Latent variable models for tiling array data

Applications to ChIP-chip and transcriptome experiments

Caroline Bérard

Advisors: Marie-Laure Martin-Magniette and Stéphane Robin

INRA MIA, DGAP, MICA

Doctoral school ABIES

UMR AgroParisTech/INRA MIA 518, Statistics and genome team, Paris.

Biological advances

- 1953: discovery of the double helix structure of DNA
- 1970: boom of molecular biology to understand the cell mechanisms
- 1972: first sequencing of a genome
  - structural annotation: prediction of genes structure and position
  - functional annotation: prediction of genes function
- 1960-Today: Evolution of high-speed technologies
  $\Rightarrow$ microarrays (1995), tiling arrays (2003), NGS (2008)

▶ Genome-wide study

Context of the thesis

▶ ANR Genoplante TAG Project

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

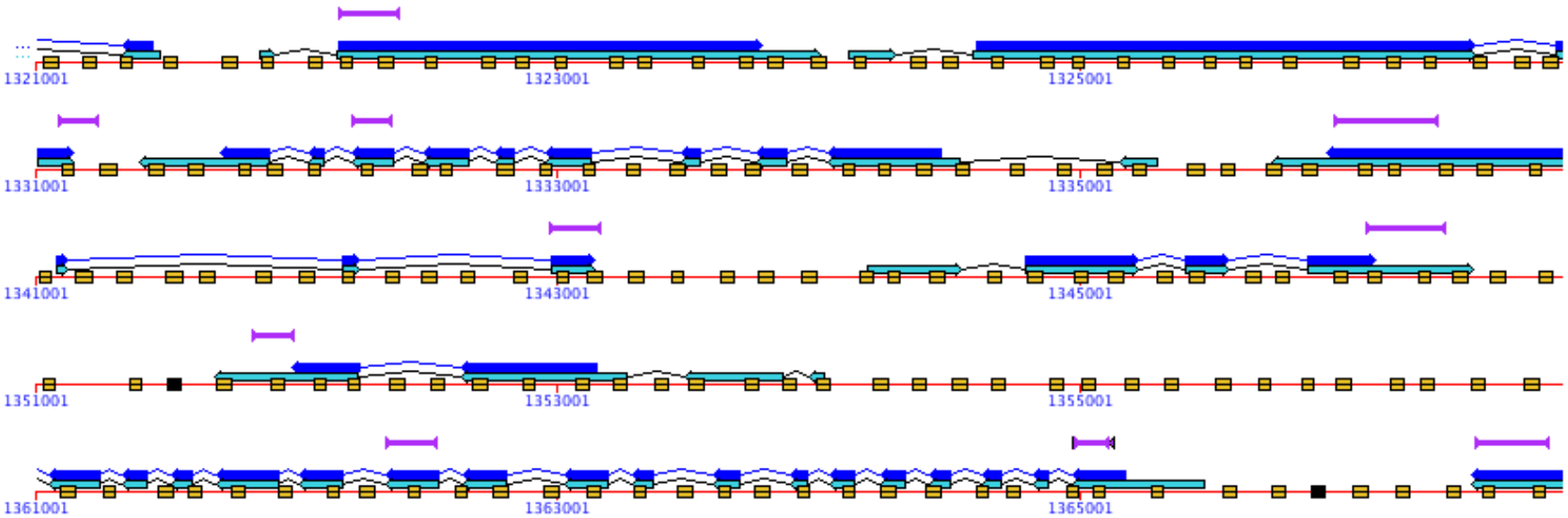
- Different types of application
  - **Transcriptome**: detection of transcripts, conditions of gene expression
  - **ChIP-chip**: study of control mechanism of gene expression (DNA methylation, histone modifications, transcription factor)

  $\longrightarrow$ Development of adapted statistical methods

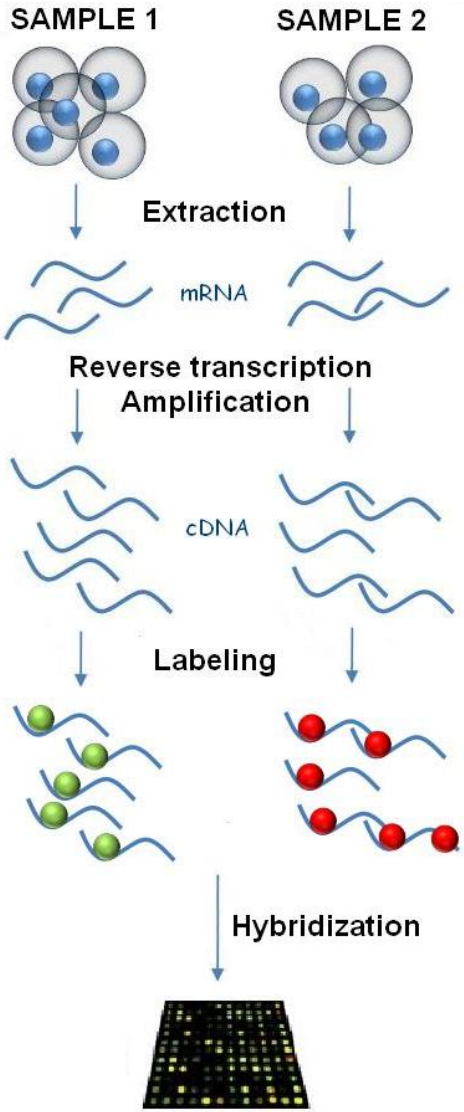
- Visualization of probe features and integration of the statistical results in the FLAGdb++ environment.

Tiling array features

- Probes are regularly distributed along the whole genome
- $\simeq$ 700 000 probes per array, $\simeq$ 100 000 probes per chromosome
- Resolution of 160 bp
- Distribution of annotation types: 67% intergenic, 14% exonic, 4% intronic



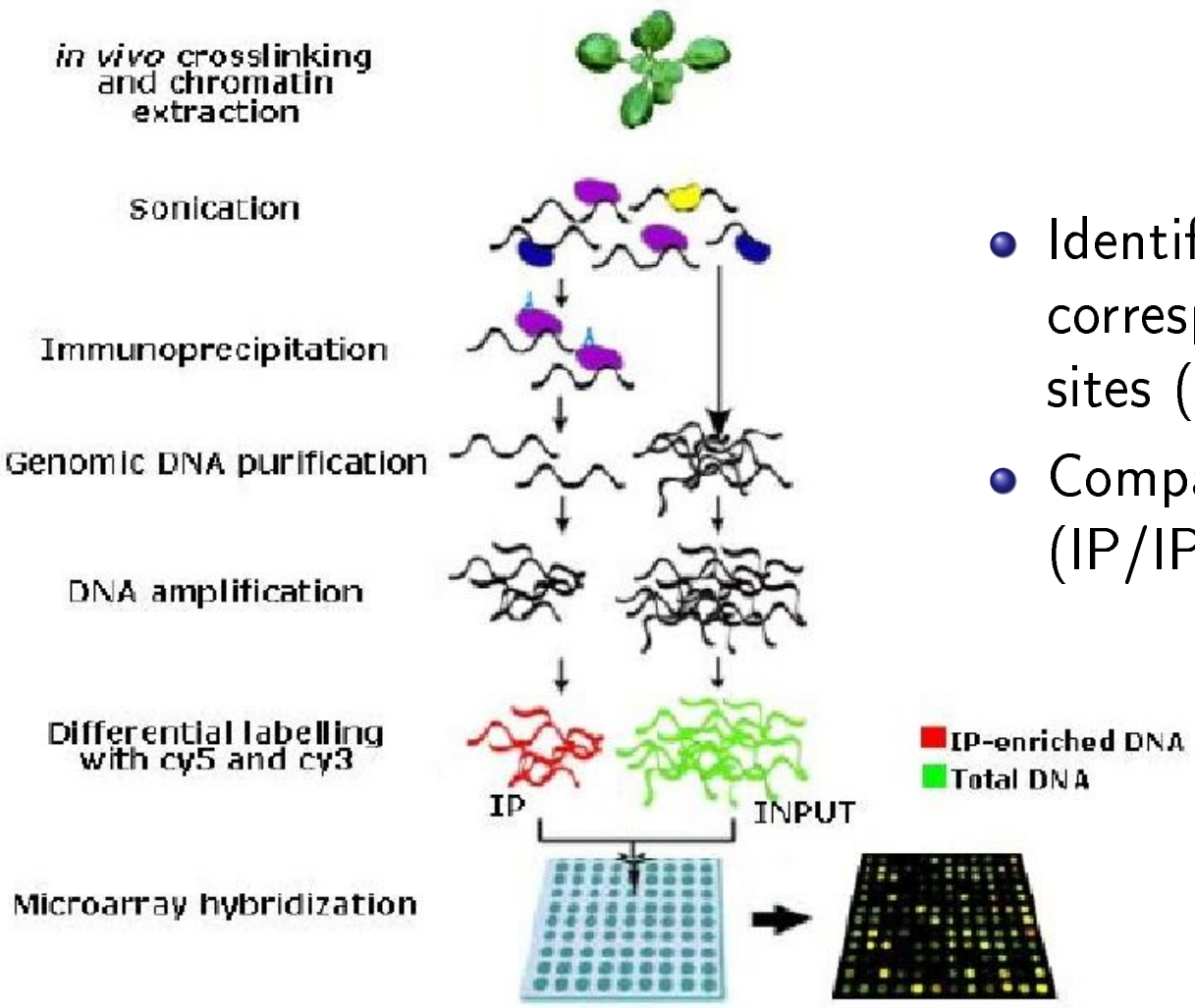
Transcriptome experiments



Objectives

- Detection of transcripts
- Gene expression profiles

ChIP-chip experiments



Objectives

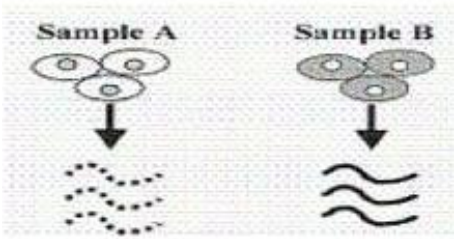
- Identification of DNA sequences corresponding to protein binding sites (IP/INPUT)
- Comparison of the two conditions (IP/IP)

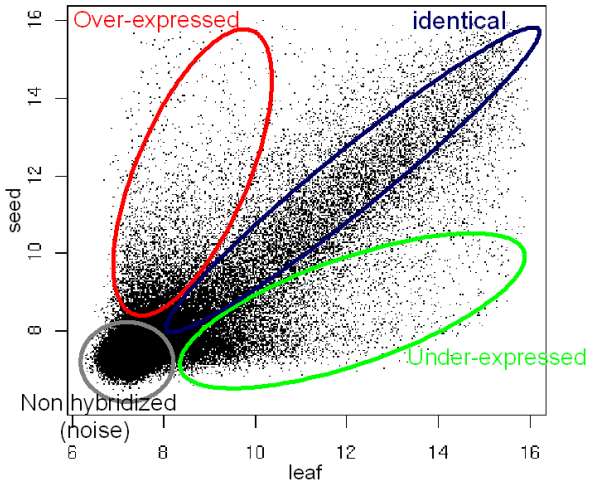
Previous work

- Transcriptome
  - Transcribed regions detection
    - ★ Segmentation methods *(Huber et al., 2006; Zeller et al., 2008)*
    - ★ Statistical tests *(Halasz et al., 2006)*
    - ★ Hidden Markov Models (Nicolas *et al.*, 2009)
  - Expression difference analysis (few methods)
    - ★ Statistical test on the log-ratio for each probe (Ji and Wong, 2005) or for given region (Ghosh *et al.*, 2007)
- ChIP-chip
  - IP/INPUT
    - ★ Several methods based on the logratio (Buck, Nobel and Lieb, 2005; Johnson *et al.*, 2006; Humburg *et al.*, 2008)
  - IP/IP
    - ★ Mixture models (Johannes *et al.*, 2010)

