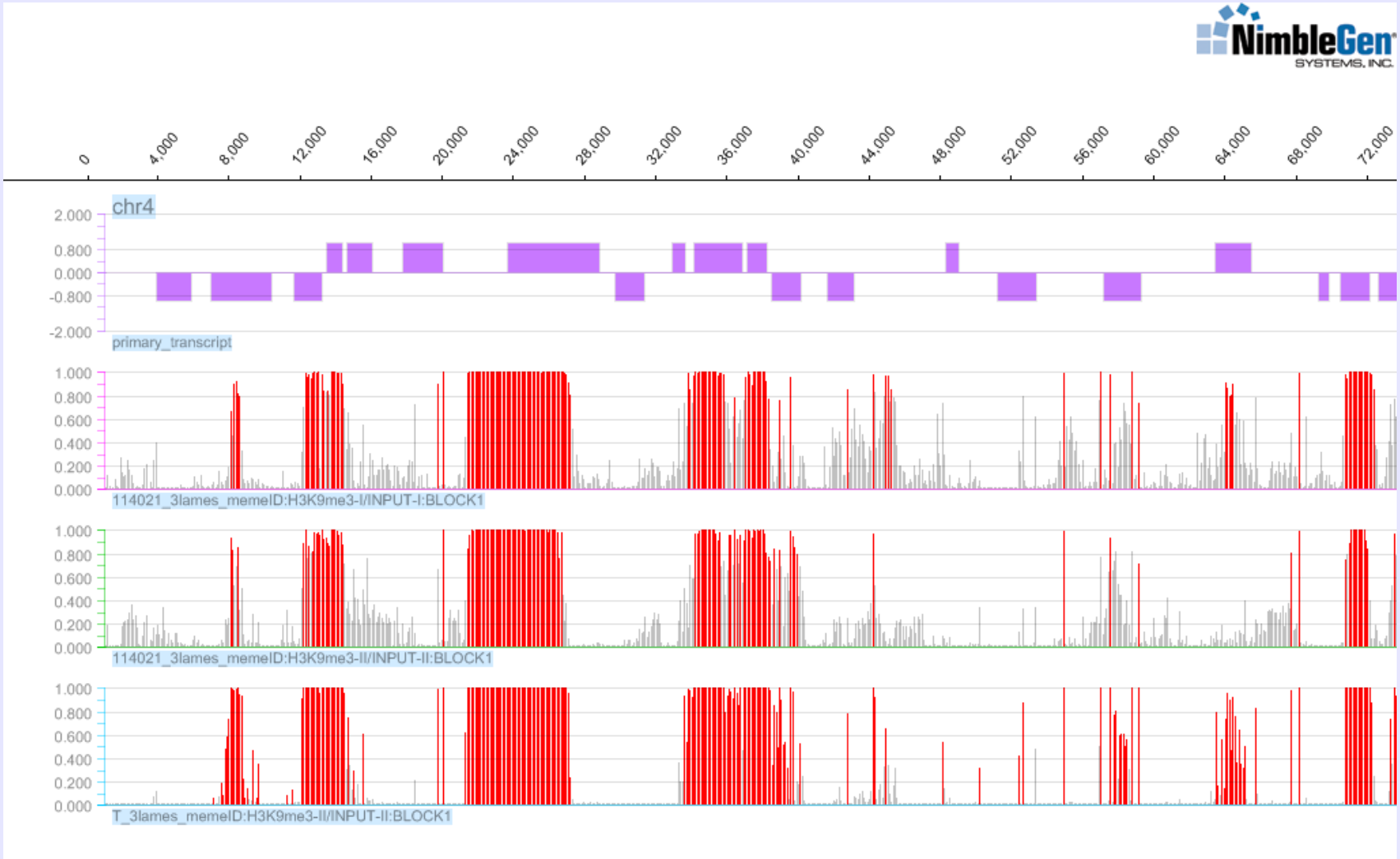




Left : Histogram of the posterior probabilities τ_i, computed on one biological replicate.

Right : Histogram of the posterior probabilities τ_i, computed with the two biological replicates.

Comparaison between single and multi-replicate analyses



Conclusions

Applications to ChIP-chip data from different arrays and different organisms are promising : our results are coherent with results already published and new candidates are found.

Future extensions :

⋆ Other explicative variables in the regressions can be added.

⋆ Control of the FDR rather than the nominal error rate level.

References :

⋆ ChIPmix : Mixture model of regressions for two-color ChIP-chip analysis, *Bioinformatics* (2008) **24**, 16.

⋆ R program of ChIPmix is available at

`http ://www.agroparistech.fr/mia/outil_A.html`

Acknowledgement

We thank François Roudier that provided the Arabidopsis data, and spent a lot of time to teach biology to ignorant statisticians.