Linking different kinds of Omics data through a model-based clustering approach

V Vandewalle, Camille Ternynck, Guillemette Marot

Linking different kinds of Omics data through a model-based clustering approach

V. Vandewalle[1,2], Camille Ternynck[1], Guillemette Marot[1,2]

[1] Université de Lille, EA 2694
[2] Inria Lille, Modal team

IFCS Meeting
Thessaloniki
August 28th, 2019

Heterogeneous Omics data application

Central dogma of molecular biology

DNA $\rightarrow$ **RNA** $\rightarrow$ Proteins

Gene expression measurements

- microarray measurements: continuous variables
- RNAseq measurements: count variables

For gene g

		Patient 1	Patient 2	$\cdots$	Patient n
Gene g	microarray 1	x_{g11}	x_{g12}	$\cdots$	x_{g1n}
	$\vdots$				
	microarray n_g	x_{gn_g1}	x_{gn_g2}	$\cdots$	x_{gn_gn}
	RNAseq 1	y_{g11}	y_{g12}	$\cdots$	y_{g1n}
	$\vdots$				
	RNAseq m_g	y_{gm_g1}	x_{gm_g2}	$\cdots$	x_{gm_gn}

Usual related questions

- High dimensional framework with many genes
- Low number of subjects
- Differential expression analysis: find gene with expression related to some disease
- Normalization of the data: remove some undesirable bias
- Often RNA transformations performed (log-transformation, normalized log-transformation, variance stabilizing transformation, ...) to treat this count variable as a continuous one

Considered framework

Initial questions

- Data normalisation framework: make RNAseq and microarray on the same "scale".
- How can we detect that some RNAseq and microarray measurements belong to the same genes (group of genes)?

Clustering of genes

1. Use the genes memberships information of the measurements,
2. Not use the genes membership information (in many settings gene information not available) $\Rightarrow$ how to cluster variables from the same gene?

Similarity criterion

Cluster together genes which have the same distribution with respect to both microarray and RNAseq measurements.

Considered framework

Initial questions

- Data normalisation framework: make RNAseq and microarray on the same "scale".
- How can we detect that some RNAseq and microarray measurements belong to the same genes (group of genes)?

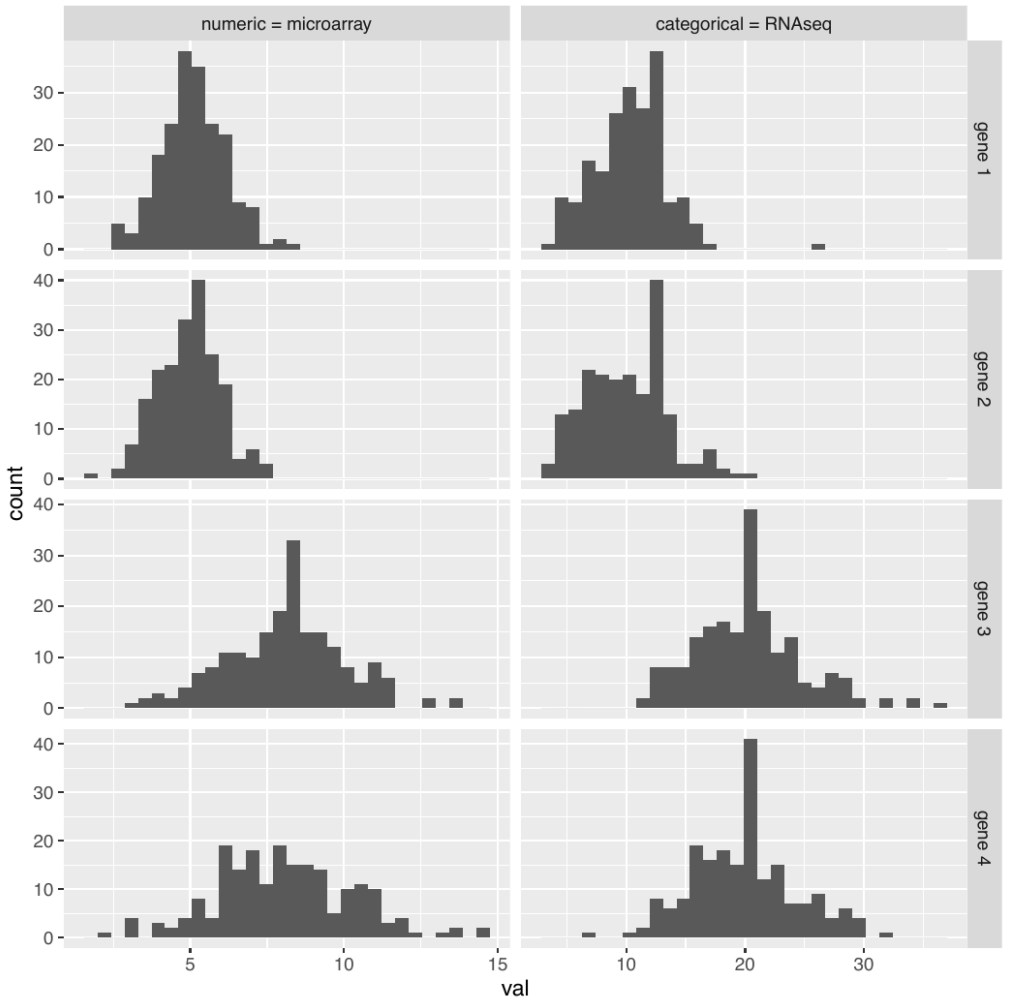
Clustering of genes

1. **Use the genes memberships information of the measurements,**
2. Not use the genes membership information (in many settings gene information not available) $\Rightarrow$ how to cluster variables from the same gene? **Semi-supervised setting, partial gene membership information!**

Similarity criterion

Cluster together genes which have the same distribution with respect to both microarray and RNAseq measurements.