Latent variable models: Comparison of 2 conditions

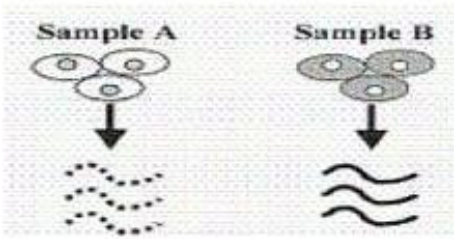
- **Transcriptome or ChIP-chip IP/IP:** 4 groups

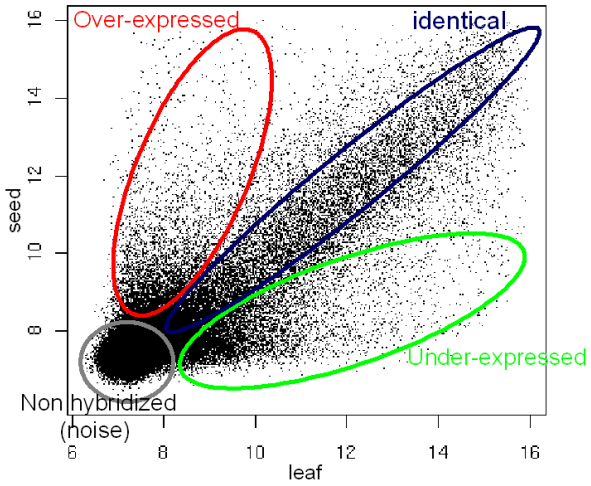




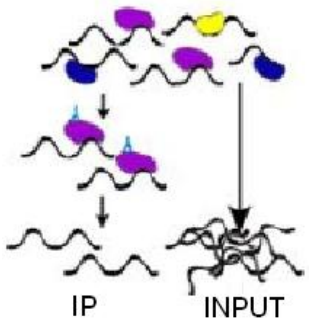
Latent variable models: Comparison of 2 conditions

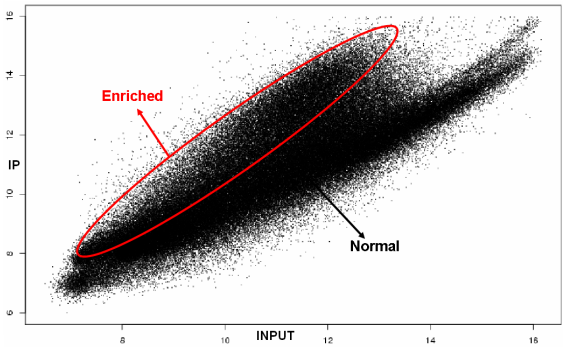
- **Transcriptome or ChIP-chip IP/IP:** 4 groups





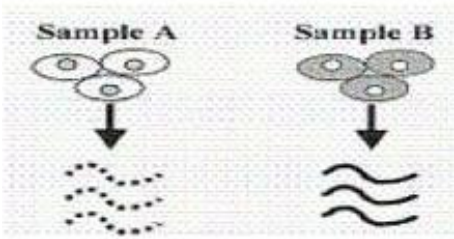
- **ChIP-chip IP/INPUT:** 2 groups (enriched, normal)

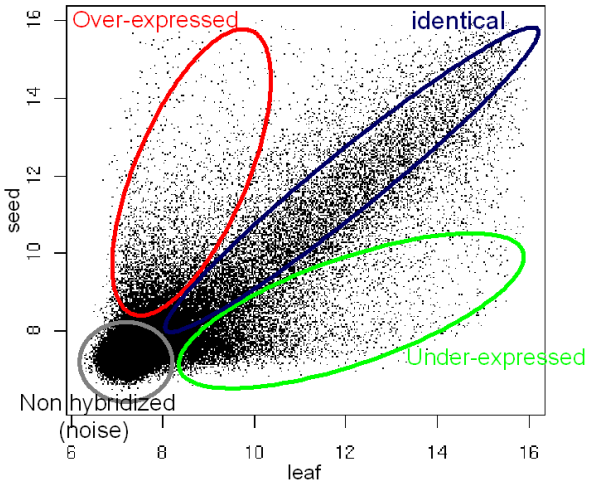




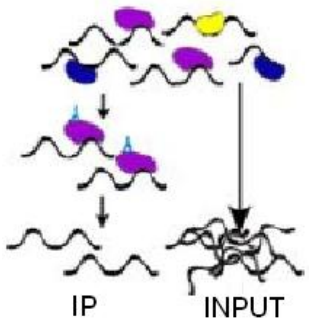
Latent variable models: Comparison of 2 conditions

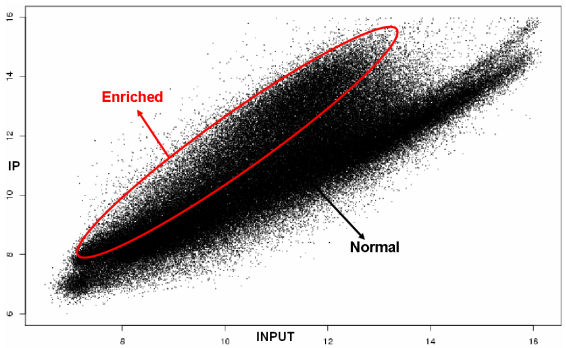
- **Transcriptome or ChIP-chip IP/IP:** 4 groups





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





Unsupervised classification problem $\longrightarrow$ Find the status of each probe

Contents

- Modeling of the latent variable distribution
  - Integration of dependence and annotation knowledge
- Modeling of the emission distribution: joint distribution of 2 samples
  - **Mixture of regressions**
    - $\star$ $X_t = (IP, INPUT)$ Non symmetrical $\rightsquigarrow$ ChIP-chip
  - **Bidimensional Gaussian mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
  - **Mixture of mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
- Inference
- Classification by probe and by region
- Applications

Contents

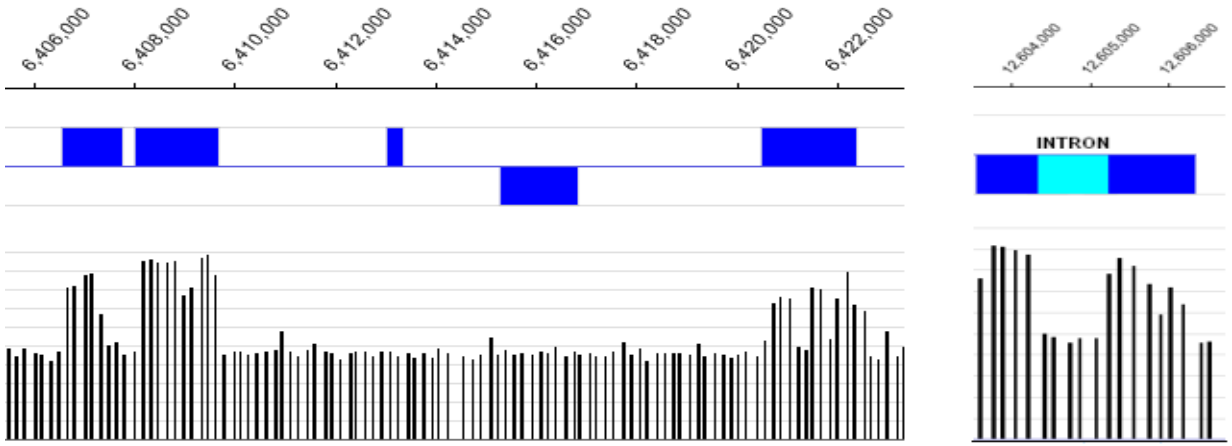
- Modeling of the latent variable distribution
  - Integration of dependence and annotation knowledge
- Modeling of the emission distribution: joint distribution of 2 samples
  - **Mixture of regressions**
    - $\star$ $X_t = (IP, INPUT)$ Non symmetrical $\rightsquigarrow$ ChIP-chip
  - **Bidimensional Gaussian mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
  - **Mixture of mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
- Inference
- Classification by probe and by region
- Applications

Contents

- Modeling of the latent variable distribution
  - Integration of dependence and annotation knowledge
- Modeling of the emission distribution: joint distribution of 2 samples
  - **Mixture of regressions**
    - $\star$ $X_t = (IP, INPUT)$ Non symmetrical $\rightsquigarrow$ ChIP-chip
  - **Bidimensional Gaussian mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
  - **Mixture of mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
- Inference
- Classification by probe and by region
- Applications

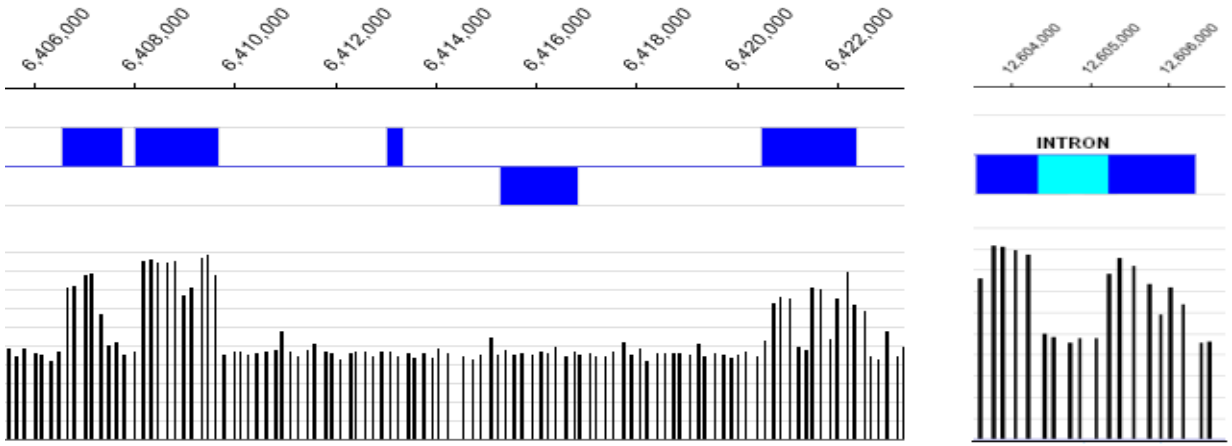
Available information

- Visualization of the signal intensity



Available information

- Visualization of the signal intensity



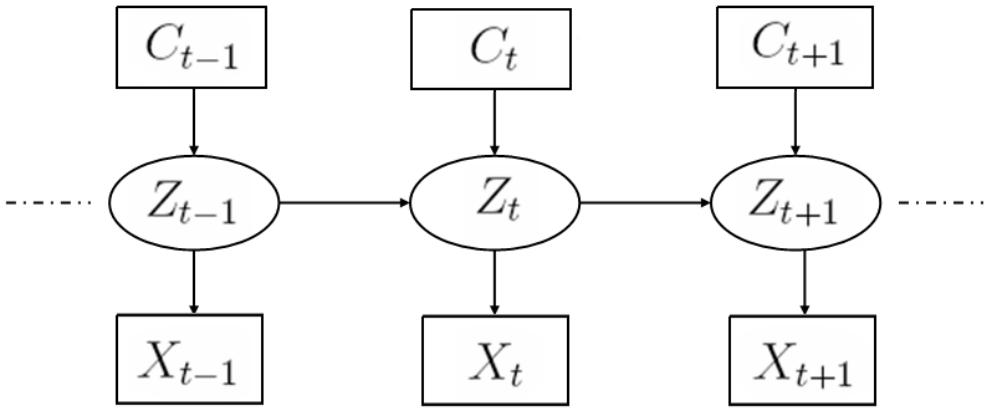
- Position of the probes along the genome $\rightsquigarrow t$

