Model comparison criteria in a global analysis of a microarray experiment

C. Díaz [1], N. Moreno-Sánchez [1], J. Rueda [2], A. Reverter[3], YH Wang[3], MJ Carabaño[1]

[1] Depto. de Mejora Genética Animal, INIA Madrid, Spain.
[2] Depto. de Genética, Universidad Complutense, Madrid, Spain.
[3] CSIRO Livestock Industries and Cooperative Research Centre for Cattle and Beef Quality, Brisbane, Australia.

INTRODUCTION (I)

- Microarray experiments allow characterization of overall patterns of gene expression to understand which genes are transcribed and how this process is regulated.

- Noisy technique: spot to spot variability, labelling efficiencies that may affect the definition of differences for a given gene under two (o more) conditions.

- The data normalization and analysis processes aim at identifying what part of measures transcript values are due to the biological variation.

- Normalization and analysis at gene vs. global level (Kerr, 2003)

- Global in two (Wolfinger et al., 2001) vs. one step (Reverter et al., 2003; 2004)

INTRODUCTION (II)

- Heterogeneity of variances according to level of intensity (Dudoit et al. 2002; Kerr et al., 2002). Such low intensity readings show large variability.

- There are not many attempts to study models for normalization and data analysis jointly and accommodate heterogeneous variance according to level of intensity.

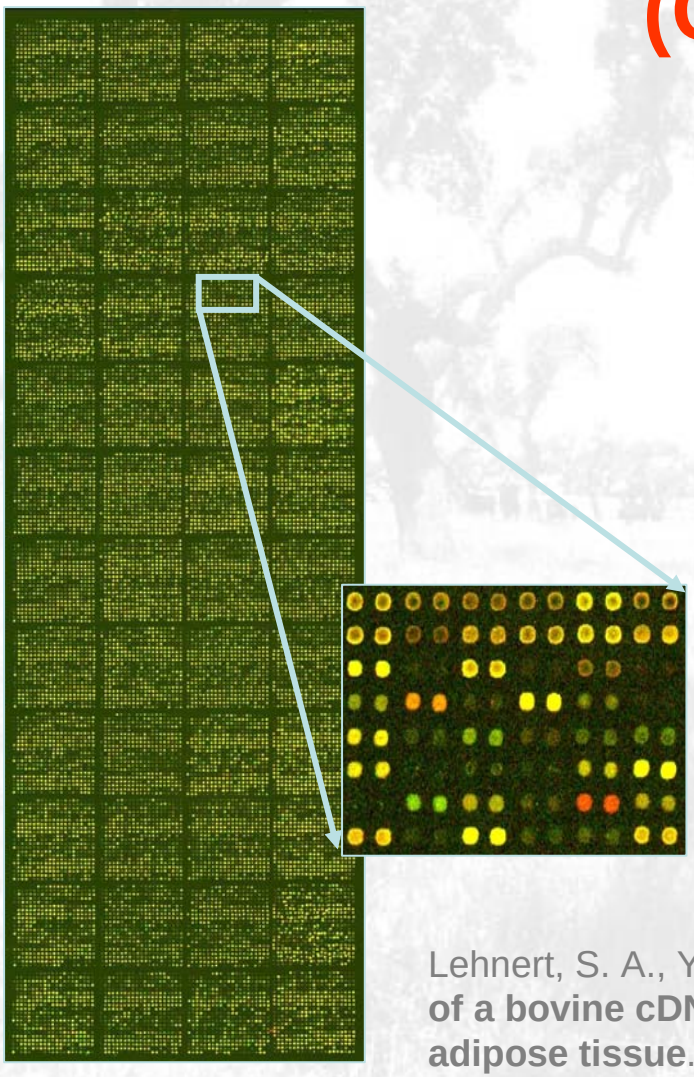
- Bayesian analysis is a flexible tool to approach model selection.

OBJECTIVE

To **assess alternative models for data normalization and analyses** of an experiment to identify DE genes between two skeletal muscles in Avileña Negra Ibérica calves

Bovine Fat & Muscle cDNA microarray

(CSIRO & CRC)



Number of array probes	9,934 in duplicate
Array probes with functional annotation	3,411 probes
Array probes for genes of unknown function	Ca. 6,500
Candidate genes	300

Lehnert, S. A., Y. H. Wang, and K. Byrne. 2004. **Development and application of a bovine cDNA microarray for expression profiling of muscle and adipose tissue**. Australian Journal of Experimental Agriculture, 44, 1127ø1133.