Illustration



- Information on the distribution

$\Rightarrow$ forget the individual level

$\simeq$ symbolic data in some sense

Data

- n patients, described by G genes
- for each patient i, each gene g is described by:
  - n_g continuous microarray measurements $x_{g1i}, \ldots, x_{gn_gi}$
  - m_g count RNAseq measurements $y_{g1i}, \ldots, y_{gm_gi}$
- assume that the genes come from K different clusters
- $z_g \in \{0,1\}^K$ denoting the cluster for gene g in binary coding.

Model

- For each gene g, z_g is a realization of the random variable $Z_g \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$
- $Z_1, \ldots, Z_G$ are assumed independent and identically distributed
- x_{gji} the microarray measurement for the measure number j ($j \in \{1, \ldots, n_g\}$) of the patient number i for the gene g is a realization of the random variable X_{gji}.
- $y_{g\ell i}$ the RNAseq measurment for the measure number ℓ ($\ell \in \{1, \ldots, m_g\}$) of the patient number i for the gene g is a realization of the random variable Y_{gji}.

Detail of the model assumptions

Parametric model

- measures on each patient are assumed independent
- given the gene cluster each measure are independent
- $X_{gji}|Z_g = k \sim \mathcal{N}(\mu_k; \sigma_k)$
- $Y_{g\ell i}|Z_g = k \sim \mathcal{P}(\lambda_k)$

Remarks

1. λ_k should be an increasing function of μ_k, but this contraint is not imposed to simplify the estimation process (no need of a link function between λ_k and μ_k)
2. gene membership of each probe assumed to be known: often the case for human expression data, however not for bacterial data ...
3. model able to cope with partial gene information, however for identifiability reasons it is needed that each cluster contains at least one gene for which we have observed both RNAseq and microarray measurements.

Parameters estimation (1/2)

Parameters estimated by maximum likelihood through the EM algorithm

E-step

$t_{gk}^{(r)} = P(Z_{gk} = 1 | \mathbf{x}_g, \mathbf{y}_g; \theta^{(r)})$

with

$$P(Z_{gk} = 1 | \mathbf{x}_g, \mathbf{y}_g) \propto \pi_k \prod_i^n \left[\left(\prod_{j=1}^{n_g} f(x_{gji}; \alpha_k) \right) \times \left(\prod_{\ell=1}^{m_g} h(y_{g\ell i}; \beta_k) \right) \right]$$

where f is the pdf for microarray data and h the pdf for RNAseq data.

M-step

$$\pi_k^{(r+1)} = \frac{\sum_{g=1}^G t_{gk}^{(r)}}{G}$$
$$\alpha_k^{(r+1)} = \arg\max_{\alpha_k} \sum_{g=1}^G t_{gk}^{(r)} \sum_{i=1}^n \sum_{j=1}^{n_g} \ln f(x_{gji}; \alpha_k)$$
$$\beta_k^{(r+1)} = \arg\max_{\beta_k} \sum_{g=1}^G t_{gk}^{(r)} \sum_{i=1}^n \sum_{\ell=1}^{m_g} \ln h(y_{g\ell i}; \beta_k),$$

Remarks

- the model collapses all the data of a same gene g
- clustering of genes $\simeq$ clustering of distributions

Parameters estimation (2/2)

Computations considerably speed up by the use of sufficient statistics:

$$\mu_k^{(r+1)} = \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \frac{1}{n \times n_g} \sum_{i=1}^{n} \sum_{j=1}^{n_g} x_{gji} = \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \bar{x}_g$$

$$\begin{aligned} \sigma_k^{2(r+1)} &= \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \frac{1}{n \times n_g} \sum_{i=1}^{n} \sum_{j=1}^{n_g} (x_{gji} - \mu_k^{(r+1)})^2 \\ &= \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \sigma_{x_g}^2 + \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} (\mu_k^{(r+1)} - \bar{x}_g)^2 \end{aligned}$$

$$\lambda_k^{(r+1)} = \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \frac{1}{n \times m_g} \sum_{i=1}^{n} \sum_{\ell=1}^{m_g} y_{g\ell i} = \frac{1}{n_k^{(r+1)}} \sum_{g=1}^{G} t_{gk}^{(r)} \bar{y}_g$$

gene g can be summarized by:

- M-step: $(\bar{x}_g, \sigma_{x_g}^2, \bar{y}_g)$.
- E-step: n_g and m_g additionally needed.
- log-likelihood: $\sum_{\ell=1}^{m_g} \sum_{i=1}^{n} \log(y_{g\ell i}!)$ additionally needed.

