A hierarchical Bayesian model to study temperature-dependent variation of sequence-specific hybridization to cDNA Microarray

Annibale Biggeri, S. Toti, E. Dreassi *Dept of Statistics, Florence*
C. Lagazio *Dept of Statistical Science Udine*
M. Gasparini *Dept of Mathematics Politecnico Turin*
C. DeFilippo *Dept of Pharmacology Florence*
K.Morneau, A. Bergerat, and D. Cavalieri
Bauer Center for Genomic Research, Harvard University

Introduction

An open issue in Microarray experiments on samples, whose genomic sequence is not necessarily identical to that used to design the probes printed on the array, is the extent of signal variation related to differences in the sequence of the experimental sample.

A second related issue regards the consequences of such phenomena on the inference about relative intensity levels.

To address these topics we designed a microarray experiment on two *Saccharomyces cerevisiae* strains (denoted as L and M) varying the hybridization temperatures.

The extent of signal variation between L and M can depend on hybridization temperature reflecting two distinct issues:

- **differences in the sequence of the experimental sample;**
- **cross hybridization among highly homologous sequences.**

Families of duplicated paralogous genes can share homology up to 98%, lowering the hybridization temperature increases the effect of cross hybridization, and we expect this effect to be multiplicative with the sequence divergence effect.

We predict that the second effect is accounted by the genes that vary at the variation of the temperature in the lab strain itself.

These information is extremely relevant to the use of microarrays in medicine and in population genetics.

Indeed, using expression arrays in expression studies is of paramount importance to be able to distinguish differences in signal expression related from differences associated with sequence variation among individuals, reflecting genetic polymorphism.

Scheme of the presentation

1. Study design
2. Rationale of the study
3. Methods
   1. Single-array inference
   2. Hierarchical Bayesian approach
4. Results
5. Discussion
6. Conclusions

Study design

S. cerevisiae DNA microarrays have been produced amplifying all the genes of genome of the *S. cerevisiae* strain S288c (6218 ORFs) and spotting the purified PCR products on poly-L-lysine coated slides using a robot arrayer.

S288c is the laboratory strain whose sequence has been determined according to the *S. cerevisiae* genome sequencing project.

DNA from two different strains, S288c (**L** from now on) and a wine strain isolated from Montalcino grapes (**M** from now on), was allowed to hybridise to a microarray representing all the yeast coding regions.

Three different hybridization temperatures (50, 55, 60 °C) were used.

Standard Reference design (Kerr and Churchill) with dye swap was used.

Biological rationale

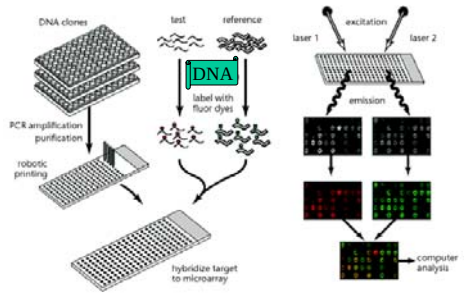
Since we considered two strains of *S. cerevisiae* whose genomes show differences in the genomic sequences, we expect differential hybridisation to occur.

Differential hybridisation can be the result of non-completely specific binding of the wine strain DNA (**M**) to the array containing probes designed on the S288c DNA (**L**) sequence. It reflects the amount of sequence variation in a given probe.

Variation of hybridisation temperature could modulate such partially specific binding reaction to an extent to be quantified, but might also modulate aspecific binding to other probes on the array.

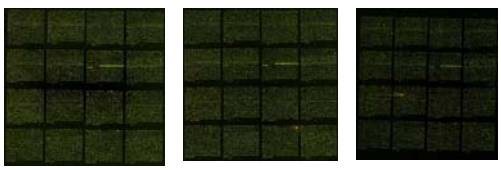
7

Production of the microarray.



8

the three arrays with Green labeled L strain and Red M



Temperature: 50 °C | **Temperature**: 55 °C | **Temperature**: 60 °C

9

Sources of variability

- Gene effects (1,…,G=6218)
- Temperature effects (1=50, 2=55, 3=60)
- Strain effects (L=laboratory, M=Montalcino wine strain)

of interest

- Dye effects (1,2)
- Pin-within slide effects (16 print tips x 6 slides)

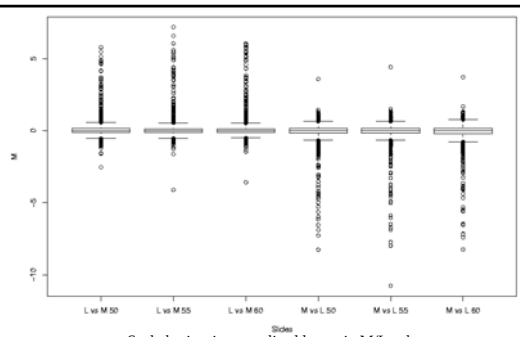
to be removed

10

Methods - 1

- We first conducted a single array analysis. For each array the fluorescence intensities from Genepix were read and elaborated using the system R http://cran.r-project.org and the library SMA http://www.stat.berkeley.edu/users/terry/zarray/Html/smacode.html
- We checked signal vs background and performed simple background subtraction. We analysed the $\log_2$ R/G against the average RG intensities (on the $\log_2$ scale) and performed a within-slide print tip lowess normalization (all spots were used in this phase).
- We did single array inference on the (scale normalized) $\log_2$ratios using the method of Newton (2001).