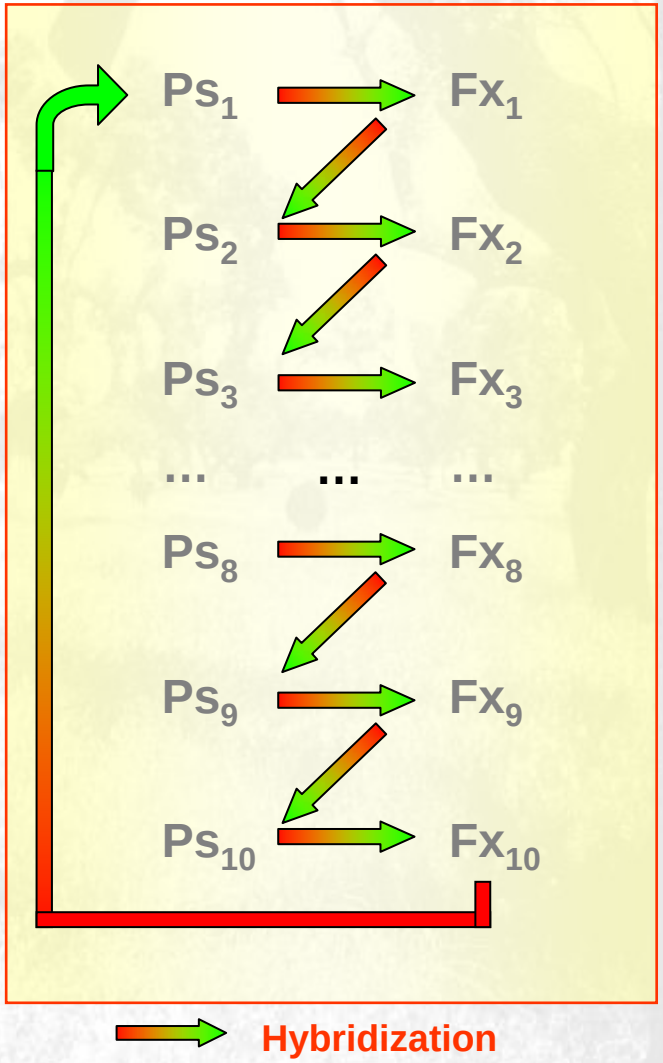
Material and Experimental Design

Samples:

- Two skeletal muscles (Ps and Fx).
- 10 male calves under same fattening conditions and average slaughter age 426 days.

20 Hybridizations:

- Loop design
- Dye swapping



DATA ACQUISITION

1. Flagged spots = out.
2. Spots with Bg > Fg = out.
3. Signals with S2N<1 & M2N<0.85 = out
4. Within or between arrays unreplicated records = out.

Initial # of spots	Final # of spot
402,192	134,856
100%	33.5%

269,712 intensity readings remained for the data normalization process

8538 clones

STATISTICAL ANALYSIS - MODELS

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{g}}\mathbf{g} + \mathbf{Z}_{ag}\mathbf{ag} + \mathbf{Z}_{dg}\mathbf{dg} + \mathbf{Z}_{\mathbf{mg}}\mathbf{mg} + \mathbf{Z}_{cg}\mathbf{cg} + \mathbf{e}$$

ß **= vector systematic effects (dye, muscle, array/block, interactions). Normalization part of model**

g **= the random vector of clones.**

ag **= the random vector of clones by array.**

dg **= the random vector of clones by dye.**

mg **= the random vector of clones by muscle. Objective**

cg = the random vector of clones by animal.

BAYESIAN INFERENCE

- **Sampling distribution**

  y | β, g, ag, mg, dg, g, σ^{2}'s , R ~ MVN

- **Prior distributions for the unknowns**

  Location parameters, β ~ MVN
  Residual variance: ~ χ-2
  Gene variances ~ χ-2

- **Gibbs sampling: Coupled chains (20,000 burn-in/ 100,000 total)**

STATISTICAL ANALYSIS - MODELS

Effects/ Models	A / AB	D	M	DM	AD/ABD	AG/ABG	MG	AG
# levels	19/912	2	2	4	38/1824	60998/73523	17084	73523
M1	ø/+	+	+	+	ø/+	ø/+	+	+
M2	ø/+	+	+	ø	ø/+	ø/+	+	+
M3	ø/+	+	ø	ø	ø/+	ø/+	+	+
M4	ø/+	+	ø	+	ø/+	ø/+	+	+
M5	ø/+	ø	+	ø	ø/+	ø/+	+	+
M6	ø/+	ø	+	+	ø/+	ø/+	+	+
M7	ø/+	+	+	+	ø	ø/+	+	+
M8	+/ ø	+	+	+	+/ ø	+/ ø	+	+
M9	+/ ø	+	+	+	+/ ø	ø/+	+	+
M10	ø/+	+	+	+	ø/+	+/ ø	+	+
M11	+/ ø	+	+	+	ø/+	+/ ø	ø	+
M12	+/ ø	+	+	+	ø/+	+/ ø	+	ø
M13	+/ ø	+	+	+	ø/+	+/ ø	+	+

STATISTICAL ANALYSIS – Bayesian criteria for model comparison

- Log-Marginal Density of Data(LMD)–“Goodness of fit”
- Cross validation predictive densities.