  $\longrightarrow$ Dependence between neighboring probes

- Structural annotation $\rightsquigarrow C_t$

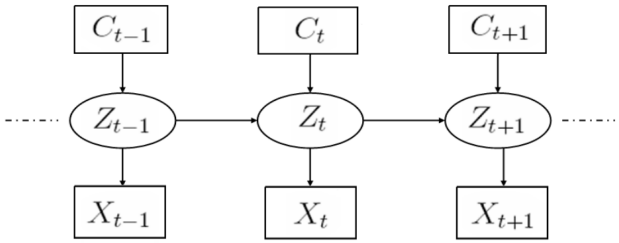
Model with HMM and Annotation

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



$\Rightarrow$ Inference: Forward/Backward algorithm for heterogeneous Markov chain

Four models



Mixture model

- Status $Z_t \sim \mathcal{M}(p)$
- Data X_t

HMM

- Status $Z_t \sim CM(\pi)$
- Data X_t

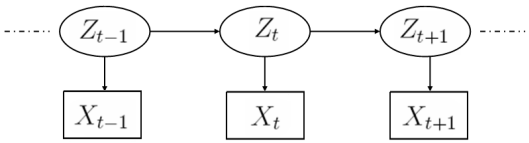
Mixture model + Annotation

- Annotation C_t
- Status $Z_t \sim \mathcal{M}(p^a)$
- Data X_t

HMM + Annotation

- Annotation C_t
- Status $Z_t \sim CM(\pi^a)$
- Data X_t

Four models



Mixture model

- Status $Z_t \sim \mathcal{M}(p)$
- Data X_t

HMM

- Status $Z_t \sim CM(\pi)$
- Data X_t

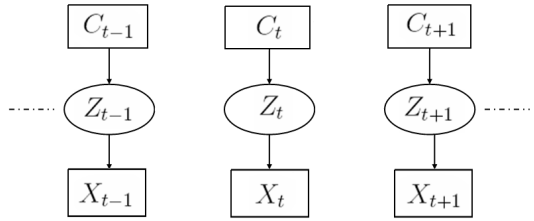
Mixture model + Annotation

- Annotation C_t
- Status $Z_t \sim \mathcal{M}(p^a)$
- Data X_t

HMM + Annotation

- Annotation C_t
- Status $Z_t \sim CM(\pi^a)$
- Data X_t

Four models



Mixture model

- Status $Z_t \sim \mathcal{M}(p)$
- Data X_t

HMM

- Status $Z_t \sim CM(\pi)$
- Data X_t

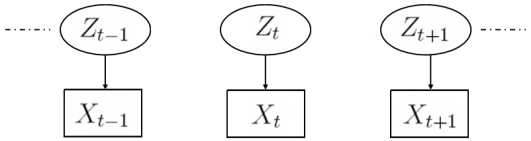
Mixture model + Annotation

- Annotation C_t
- Status $Z_t \sim \mathcal{M}(p^a)$
- Data X_t

HMM + Annotation

- Annotation C_t
- Status $Z_t \sim CM(\pi^a)$
- Data X_t

Four models



Mixture model

- Status $Z_t \sim \mathcal{M}(p)$
- Data X_t

HMM

- Status $Z_t \sim CM(\pi)$
- Data X_t

Mixture model + Annotation

- Annotation C_t
- Status $Z_t \sim \mathcal{M}(p^a)$
- Data X_t

HMM + Annotation

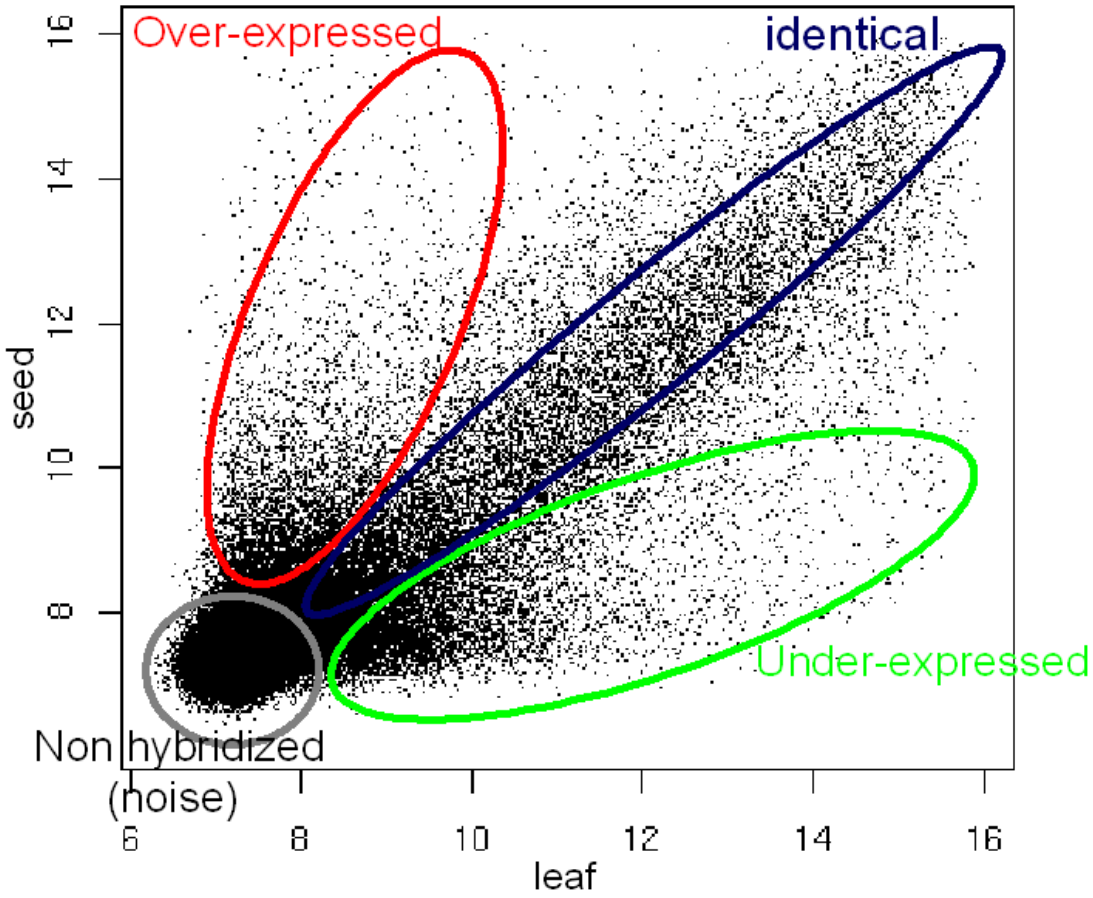
- Annotation C_t
- Status $Z_t \sim CM(\pi^a)$
- Data X_t

Contents

- Modeling of the latent variable distribution
  - Integration of dependence and annotation knowledge
- Modeling of the emission distribution: joint distribution of 2 samples
  - **Mixture of regressions**
    - $\star$ $X_t = (IP, INPUT)$ Non symmetrical $\rightsquigarrow$ ChIP-chip
  - **Bidimensional Gaussian mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
  - **Mixture of mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
- Inference
- Classification by probe and by region
- Applications

Bidimensional Gaussian mixture

- Data $X_t = (X_{1t}, X_{2t})$
- $K = 4$ biologically interpretable groups
- $(X_t | Z_t = k) \sim \mathcal{N}(\mu_k, \Sigma_k)\ \forall k = 1, ..., K$



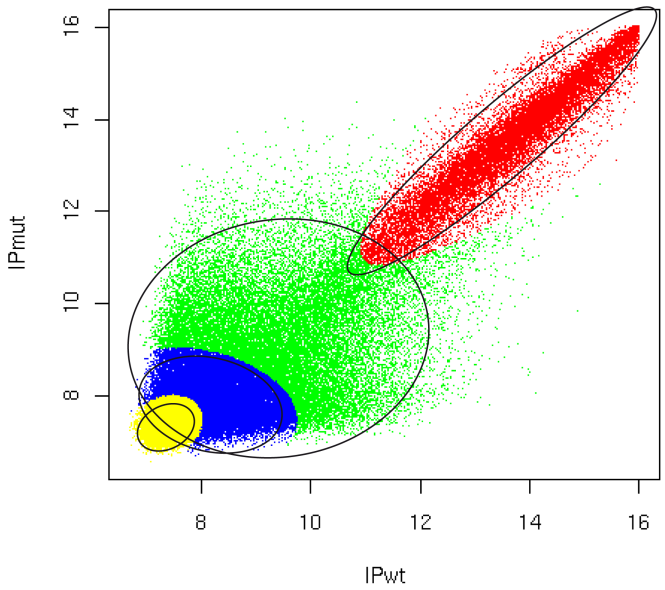
Eigenvalue decomposition of Σ_k (Banfield & Raftery, 1993)

$$\boxed{\Sigma_k = \lambda_k D_k A_k D_k'}$$

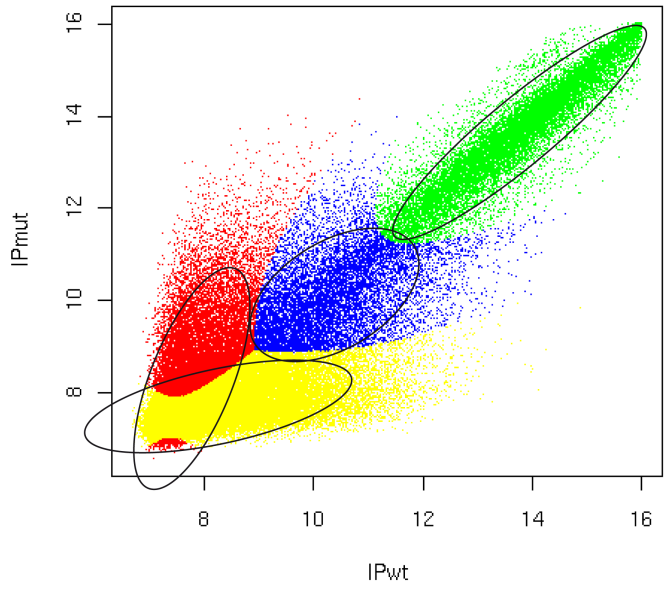
- $\lambda_k = \det(\Sigma_k)^{1/2} \rightsquigarrow$ volume
- $D_k =$ matrix of eigen vectors of $\Sigma_k \rightsquigarrow$ orientation
- $A_k =$ matrix of normalised eigen values of $\Sigma_k \rightsquigarrow$ shape