11



Scaled print-tip normalized $\log_2$ratio M/L values for each of the six arrays

12

Methods – 2: hierarchical Bayesian model

We used cDNA microarray to investigate changes in gene copy number (duplications-deletions) and sequence homology in the laboratory strain (**L**) and the Montalcino strain (**M**).

Gene background-adjusted intensities of the two strains are assumed Gamma distributed (with common Coefficient of Variation = $1/\sqrt{a}$).

$L_i \sim \text{Gamma}(\theta_i^L, a)$

$[L_i|\theta_i^L, a] = \frac{1}{\Gamma(a)}\left(\frac{aL_i}{\theta_i^L}\right)^a \exp\left(-\frac{aL_i}{\theta_i^L}\right)$

$E(L_i) = \theta_i^L$

$Var(L_i) = \theta_i^{L2}/a$

$M_i \sim \text{Gamma}(\theta_i^M, a)$

$[M_i|\theta_i^M, a] = \frac{1}{\Gamma(a)}\left(\frac{aM_i}{\theta_i^M}\right)^a \exp\left(-\frac{aM_i}{\theta_i^M}\right)$

$E(M_i) = \theta_i^L$

$Var(M_i) = \theta_i^{L2}/a$

where

L_i is the intensity for the i-th gene of the laboratory strain and M_i is the intensity for the corresponding gene of the Montalcino strain (i=1,…,G);

We adopted the nonstandard Gamma parametrization of McCullagh and Nelder (1989 page 287), to formulate a Generalized Linear Model for the sequence divergence-related hybridisation intensities.

L_i M_i L_j M_j are independent given the parameters for each *i* and *j* .

The likelihood is given by the product of the two gamma densities.

Similar to Newton (2001).

14

Such model implies that (in a single experiment with no replicates) the **β** terms have a nice interpretation as log fold change:

$$\log_2(\theta_i^L) = \mu_i \qquad \log_2(\theta_i^M) = \mu_i + \beta_i$$

$$\beta_i = \log_2 \frac{E(M_i)}{E(L_i)}$$

15

The model can easily extended to complex study designs

In the experiment considered here, there are three different hybridization temperatures (j=1,…,3) and dye swap (k=1,2). We have six arrays and 16 pins, whose effect must be accounted for (p=1,…,6x16=96).

$$L_{ijkp} \sim \text{Gamma}(\theta_{ijkp}^L, a)$$

$$M_{ijkp} \sim \text{Gamma}(\theta_{ijkp}^M, a)$$

16

Two different models were considered, without and with Temperature effects

1 - $\log_2(\theta_{ijkp}^L) = \mu_i^g + \mu_j^t + \mu_k^d + \mu_p^p$

$\log_2(\theta_{ijkp}^M) = \mu_i^g + \mu_j^t + \mu_k^d + \mu_p^p + \beta_i$

log fold change

The greater the temperature, the lower the absolute intensities (lesser cross-hybridization). Such phenomenon could be strain-specific.

Strain-specific temperature effects

2 - $\log_2\left(\theta_{ijkp}^M\right) = \mu_i^g + \mu_j^t + \mu_k^d + \mu_p^p + \beta_j^t + \beta_i$

17

Two different mixture priors were considered, without and with Temperature effects

- μ terms (and β_j^t terms) are assumed to be distributed as Normal
- Second level Gamma priors are specified on precision parameters of the Normal distributions
- The (squared) coefficient of variation is assumed Gamma distributed
- Gene-strain effects are assumed $N\left(\mu_{\beta_i}, \tau_\beta^*\right)$ (TN=truncated normal)

$$\mu_{\beta_i} \sim \begin{cases} \text{TN}(-7, \tau_\beta) & \text{with prob. } \pi^- \\ \text{TN}(7, \tau_\beta) & \text{with prob. } \pi^+ \\ N(0, \tau_\beta) & \text{with prob. } \pi = 1 - \pi^+ - \pi^- \end{cases}$$

$$(\pi, ^-, \pi^+, 1 - \pi^- - \pi^+) \sim \text{Dirichlet}(\nu_1, \nu_2, \nu_3)$$