BAYESIAN MODEL BASED CLUSTER WITH KNOWM NUMBER OF COMPONENTS (Bayesmix, Reverter et al., 2003)

$$\mathbf{d_g = \hat{m}_{g,Ps} - \hat{m}_{g,Fx}}$$

d_g were assumed to be independent measures from a mixture of normal densities such as

$$f(d;\phi_k) = \sum_{j=1}^{k} \pi_j \phi(d_j; \mu_j, V_j)$$

π_j **= mixing proportions**

$\phi(d_j; \mu_j, V_j)$ **= normal density function**

Normalization- Spot to Spot (A vs. AB)

Effects/ Models	A / AB	AG/ABG
# levels	19/912	60998/73523
M1	ø/+	ø/+
M8	+/ ø	+/ ø
M9	+/ ø	ø/+
M10	ø/+	+/ ø

Models	LMD	D	σ_g^2	σ_{ag}^2	σ_{dg}^2	σ_{mg}^2	σ_{cg}^2	σ_r^2	σ_T^2
M1	-99273.46	0.19	3.00	0.48	0.03	0.15	0.07	0.12	3.85
M8	-255354.43	0.46	3.07	0.20	0.03	0.13	0.03	0.38	3.85
M9	-104132.49	0.20	3.00	0.53	0.04	0.15	0.07	0.12	3.92
M10	-244435.28	0.43	3.06	0.17	0.02	0.13	0.03	0.36	3.78

NORMALIZATION– Labelling (DM)

Effects/ Models	AB	D	M	DM
# levels	912	2	2	4
M1	+	+	+	+
M2	+	+	+	ø
M3	+	+	ø	ø
M4	+	+	ø	+
M5	+	ø	+	ø
M6	+	ø	+	+

Models	LMD	D	σ_g^2	σ_{ag}^2	σ_{dg}^2	σ_{mg}^2	σ_{cg}^2	σ_r^2	σ_T^2
M1	-99273.4617	0.1903	3.00	0.48	0.03	0.15	0.07	0.12	3.8
M2	-99155.0005	0.1879	3.00	0.48	0.03	0.15	0.07	0.12	3.8
M3	-99237.4159	0.1906	3.00	0.48	0.03	0.15	0.07	0.12	3.8
M4	-99349.4948	0.1904	3.00	0.48	0.03	0.15	0.07	0.12	3.8
M5	-99288.1110	0.1909	3.00	0.48	0.03	0.15	0.07	0.12	3.8
M6	-99349.4948	0.1942	3.00	0.48	0.03	0.15	0.07	0.12	3.8

NORMALIZATION – Labelling (DA)

Effects/ Models	AB	D	M	DM	ABD	ABG
# levels	912	2	2	4	1824	73523
M1	+	+	+	+	+	+
M7	+	+	+	+	ø	+

Models	LMD	D	σ_g^2	σ_{ag}^2	σ_{dg}^2	σ_{mg}^2	σ_{cg}^2	σ_r^2	σ_T^2
M1	-99273.46	0.19	3.00	0.48	0.03	0.15	0.07	0.12	3.85
M7	-110832.52	0.21	2.98	0.48	0.04	0.15	0.10	0.13	3.88