$\Rightarrow$ 14 easily interpretable models (Celeux & Govaert, 1995)

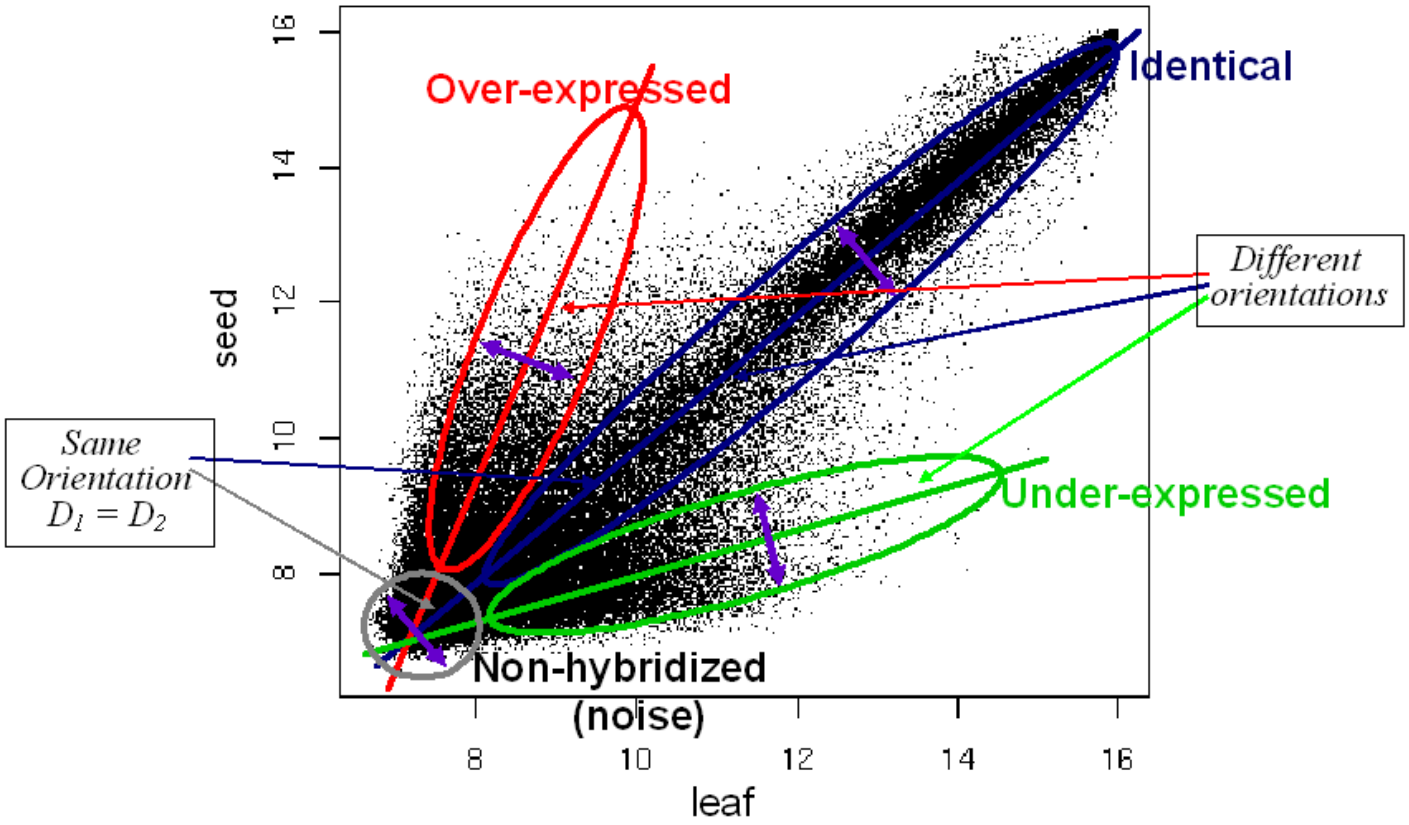
Model $\lambda_k D_k A_k D_k'$



Model $\lambda D_k A_k D_k'$



Specific modeling of the variance matrix



- 2 groups have the same orientation
- Same noise in each group $\Leftrightarrow$ fixed $2nd$ eigen value of Σ_k

Specific modeling of the variance matrix

- Constraints:

$$\begin{cases} \Sigma_k &= \lambda_k D_k A_k D_k' = D_k \Lambda_k D_k', \text{ for } k = 1, .., 4, \text{ with } \Lambda_k = \lambda_k A_k \\ D_1 &= D_2 = D \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \text{ with } u_{1k} > u_2, \text{ for } k = 1, .., 4. \end{cases}$$

- Estimates of D, D_k, Λ_k using the EM algorithm

- *TAHMMAnnot* package freely available from CRAN

Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipité. C. Bérard, M-L. Martin-Magniette, A. To, F. Roudier, V. Colot and S. Robin. *La revue de MODULAD* (2009)

Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome. C. Bérard, M-L. Martin-Magniette, V. Brunaud, S. Aubourg and S. Robin. *SAGMB* (2011)

Application on *Arabidopsis thaliana* transcriptomic dataset: Seed VS Leaf

- Comparison of the 4 models, with 3 annotation categories

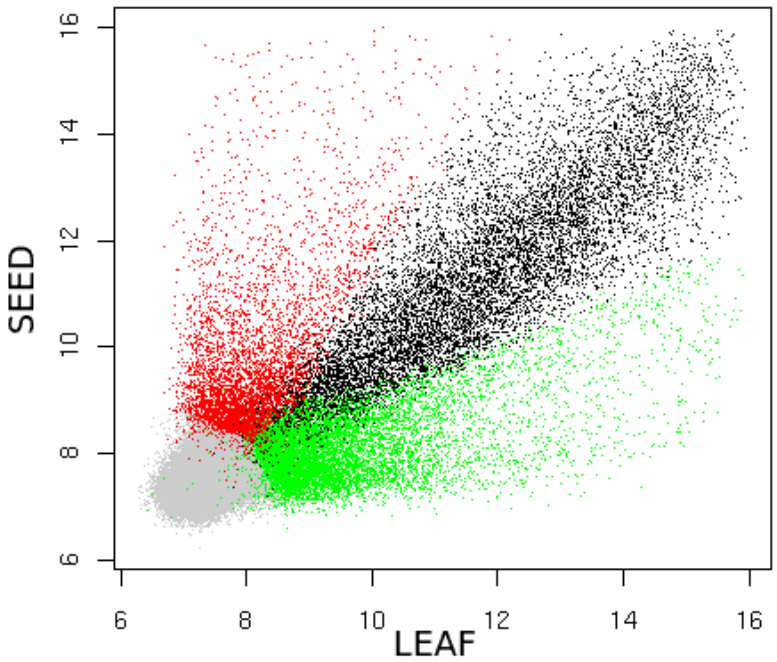
	Mixture	HMM	Mixture+Annot	HMM+Annot
#Param.	19	31	25	61
BIC	406469 (+ 49146)	371668 (+ 14345)	373573 (+ 16250)	**357323**
ICL	436197 (+ 37925)	412706 (+ 14434)	399986 (+ 1714)	**398272**

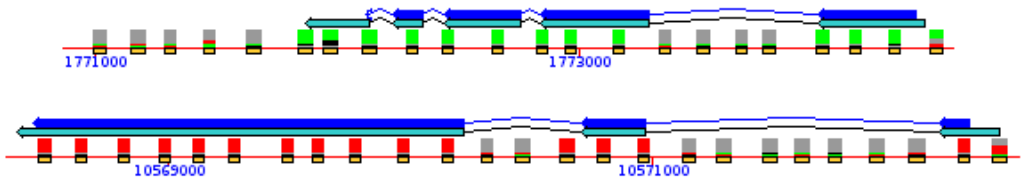
Application on *Arabidopsis thaliana* transcriptomic dataset: Seed VS Leaf

- Comparison of the 4 models, with 3 annotation categories

	Mixture	HMM	Mixture+Annot	HMM+Annot
#Param.	19	31	25	61
BIC	406469 (+ 49146)	371668 (+ 14345)	373573 (+ 16250)	**357323**
ICL	436197 (+ 37925)	412706 (+ 14434)	399986 (+ 1714)	**398272**

- Probe classification and visualization in Flagdb++





Estimation of the transition matrices and the proportions

Transition matrix of **intergenic** category:

(in %)	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 **intronic** category:

(in %)	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

Transition matrix of **exonic** category:

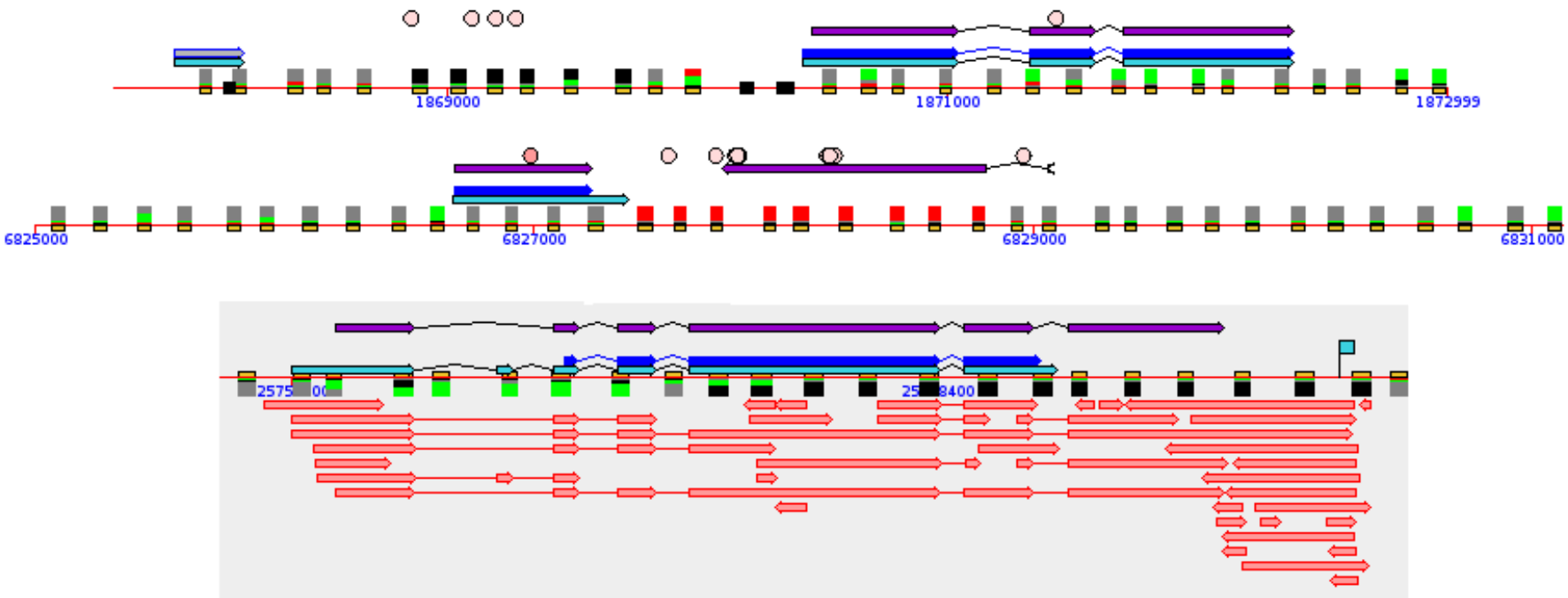
(in %)	Noise	Ident.	Under-exp	Over-exp
Noise	83	14	3	0
Ident.	2	90	6	2
Under-exp	7	5	87	1
Over-exp	8	6	1	85