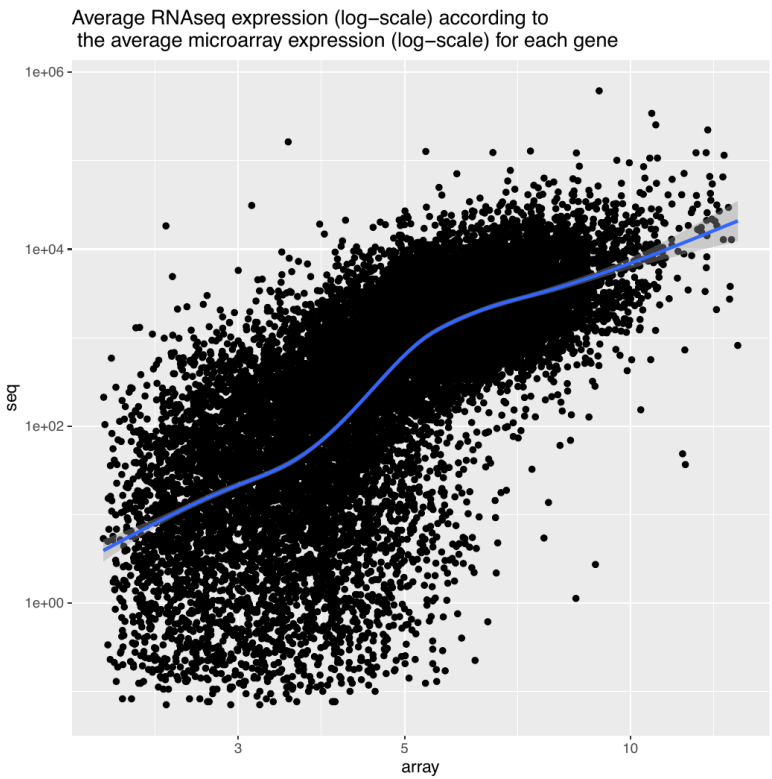
Presentation of the data

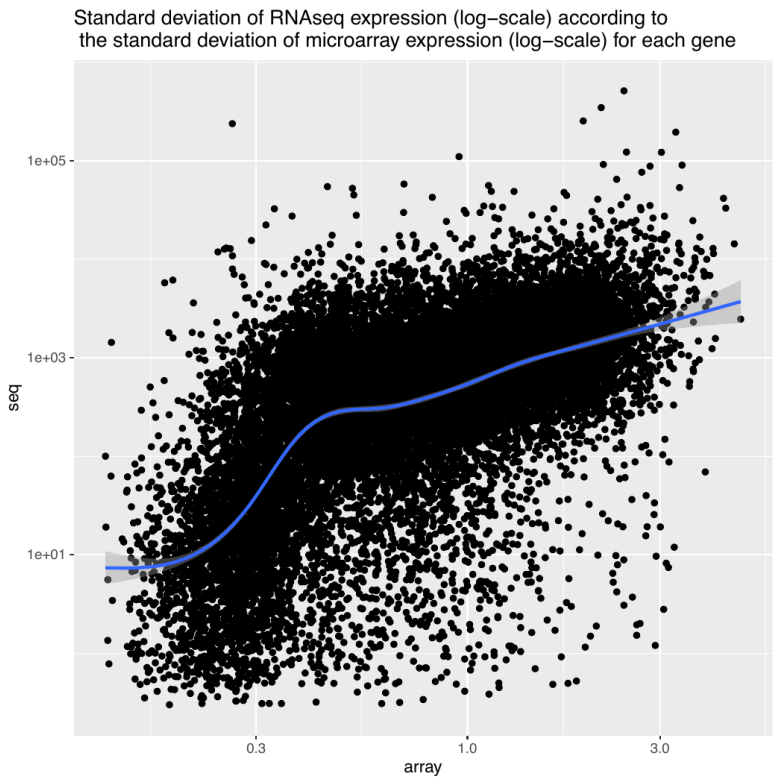
The Cancer Genome Atlas (TCGA) dataset

- Data about the Acute Myeloid Leukemia [1]
- 169 patients
- Keep genes with:
  - Only one RNAseq measure
  - Non null RNAseq values for at least 9 patients
  - At least one microArray measure
  - Keep probes related to only one gene
- At the end: 16,741 genes with an RNAseq measure and one or more microarray measures

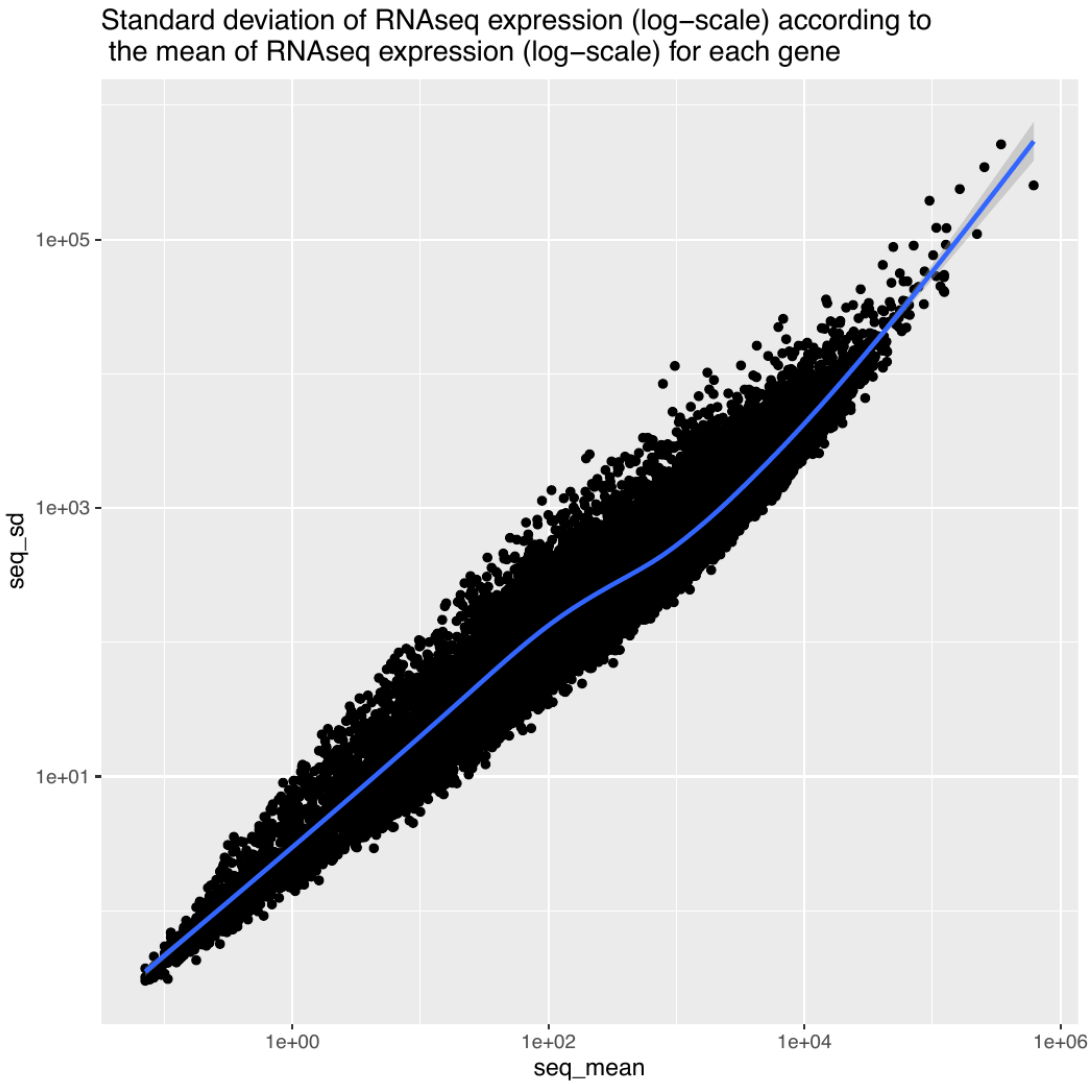
[1] https://portal.gdc.cancer.gov/projects/TCGA-LAML