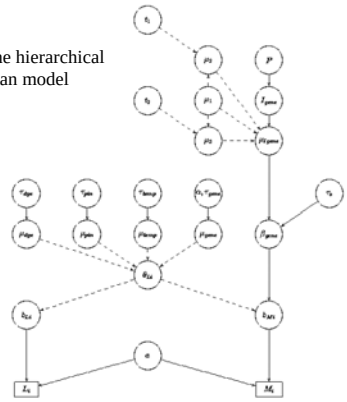
- To model strain-gene specific temperature effects β_{ij} we modified the mixture prior as follows:

$$\mu_{\beta_{ij}} \sim \begin{cases} \mathrm{TN}(\mu_{1ij}, \tau_\beta) & \text{with prob. } \pi^- \\ \mathrm{TN}(\mu_{2ij}, \tau_\beta) & \text{with prob. } \pi^+ \\ N(0, \tau_\beta) & \text{with prob. } \pi = 1 - \pi^+ - \pi^- \end{cases}$$

$$\mu_{rij} = \alpha_r + \gamma_{ij}^t$$

19

Graph of the hierarchical bayesian model



20

Inference

Inference is based on the full posterior distributions approximated by MonteCarlo Markov Chain simulations.

Sampling was performed using a Metropolis algorithm. Acceptance rate was monitored and convergence checked by Gelman-Rubin approach.

We used WinBugs 13 http://www.mrc-bsu.cam.ac.uk/bugs

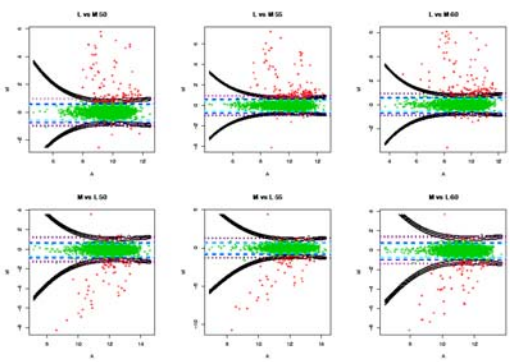
21

Results – Model 1: differential intensities M/L

- Single array analysis identified from 50 to 106 spots, mostly showing genes which are either absent or highly diverged in the Montalcino strain
- Bayesian analysis identified 35 spots, 34 diverged or deleted and 1 present in higher copy number.
- There is strong concordance (32/35 spots) and evident shrinkage in the estimated $\log_2$ratios.

22

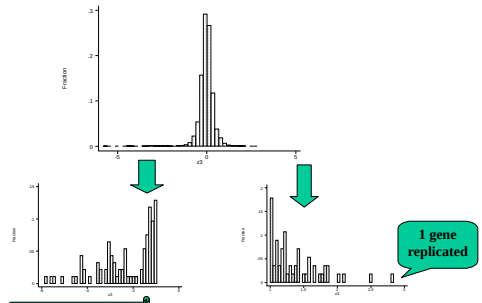
single array analysis (Newton 2001; 99% significant bands)



Gene	Tot-Newton	Bayes	Gene	Tot-Newton	Bayes
YAL064W	6	0	YJL222W	6	0
YAR010C	6	1	YJR026W	6	1
YAR031W	6	1	YJR028W	6	1
YBL005W-A	6	0	YJR029W	6	1
YBL005W-B	3	1	YJR153W	6	1
YBR012W-A	6	1	YLR155C	6	1
YDL246C	6	0	YLR156W	6	1
YDR038C	6	0	YLR157C	6	1
YDR039C	6	0	YLR158C	6	1
YDR040C	6	0	YLR159W	6	1
YEL021W	6	1	YLR160C	6	1
YGL053W	6	1	YLR161W	6	1
YHR054C	6	1	YLR465C	6	0
YHR055C	6	1	YML039W	6	1
YIL015C-A	6	1	YML040W	6	1
YIL080W	4	1	YML045W	6	1
YIL082W	6	1	YMR046C	6	1
YIR042C	6	1	YMR051C	6	1
YJL114W	6	1	YOL156W	6	0
YJL217W	6	1	YOL162W	6	1
YJL218W	6	1	YOL163W	6	1
YJL219W	6	1	YOL164W	5	1

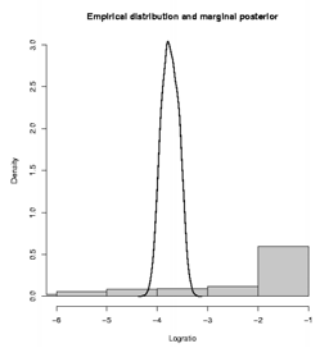
24

empirical distribution of $\log_2$ratios



25

We estimated a log fold change of about –3.7, on average, for the "deleted/diverged" Montalcino genes