Proportions (%)
22
41
23
14

Detection of new transcripts

- 143 expressed regions of more than 850 *bp* found in intergenic in 2 biological replicates
- 82 validated by TAIR10: *other*RNA, *sn*RNA, *sno*RNA, *r*RNA, *t*RNA
- Analysis of the 61 other regions: EST, MPSS mRNA, gene Eugene

  $\longrightarrow$ 47 with at least an indication of transcription and 14 with nothing

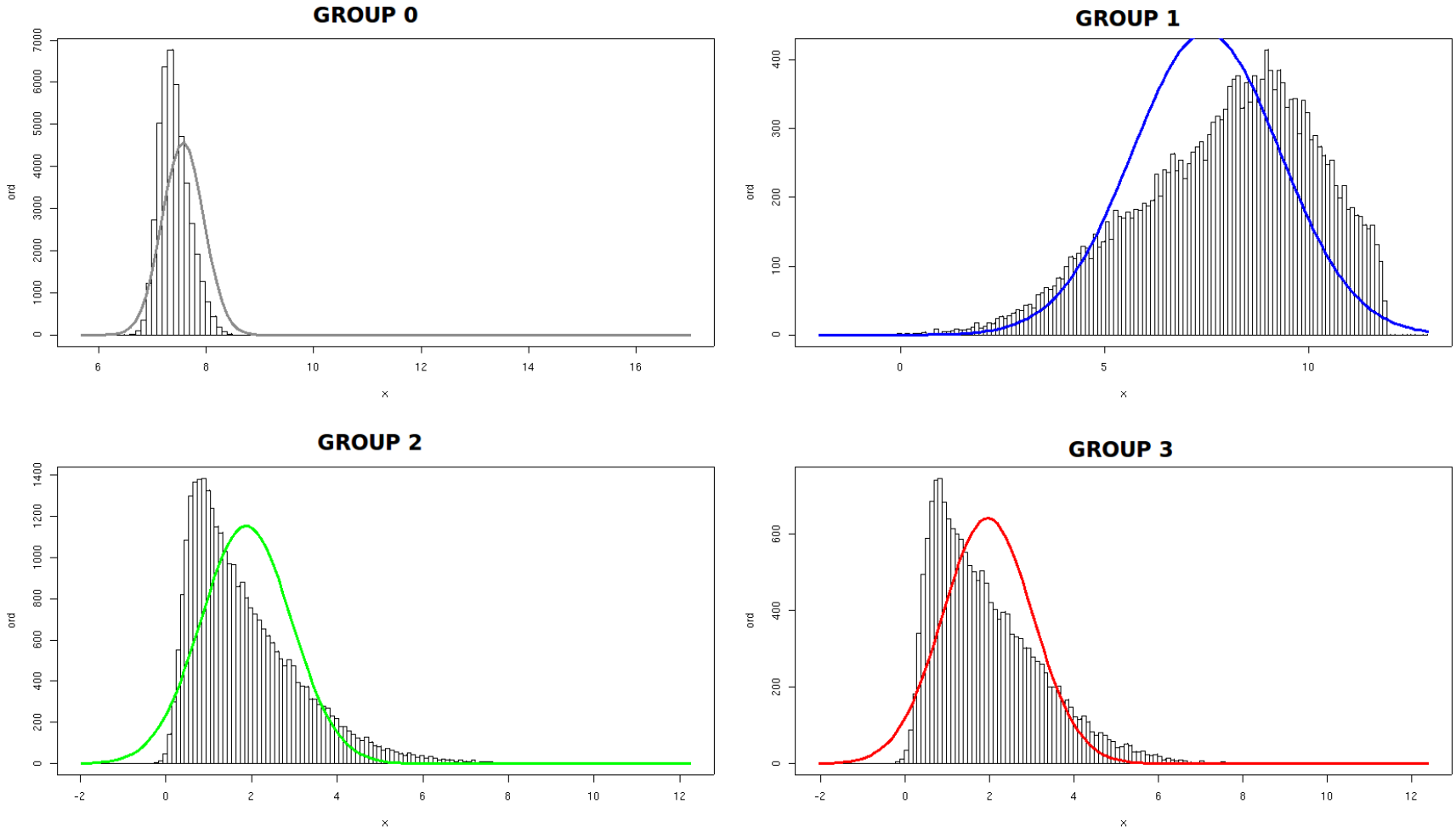


Contents

- Modeling of the latent variable distribution
  - Integration of dependence and annotation knowledge
- Modeling of the emission distribution: joint distribution of 2 samples
  - **Mixture of regressions**
    - $\star$ $X_t = (IP, INPUT)$ Non symmetrical $\rightsquigarrow$ ChIP-chip
  - **Bidimensional Gaussian mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
  - **Mixture of mixture**
    - $\star$ $X_t = (X_{t1}, X_{t2})$ Symmetrical $\rightsquigarrow$ Transcriptome or IP/IP
- Inference
- Classification by probe and by region
- Applications

Mixture of Mixture

Histograms of weigthed data projected on the main axis of each group



Model

- $X_t = (X_{1t}, X_{2t})$
- $\{Z_t\}$ is a K-state homogeneous Markov chain (Π, m)
- The observations $\{X_t\}$ are independent conditionally to Z
- Conditional distribution:

$$(X_t | Z_t = k) \sim \phi_k \quad \text{and} \quad \phi_k = \sum_{\ell=1}^{L_k} \eta_{k\ell} f(.; \theta_{k\ell})$$

  - $\eta_{k\ell}$ is the mixing proportion of the ℓ-th component for the group k
  - L_k is the number of components within the group k and $\sum_\ell \eta_{k\ell} = 1$
  - L is the total number of components of the model

Vector of model parameters: $\Theta = (\Pi, m, \{\eta_{k\ell}\}_{k,\ell}, \{\theta_{k\ell}\}_{k,\ell})$

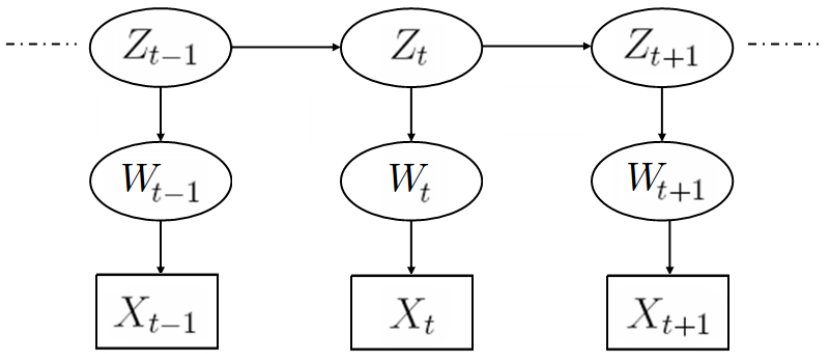
Another view of the model

- $\{Z_t\}$ is a Markov chain taking its values in $\{1, ..., K\} \rightarrow$ groups
- $\{W_t\}$ is a Markov chain taking its values in $\{1, ..., L\} \rightarrow$ components
- Z and W are two nested Markov chains

$$\forall\, t,\ (X_t | W_{tk} = \ell) \sim f(.; \theta_{k\ell})$$

The transition matrix of W, $\Omega = \{\omega_{k,\ell;k',\ell'}\}$ with $(k, k') \in \{1, ..., K\}^2$ and $(\ell, \ell') \in \{1, ..., L_k\}^2$ is of the form:

$$\omega_{k,\ell;k',\ell'} = \pi_{k,k'} \times \eta_{k'\ell'}$$



Inference

▶ EM Algorithm

- E step: Forward/Backward algorithm to estimate $P(Z|X;\Theta^h)$
- M step: maximizing $\mathbb{E}_{Z|X}\left[\log P(X,Z;\Theta)\right]$ in Θ

$$\mathbb{E}_{Z|X}\left[\log P(X,Z;\Theta)\right] = \underbrace{\mathbb{E}_{Z|X}\left[\log P(Z;\Theta)\right]}_{\Pi^{(h+1)},m^{(h+1)}} + \mathbb{E}_{Z|X}\left[\log P(X|Z;\Theta)\right]$$

$$\mathbb{E}_{Z|X}\left[\log P(X|Z;\Theta)\right] = \sum_{t=1}^{n}\sum_{k=1}^{K}\tau_{tk}\log \phi_k(X_t)$$

where $\tau_{tk} = P(Z_t = k|X;\Theta^h)$

Inference

▶ EM Algorithm

- E step: Forward/Backward algorithm to estimate $P(Z|X;\Theta^h)$
- M step: maximizing $\mathbb{E}_{Z|X}\left[\log P(X,Z;\Theta)\right]$ in Θ

$$\mathbb{E}_{Z|X}\left[\log P(X,Z;\Theta)\right] = \underbrace{\mathbb{E}_{Z|X}\left[\log P(Z;\Theta)\right]}_{\Pi^{(h+1)},m^{(h+1)}} + \mathbb{E}_{Z|X}\left[\log P(X|Z;\Theta)\right]$$

$$\mathbb{E}_{Z|X}\left[\log P(X|Z;\Theta)\right] = \sum_{t=1}^{n}\sum_{k=1}^{K}\tau_{tk}\log\left[\sum_{\ell=1}^{L_k}\eta_{k\ell}f(X_t;\theta_{k\ell})\right]$$

where $\tau_{tk} = P(Z_t = k|X;\Theta^h)$

Inference with two latent variables

$$\log P(X) \quad = \quad \mathbb{E}_{Z|X}\left[\log P(X,Z)\right] - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right]$$

Inference with two latent variables

$$
\begin{aligned}
\log P(X) &= \mathbb{E}_{Z|X}\left[\log P(X,Z)\right] - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z|X}\left\{\mathbb{E}_{W|Z,X}\left[\log P(X,Z,W)\right] - \mathbb{E}_{W|Z,X}\left[\log P(W|Z,X)\right]\right\} \\
&\quad - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right]
\end{aligned}
$$

Inference with two latent variables

$$
\begin{aligned}
\log P(X) &= \mathbb{E}_{Z|X}\left[\log P(X,Z)\right] - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z|X}\left\{\mathbb{E}_{W|Z,X}\left[\log P(X,Z,W)\right] - \mathbb{E}_{W|Z,X}\left[\log P(W|Z,X)\right]\right\} \\
&\quad - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z,W|X}\left[\log P(X,Z,W)\right] + \mathbb{E}_{Z|X}\mathcal{H}(W|Z,X) + \mathcal{H}(Z|X)
\end{aligned}
$$