Link RNAseq / microarray

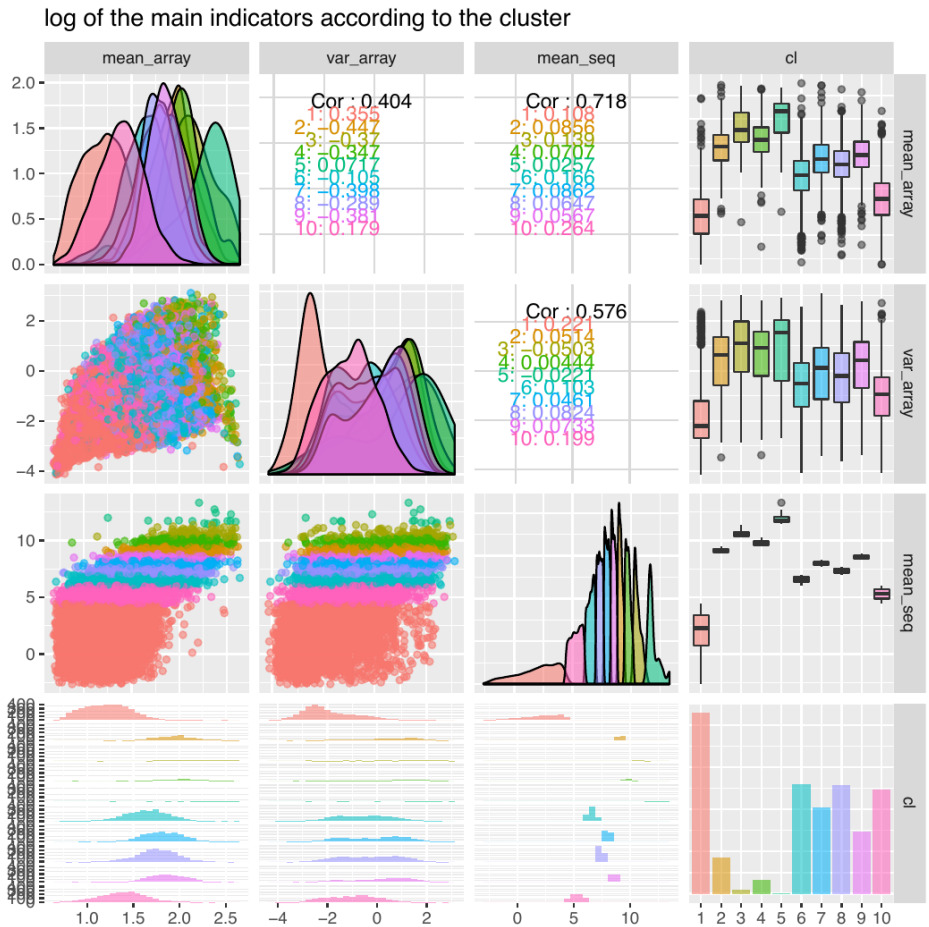




Link between standard deviation and average for RNAseq data



Results of a clustering in 10 clusters (1/2)



Remarks

- It is essentially the RNAseq which determines the cluster membership!
- How to explain this particularly big influence?

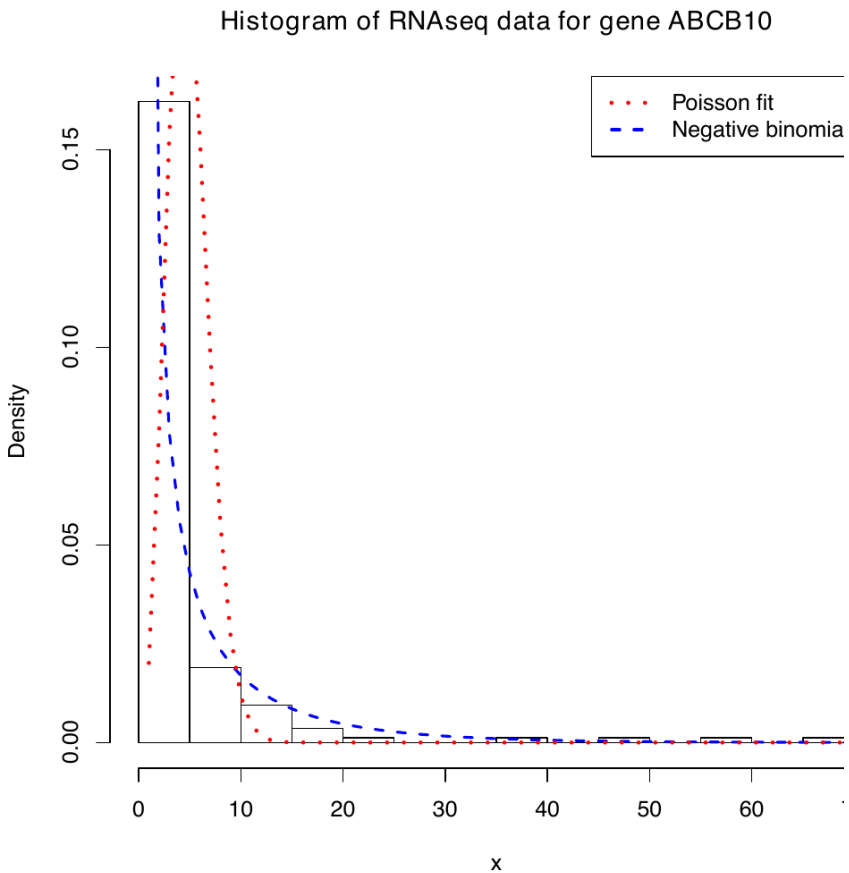
Results of a clustering in 10 clusters (1/2)