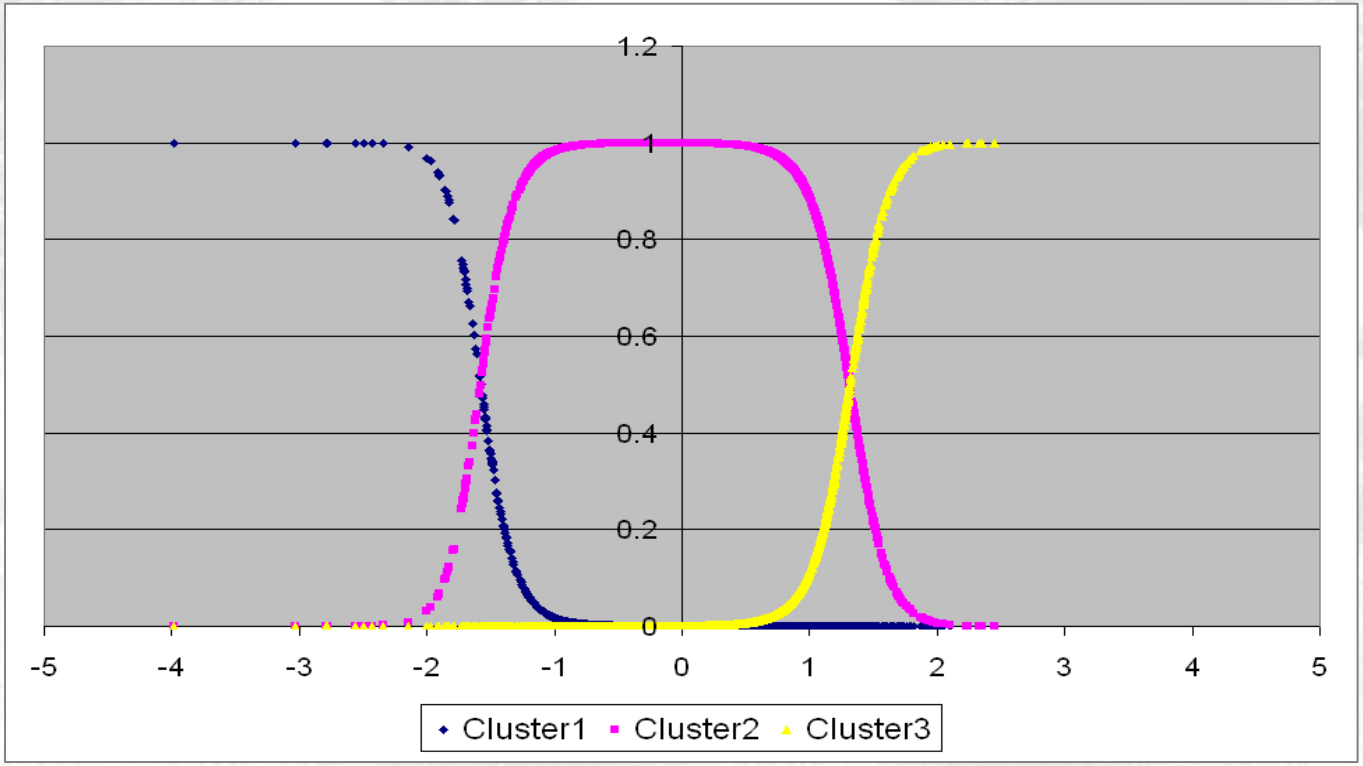
GENE EFFECTS - GxA and GxM HETEROSCEDASTIC RESIDUAL

Values of Log marginal density (LMD), predictive ability of the model (D) and estimates of variance components due to gene (σ^2_g) effect , gene by array (σ^2_{ag}), gene by dye (σ^2_{dg}), gene by muscle (σ^2_{mg}), gene by animal (σ^2_{cg}), residual variance (σ^2_r) and total variance (σ^2_T).

Models	LMD	D	σ^2_g	σ^2_{ag}	σ^2_{dg}	σ^2_{mg}	σ^2_{cg}	σ^2_r	σ^2_T
M1	-99273.46	0.19	3.00	0.48	0.03	0.15	0.07	0.12	3.85
M11	-269669.24	0.41	3.07	-	0.01	0.13	0.15	0.43	3.79
M12	-158504.64	0.29	3.04	0.47	0.05	-	0.19	0.19	3.92
M13	-98064.11	0.21	2.90	0.44	0.04	0.15	0.05	0.13*	3.73*

σ^2_r (3-5)	σ^2_r (5-7)	σ^2_r (7-9)	σ^2_r (9-11)	σ^2_r (11-13)	σ^2_r (13-16)
2.3843	0.2920	0.1116	0.0947	0.1054	0.0789

CLUSTERS



3 clusters (according to the Goodness of fit measured by log L, AIC, BIC):

- **Cluster 1 = Over- expressed in *Flexor digitorum***
- **Cluster 2 = No_ DE**
- **Cluster 3 = Over-expresed in *Psoas major***

DE CLONES

No. clones considering homogeneous variances		No. clones considering heterogeneous variances	
151		198	
No. clones in cluster 1	No. clones in cluster 3	No. clones in cluster 1	No. clones in cluster 3
50	101	41	157
No. genes in cluster 1 (clones with functional annotation)	No. genes in cluster 3 (clones with functional annotation)	No. genes in cluster 1 (clones with functional annotation)	No. genes in cluster 3 (clones with functional annotation)
9	21	8	24

CONCLUSIONS

- Model selection important to approach the normalization and analysis to identify DE genes free of bias.
- Major impact of spot to spot variation, labelling efficiencies across arrays, heterogeneity of variances according to level of intensity.
- Clones were clustered in 3 groups (Non_DE and 2 of over-expression in each muscle)
- Heterogeneity of variances allowed to identify more DE clones.