Inference with two latent variables

$$
\begin{aligned}
\log P(X) &= \mathbb{E}_{Z|X}\left[\log P(X,Z)\right] - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z|X}\left\{\mathbb{E}_{W|Z,X}\left[\log P(X,Z,W)\right] - \mathbb{E}_{W|Z,X}\left[\log P(W|Z,X)\right]\right\} \\
&\quad - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z,W|X}\left[\log P(X,Z,W)\right] \underbrace{+\mathbb{E}_{Z|X}\mathcal{H}(W|Z,X) + \mathcal{H}(Z|X)}_{\mathcal{H}(W,Z|X)}
\end{aligned}
$$

Inference with two latent variables

$$
\begin{aligned}
\log P(X) &= \mathbb{E}_{Z|X}\left[\log P(X,Z)\right] - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z|X}\left\{\mathbb{E}_{W|Z,X}\left[\log P(X,Z,W)\right] - \mathbb{E}_{W|Z,X}\left[\log P(W|Z,X)\right]\right\} \\
&\quad - \mathbb{E}_{Z|X}\left[\log P(Z|X)\right] \\
&= \mathbb{E}_{Z,W|X}\left[\log P(X,Z,W)\right] \underbrace{+\mathbb{E}_{Z|X}\mathcal{H}(W|Z,X) + \mathcal{H}(Z|X)}_{\mathcal{H}(W,Z|X)}
\end{aligned}
$$

- Maximisation in Θ

$$\mathbb{E}_{Z|X}\left[\log P(Z;\Theta)\right] = \sum_k \tau_{k1}\log(m_k) + \sum_{t\geq 2}\sum_{k,k'}\mathbb{E}\left[Z_{t-1,k}Z_{t,k'}|X\right]\log(\pi_{k,k'})$$

$$\mathbb{E}_{W,Z|X}\left[\log P(W|Z;\Theta)\right] = \sum_t\sum_k \tau_{tk}\sum_\ell \delta_{tk\ell}\log\eta_{k\ell}$$

$$\mathbb{E}_{W,Z|X}\left[\log P(X|W,Z;\Theta)\right] = \sum_t\sum_k \tau_{tk}\sum_\ell \delta_{tk\ell}\log f(X_t;\theta_{k\ell})$$

with

$$\tau_{tk} = P(Z_t = k|X) \quad \text{and} \quad \delta_{tk\ell} = P[W_{tk} = \ell|X_t, Z_t = k]$$

Selection criteria to estimate the number of groups K or the number of components L

- Parametric emission distribution (generic latent variable S):

$$BIC(K) = \log P(X; \widehat{\Theta}_K) - \frac{\nu_K}{2} \log(n)$$

$$ICL(K) = \log P(X; \widehat{\Theta}_K) - \frac{\nu_K}{2} \log(n) - \mathcal{H}(S|X)$$

Selection criteria to estimate the number of groups K or the number of components L

- Parametric emission distribution (generic latent variable S):

$$BIC(K) = \log P(X; \widehat{\Theta}_K) - \frac{\nu_K}{2} \log(n)$$

$$ICL(K) = \log P(X; \widehat{\Theta}_K) - \frac{\nu_K}{2} \log(n) - \mathcal{H}(S|X)$$

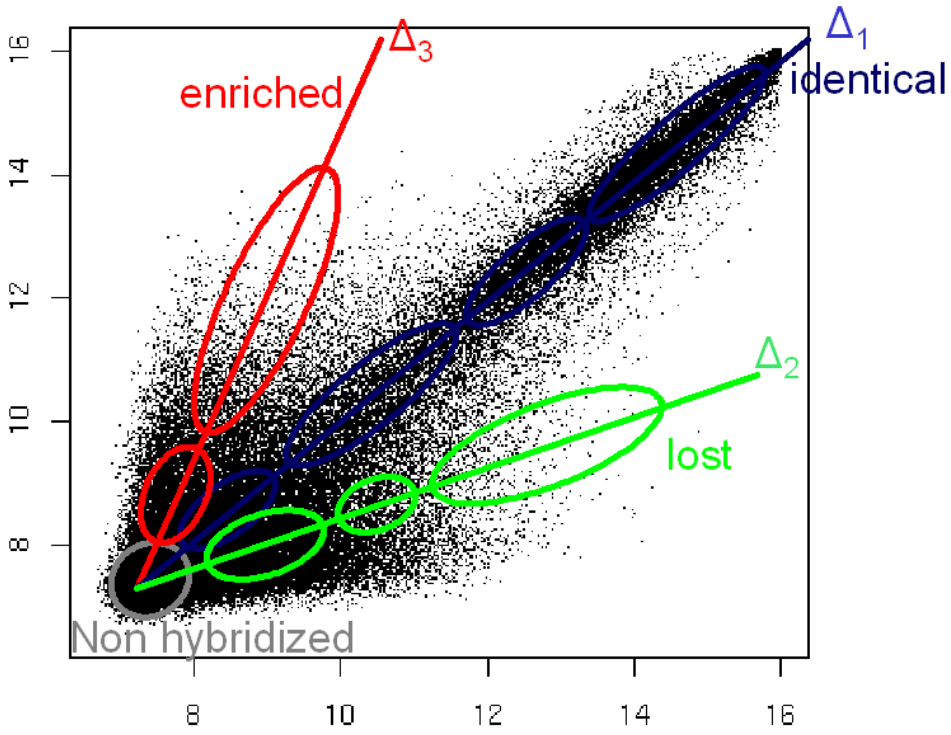
- Mixture as emission distribution:

$$BIC(K, L) = \log P(X; \widehat{\Theta}_{K,L}) - \frac{\nu_{K,L}}{2} \log(n)$$

$$ICL_W(K, L) = \log P(X; \widehat{\Theta}_{K,L}) - \frac{\nu_{K,L}}{2} \log(n) - \mathcal{H}(W, Z|X)$$

$$ICL_Z(K, L) = \log P(X; \widehat{\Theta}_{K,L}) - \frac{\nu_{K,L}}{2} \log(n) - \mathcal{H}(Z|X)$$

MODEL 1: Model with colinearity constraints



- Δ_k concurrent at the barycentre of the group 0
- The Gaussian components of the k-th cluster are forced to be colinear along Δ_k

- Group 0: spherical Gaussian

$$(X_t|Z_t = 0) \sim \mathcal{N}\left(\begin{pmatrix}\mu_0^1\\ \mu_0^2\end{pmatrix}, \sigma^2 I_2\right)$$

- The other groups are modeled by a Gaussian mixture:

  (U_{tk}, V_{tk}) = coordinates of (X_{1t}, X_{2t}) in the orthonormal basis $(\Delta_k, \Delta_k^\perp)$

$\rightsquigarrow$ Unidimensional Mixture

  - $(V_{tk}|Z_t = k) \sim \mathcal{N}(0, \sigma_k^2)$
  - $(U_{tk}|Z_t = k) \sim \psi_k$

where $\psi_k = \sum_{\ell=1}^{L_k} \eta_{k\ell} \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$

Inference algorithm

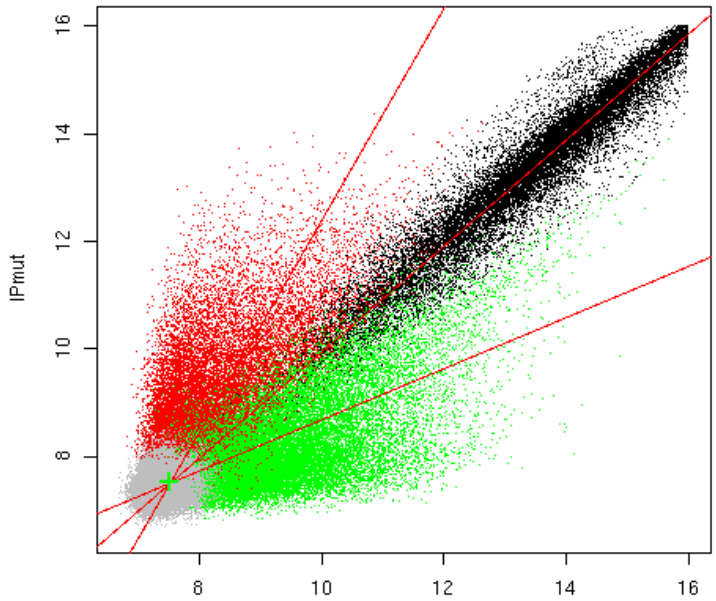
- Number of groups $K = 4$
- Initialisation of the EM algorithm: using the results of the model with a single Gaussian per group
- EM inference with Z and W
- Number of components in each group: BIC, ICL_Z

$\rightarrow$ Assumption: L_k is constant $\forall k$

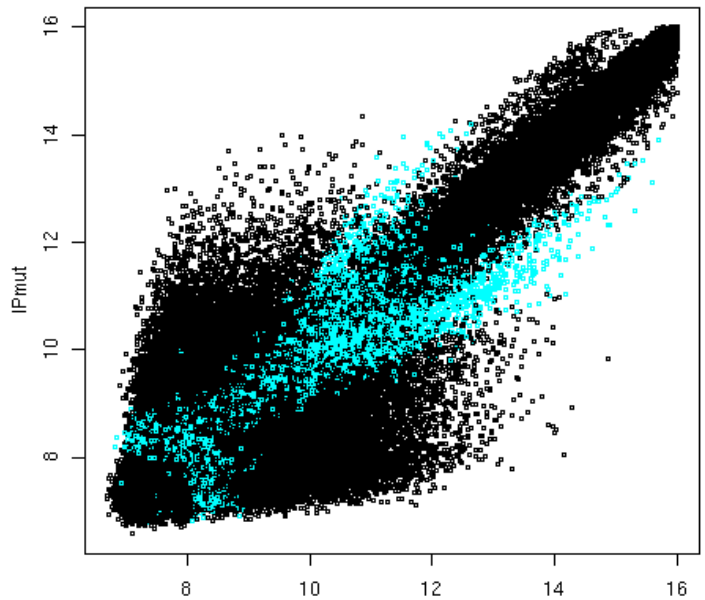
Application on Arabidopsis thaliana ChIP-chip IP/IP dataset: Wt VS Mutant

$\rightsquigarrow$ Results with $\sigma_0^2 = \sigma_1^2$ and $\sigma_2^2 = \sigma_3^2$

nbcomp	1	2	3	4	5	6	7
nbparam	31	70	124	208	304	418	550
BIC	485367	455625	453683	453104	452967	**452904**	452950
ICL_Z	516275	486285	482988	481662	481257	**481028**	481041