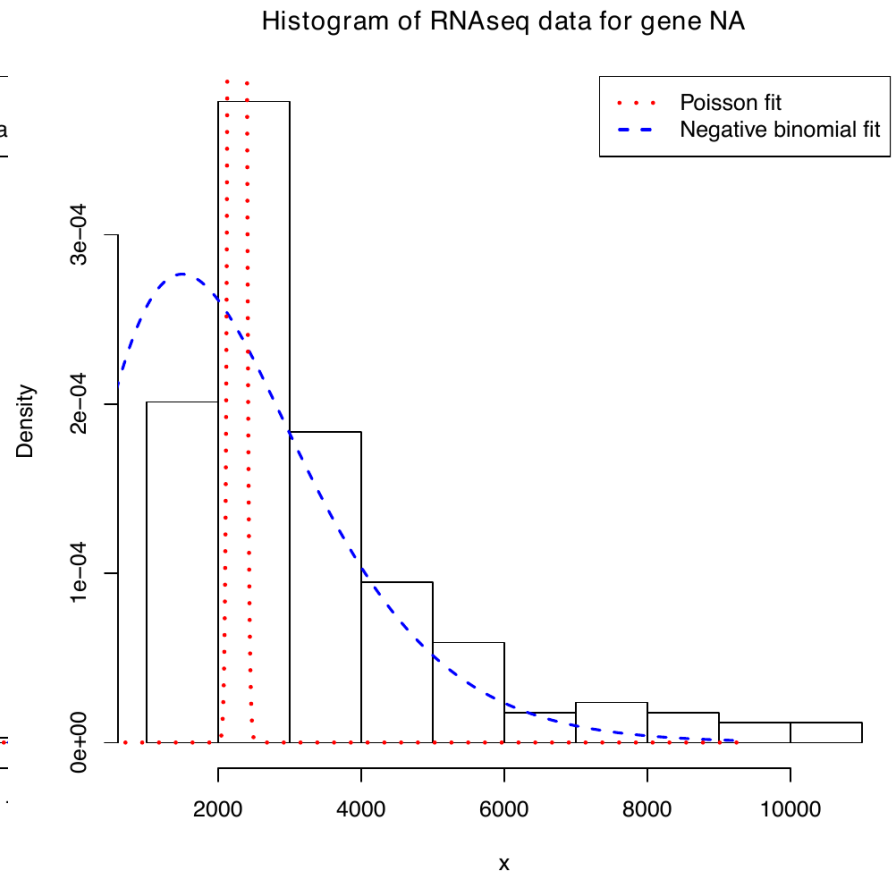
Model parameters:

cluster	prop	mean_array	var_array	mean_seq
1	0.26	3.46	1.14	19.53
2	0.15	5.27	2.64	736.30
3	0.15	5.92	2.98	1577.26
4	0.15	4.13	1.74	220.13
5	0.12	6.32	3.57	2938.47
6	0.09	6.60	3.88	5207.63
7	0.05	7.15	4.96	9332.41
8	0.02	7.80	6.45	18148.48
9	0.01	8.79	8.98	41936.76
10	0.00	9.61	12.76	179799.30

- Small clusters have an higher mean
- The mean array value is monotetic according to the mean seq value
- Huge differences according to the cluster for mean seq value

Explanation of the influence of RNAseq





- Count for RNAseq overdispersed

$\Rightarrow$ Poisson not accurate, the mixture model prioritizes to improve this bad fit,

$\Rightarrow$ better fit using negative binomial distribution.

Negative binomial distribution fit

$$p(y|r,p) = \frac{\Gamma(r+y)p^r(1-p)^y}{\Gamma(r)\Gamma(y+1)}$$

r: size, p: probability, $\mu = r(1-p)/p$: expectation

Maximum likelihood estimation

Expression of p according to r and the data

$$p = \frac{Nr}{Nr + \sum_{i=1}^{N} y_i}$$

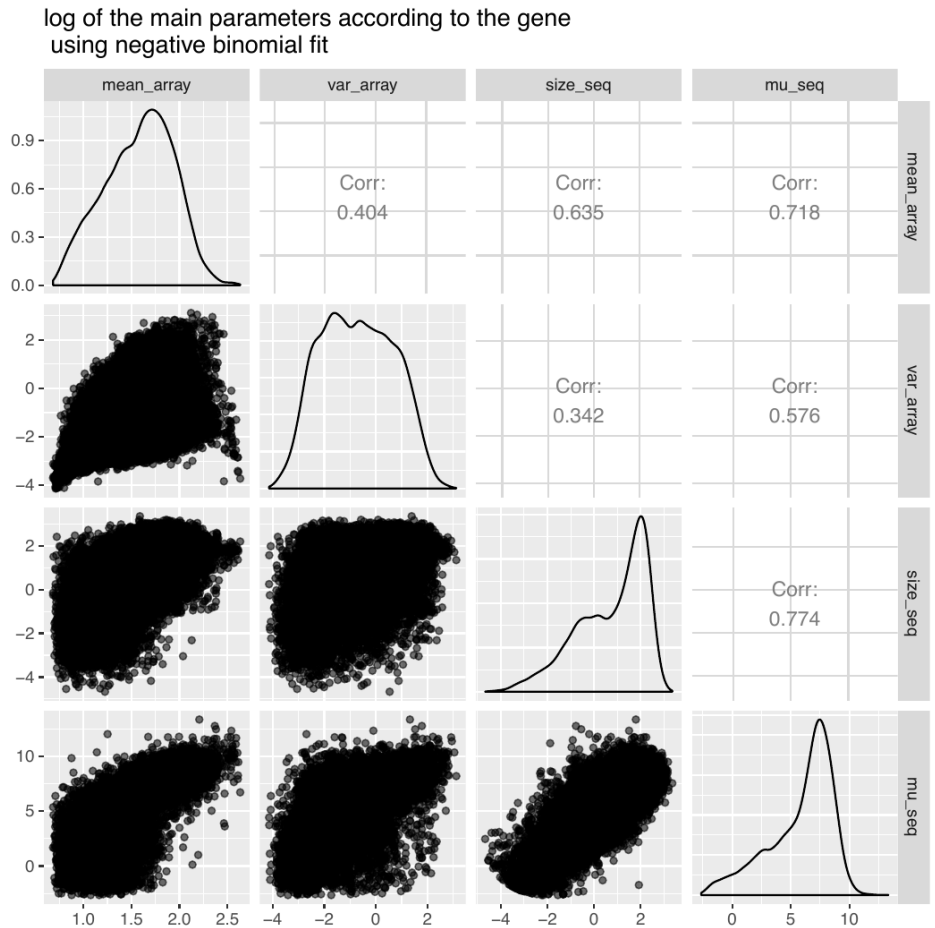
Partial derivate according to r

$$\frac{d\ell(r|y_1,\ldots,y_N)}{dr} = N \ln\left(\frac{Nr}{Nr + \sum_{i=1}^{N} y_i}\right) - N\Psi(r) + \sum_{i=1}^{N} \Psi(r+y_i) = 0,$$

Ψ: digamma function, *i.e.* derivate of the Gamma function.

- Non-closed form, found by numerical optimisation
- Algorithm implemented in fitdistrplus
- No sufficient statistic available due to the digamma function

New preliminary analysis based on a negative binomial fit on each gene

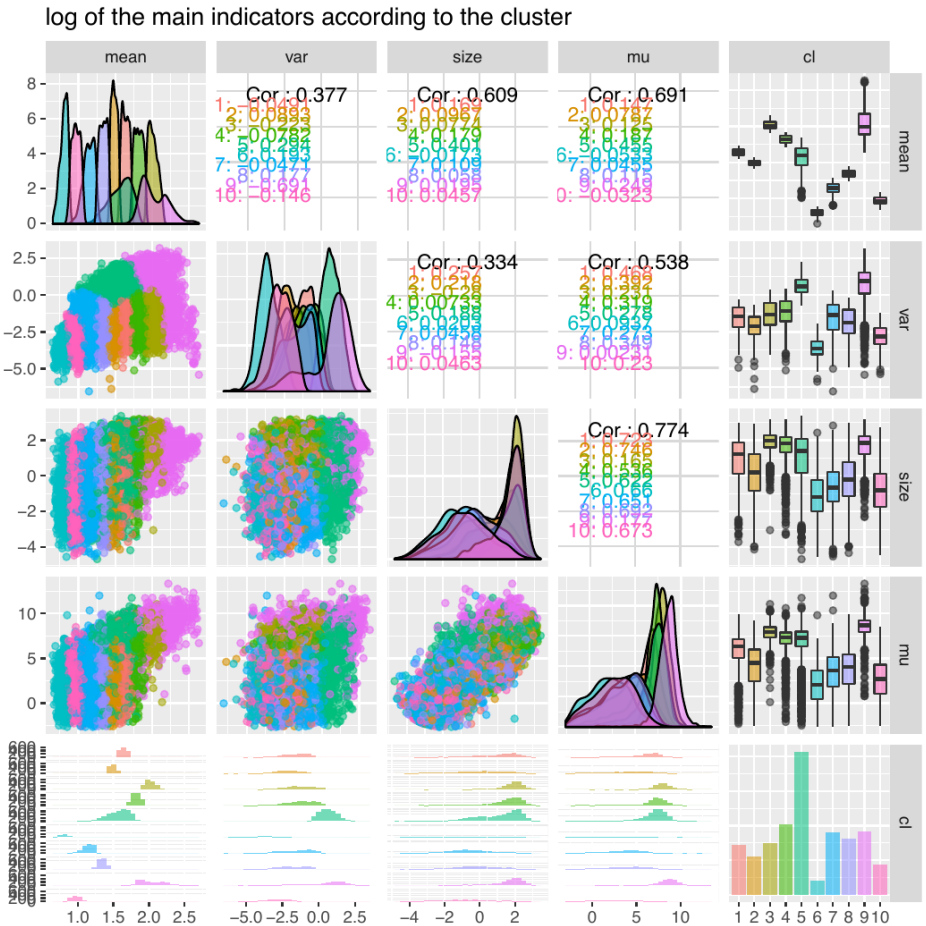


Outlier removed (log-value) :