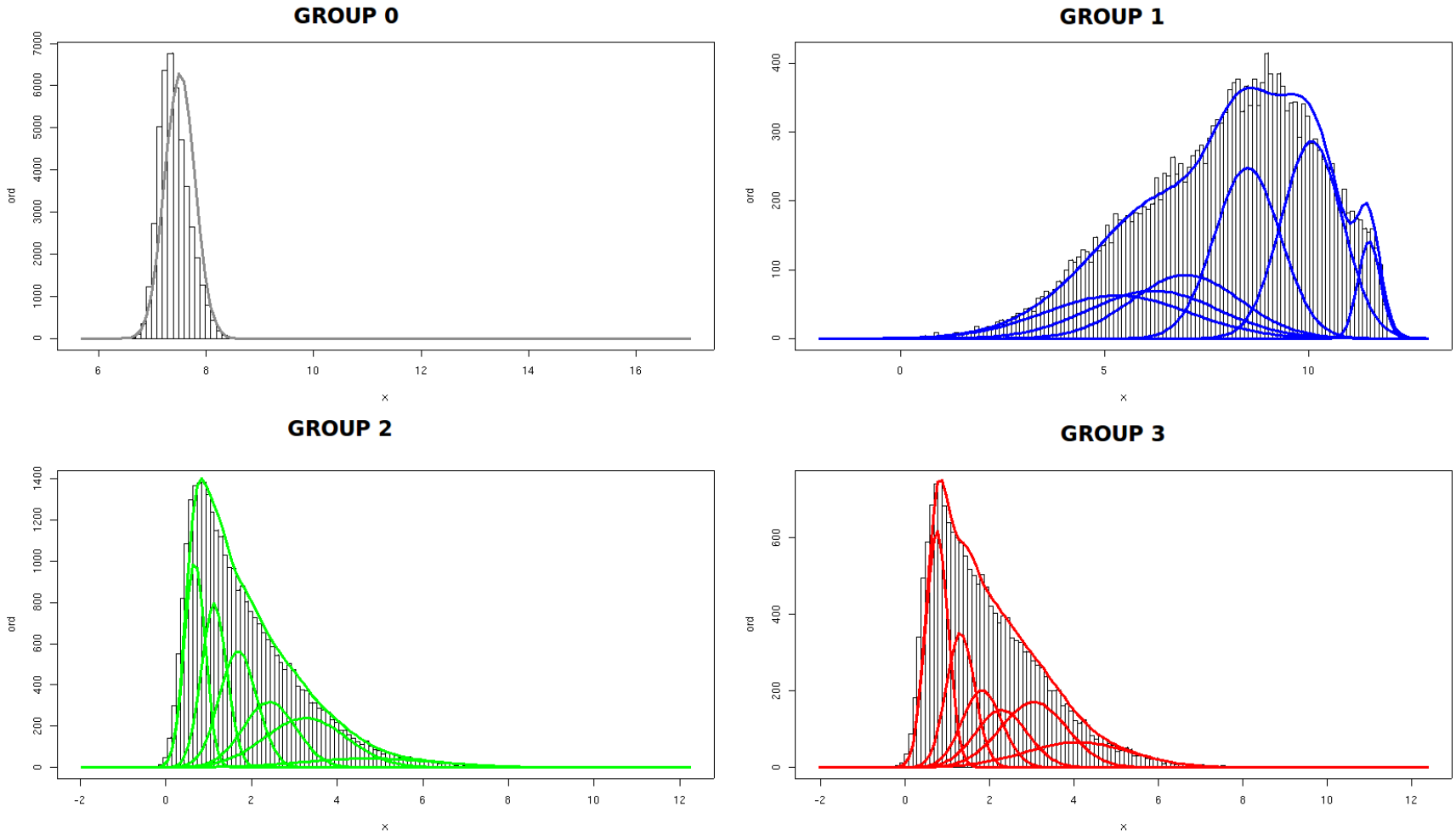


Probe classification

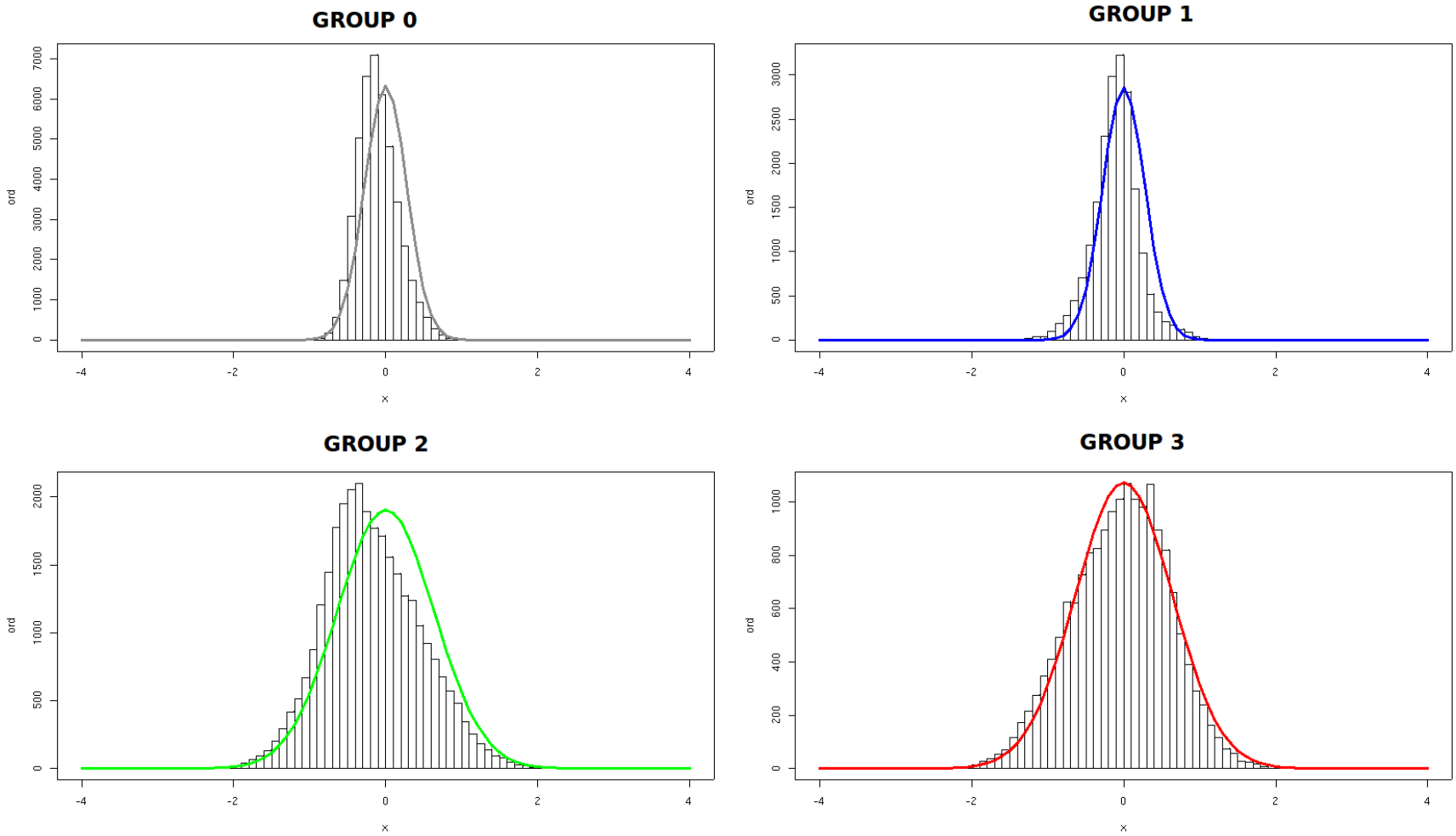


Comparison with single Gaussian

Fit of the estimated densities for U_k

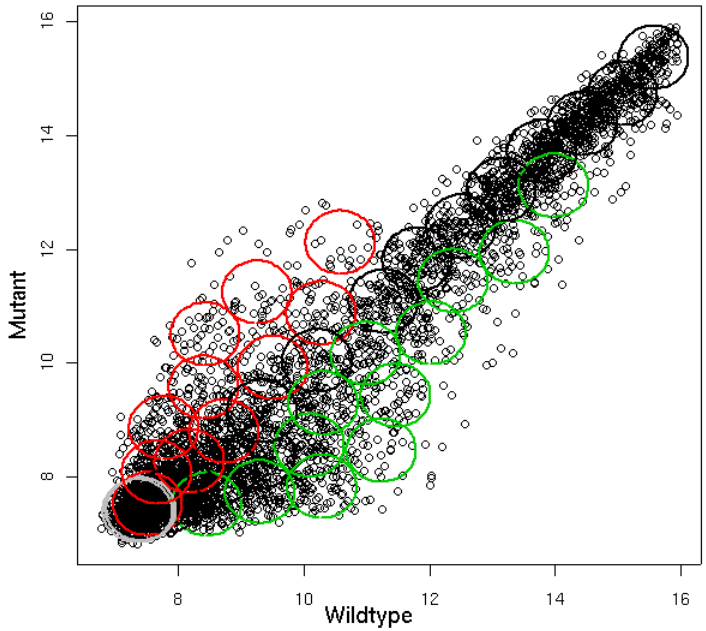


Fit of the estimated densities for V_k



MODEL 2: A more general model

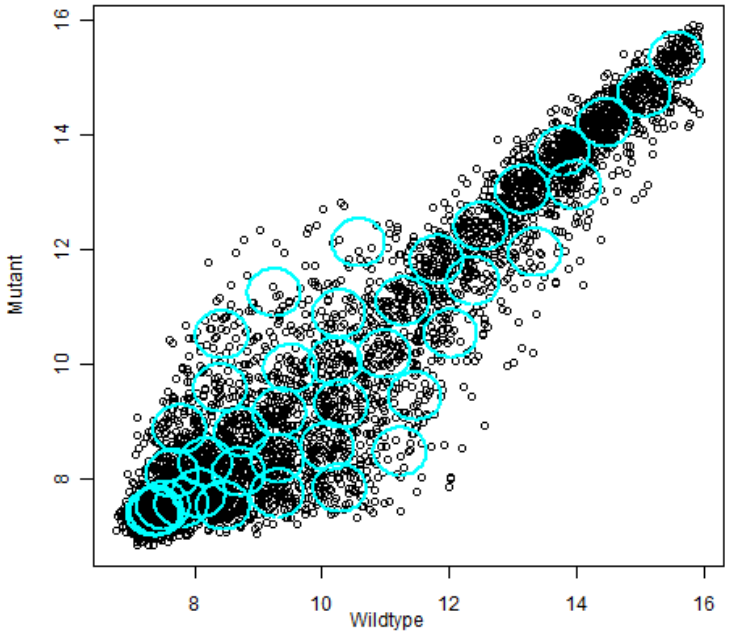
- $(X_t|Z_t = k) \sim \phi_k$, $\qquad \phi_k = \sum_{\ell=1}^{L_k} \eta_{k\ell} f(.; \theta_{k\ell})$
- f *p.d.f* of $\mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \sigma^2 I_2\right)$, no constraints on μ_1 and μ_2

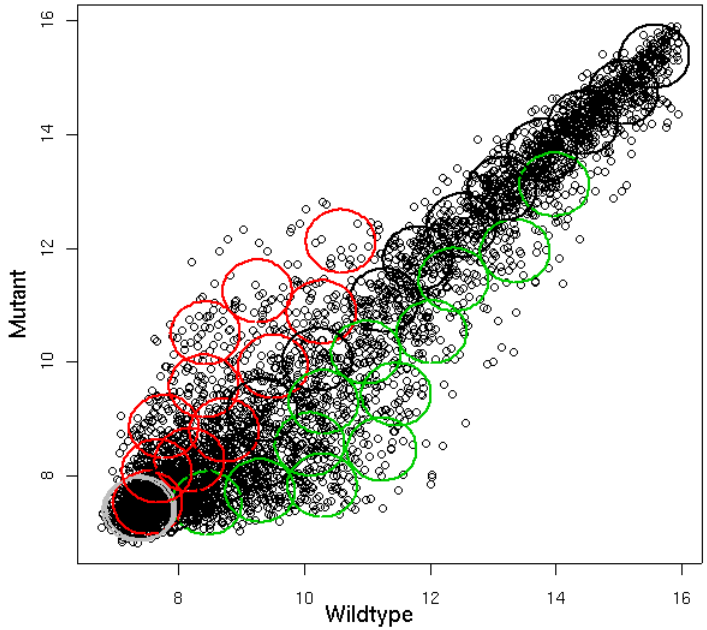


- EM inference with Z and W

Initialisation of the EM algorithm

- It is essential to know the components arrangement for each group





- Most methods deal with the combination of components
- We propose to extend the method of Baudry et al. (2010) to the case of HMM.