gene	mean_array	var_array	size_seq	mu_seq
TBC1D21	1.26	-2.26	12.55	-2.24

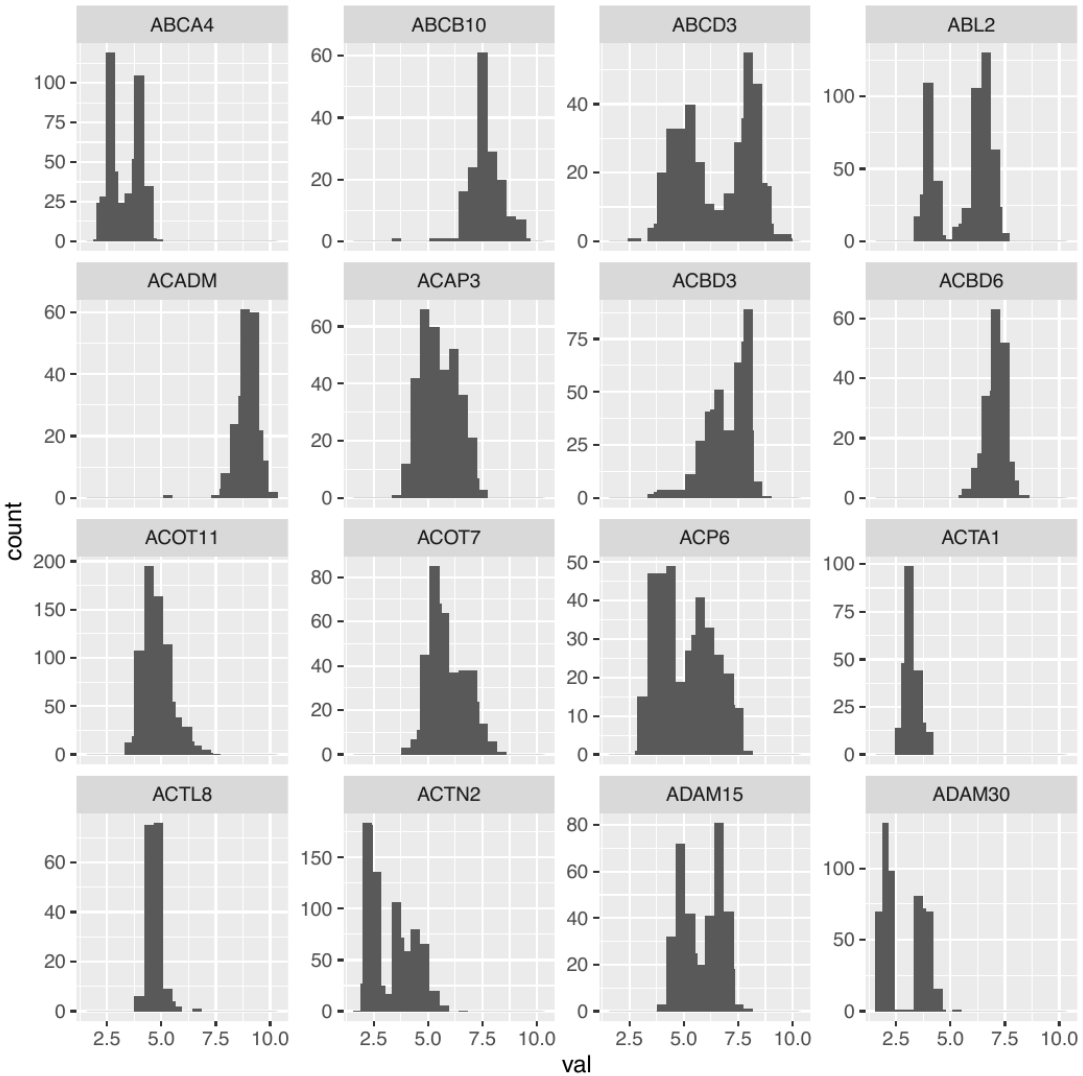
Results with the model based clustering approach

Clustering in 10 clusters



- Now it seems that microarray data dominates the clustering process
- It should be mainly due to bad fit of Gaussian distribution

Explication of the influence of microarray data



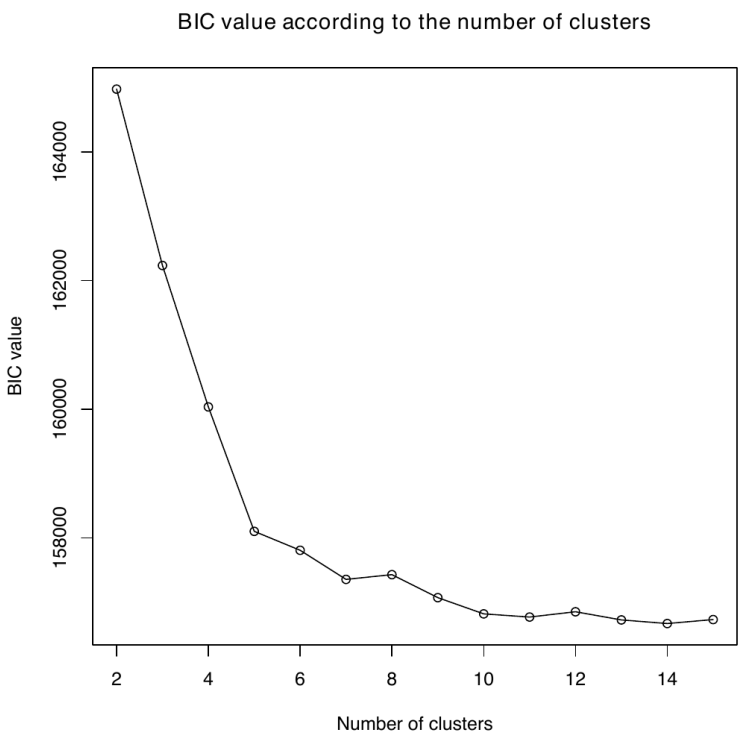
- Unimodal distribution not accurate for a lot of variables.
- Consider mixture of Gaussian

$\Rightarrow$ Increase the cost of the algorithm, work in progress

...

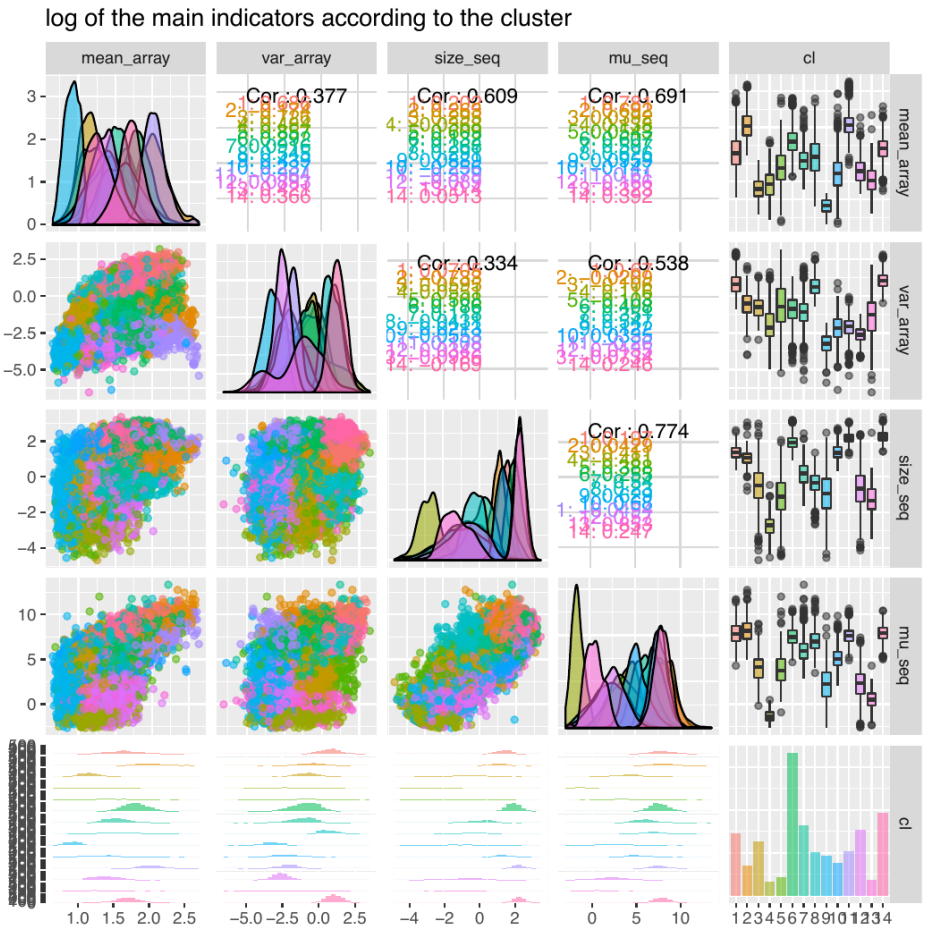
Clustering of genes based on summary of each gene

- **Rationale: as the number of patients n increases, the summary of each gene becomes consistent.**
- Use clustering of genes based on summaries provided by Gaussian and negative binomial fit for each gene.
- Model based clustering with mixture of gaussian of these summaries
- Easier to balance the influence of each kind of data.



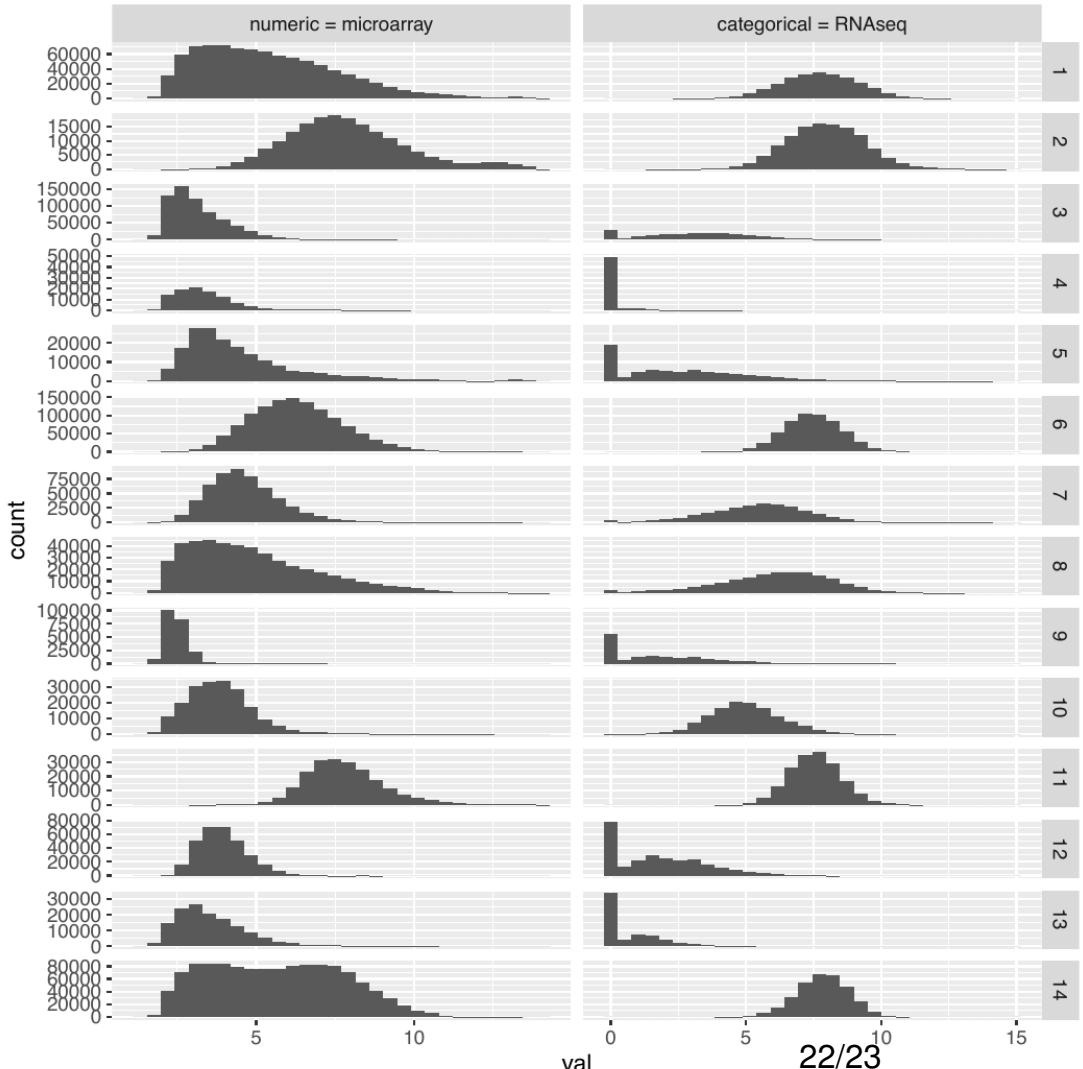
BIC indicates approximately 14 clusters

Analysis of the results of the Gaussian mixture



- More accurate results, however cluster poorly separated
- ICL selects 3 clusters

Resulting empirical distribution in each cluster on the initial data



- Can be used to cluster new data based only on RNAseq
- Then deduce the distribution of microarray data given the observed RNAseq data

Conclusion and perspective

Conclusion

- Approach very sensitive to model assumptions
- Difficulty to balance the influence of each kind of variables
- Empirical solution through clustering of summaries

Perspectives

- Work in progress to stabilize the approach
- Take into account dependency of the measurements for a patient
- Simultaneous clustering of patients and genes