Hierarchical clustering

Objective: heuristic to merge L components into K groups

- Three likelihood-based merging criteria:

$$\begin{aligned}
\nabla_{ij}^1 &= \mathbb{E}\left[\log P(X; G'_{i\cup j})|X\right] \\
\nabla_{ij}^2 &= \mathbb{E}\left[\log P(X, Z, W; G'_{i\cup j})|X\right] \\
\nabla_{ij}^3 &= \mathbb{E}\left[\log P(X, Z; G'_{i\cup j})|X\right]
\end{aligned}$$

Remark:

$$\begin{aligned}
\nabla_{ij}^1 &= BIC(G'_{i\cup j}) - BIC(G) + cst \\
\nabla_{ij}^2 &= ICL_W(G'_{i\cup j}) - ICL_W(G) + cst \\
\nabla_{ij}^3 &= ICL_Z(G'_{i\cup j}) - ICL_Z(G) + cst
\end{aligned}$$

Selection of the number of groups

- Independent framework:
  - The likelihood still remains the same when 2 components are merged
  - The number of free parameters only depends on L

  $\Rightarrow BIC$ and ICL_W do not depend on the number of groups K
  $\Rightarrow ICL_Z$ always increases with the number of groups

- HMM framework:
  - The likelihood varies with the number of groups
  - The number of free parameters depends on K and L

$\Rightarrow BIC, ICL_W$ and ICL_Z can be used to estimate the number of groups

Initialisation algorithm

1. Fit a HMM with L components.
2. From $G = L, L-1, ..., 1$
   - Select the components i and j to be combined as:

$$(i, j) = \underset{k,\ell \in \{1,...,G\}^2}{\operatorname{argmax}} \nabla^{?}_{k\ell},$$

   - Model with $G-1$ groups where the density of the component i' is fitted by the mixture distributions of components i and j.
   - Update the parameters with few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of groups $\widehat{K}$:

$$\hat{K} = \underset{\ell \in \{L,...,1\}}{\operatorname{argmax}} \textbf{crit}\ (\ell)$$

Initialisation algorithm

1. Fit a HMM with L components.
2. From $G = L, L-1, ..., 1$
   - Select the components i and j to be combined as:

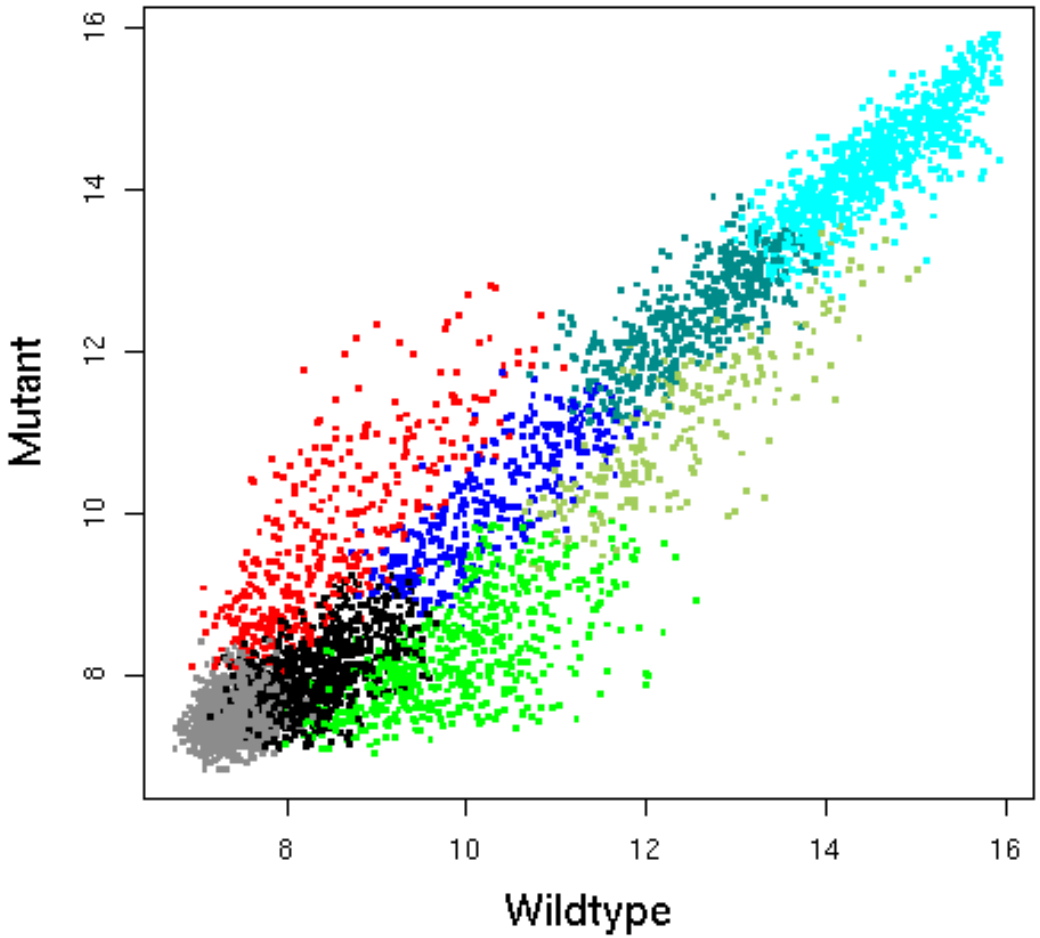
$$(i, j) = \underset{k,\ell \in \{1,...,G\}^2}{\operatorname{argmax}} \nabla^{\mathbf{1}}_{k\ell},$$

   - Model with $G-1$ groups where the density of the component i' is fitted by the mixture distributions of components i and j.
   - Update the parameters with few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of groups $\widehat{K}$:

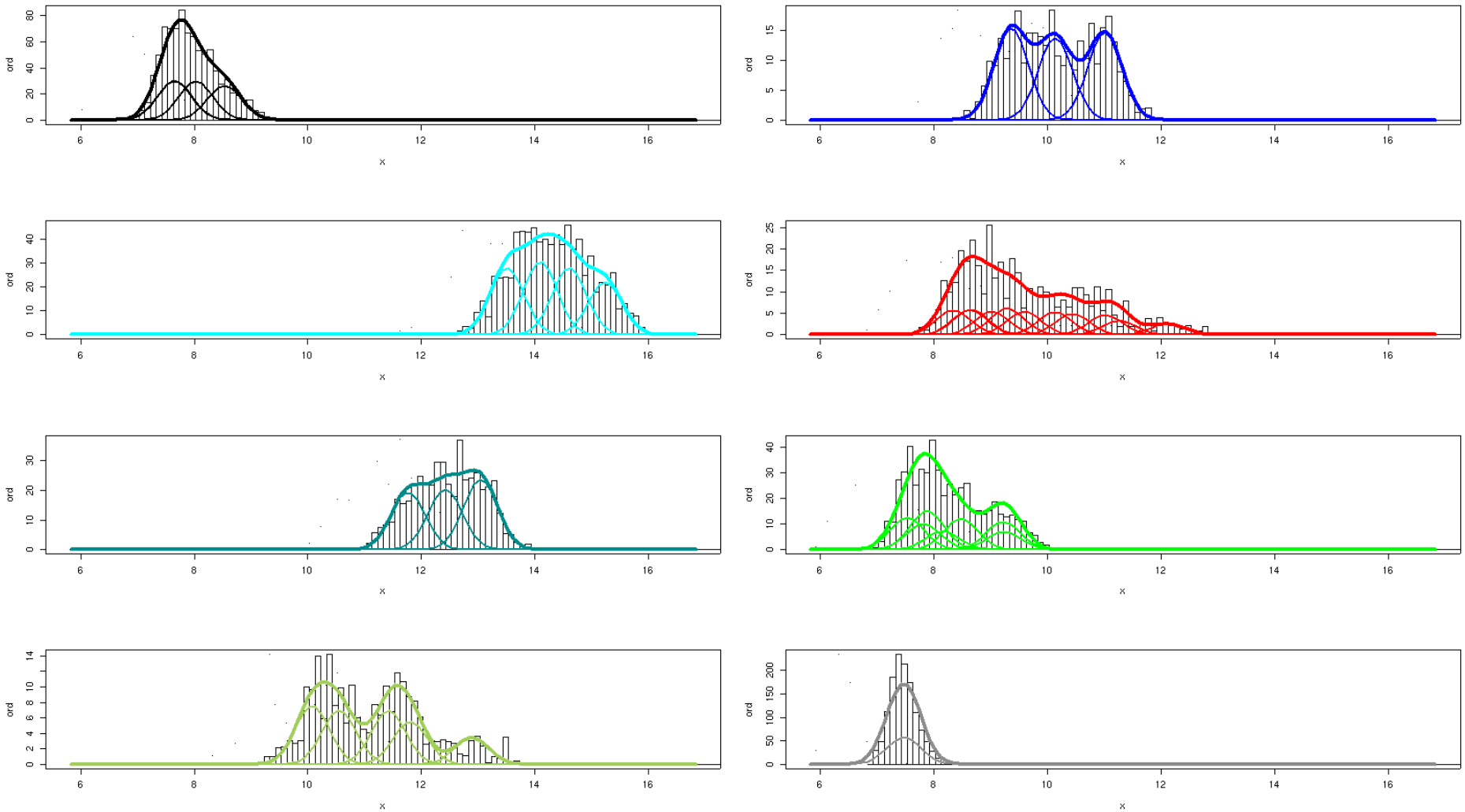
$$\hat{K} = \underset{\ell \in \{L,...,1\}}{\operatorname{argmax}} \mathbf{ICL_Z}(\ell)$$

ChIP-chip IP/IP dataset - Sample of 5000 probes

- Starting from a HMM with 40 components
- The number of groups given by ICL_Z is 8.
- The proportions of the over and under-methylated groups are 6.5% and 15.5%



Fit of the estimated densities for each group



Conclusion

- General modeling of the hybridized signal using latent variable models
  - Use the whole available information of the probes.
  - Adapted model according to the biological question
    - ★ Mixture of regressions
    - ★ Bidimensional Gaussian mixture
    - ★ Mixture of mixture
- Classification by probe and by regions
  - Control of false positive (independent case, $K = 2$)
  - Generalization of the posterior probabilities for a region

Perspectives

- Modeling issues
  - ⋆ Integration of annotation in the mixture of mixture model
  - ⋆ Next Generation Sequencing (NGS) technologies
  - ⋆ Comparison of more than 2 conditions
- Inference issues
  - ⋆ How to choose the initial number of components L
  - ⋆ Very high computational time: Pruning criterion for considering only the most likely components
- Biological issues
  - ⋆ Validation of the new transcripts detected