Results – Model 2: Temperature effects

Temperature of hybridisation modifies signal intensities

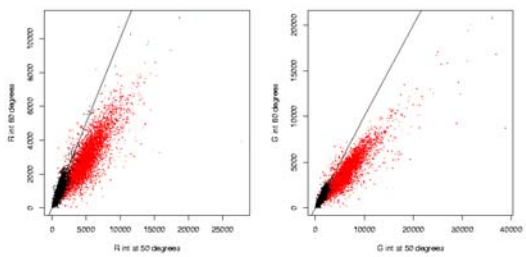
$$\log_2\left(\theta^L_{ijkp}\right) = \mu^g_i + \mu^t_j + \mu^d_k + \mu^p_p$$

$$\log_2\left(\theta^M_{ijkp}\right) = \mu^g_i + \mu^t_j + \mu^d_k + \mu^p_p + \beta^t_j + \beta_i$$

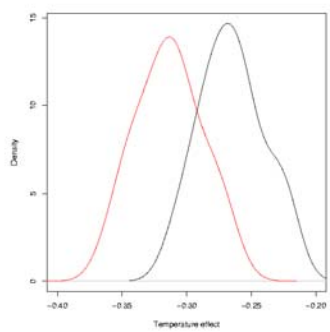
- The temperature effect is more evident for the Montalcino strain (in red) than for the Laboratory strain (in black) – the posterior densities are reported
- The $\log_2$ratios appeared weakly affected, as showed by the posterior density of β^t_j

27

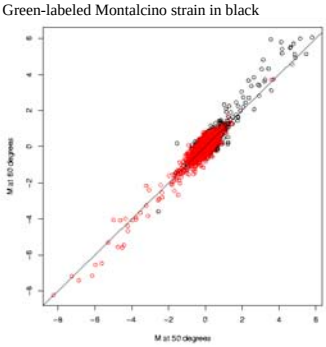
Y-axis: intensity at 60° ; X-axis intensity at 50°
Left: Red dye; Right: Green dye
Montalcino strain (in red) Laboratory strain (in black)



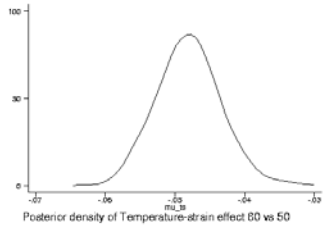
Posterior densities (strain specific temperature effects, j=60)



Y-axis: $\log_2$ratio (M/L) at 60° ; X-axis $\log_2$ratio (M/L) at 50°
Red-labeled Montalcino strain in red
Green-labeled Montalcino strain in black



Posterior density of β_j^t (j=60°)



Results – Mixture prior 2:
Strain-Gene specific Temperature effects

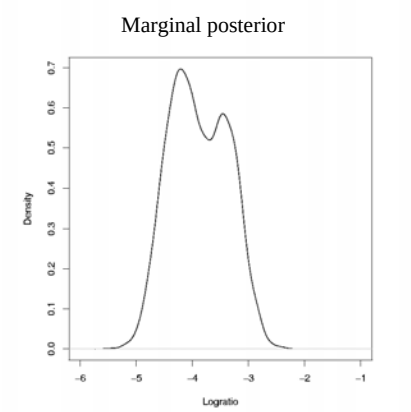
- To model strain-gene specific temperature effects β_{ij} we specified a mixture prior as follows:

$$\mu_{\beta_{ij}} \sim \begin{cases} \mathrm{TN}(\mu_{1ij}, \tau_\beta) & \text{with prob. } \pi^- \\ \mathrm{TN}(\mu_{2ij}, \tau_\beta) & \text{with prob. } \pi^+ \\ N(0, \tau_\beta) & \text{with prob. } \pi = 1 - \pi^+ - \pi^- \end{cases}$$

$$\mu_{rij} = \alpha_r + \gamma_{ij}^t$$

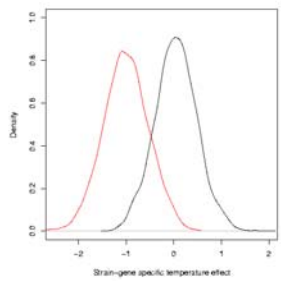
- The shrinkage effect is less evident
- The interaction effect varies among affected genes

Marginal posterior



The interaction effect varies among affected genes

- For example we report the posterior densities of γ_{ij} temperature effects for gene=YJL218W (more affected by temperature changes) and gene =YJR029W (less affected)



Discussion

- The inference borrows strenght (too much) from all the observations.
- Relax the exchangeability assumptions through mixture
- Homogeneity variance of gene effects could be unrealistic
- Incorporate background intensities modelling could improve the flexibility of the analysis

The choice of the prior for the gene-strain effects is crucial.

1. A large majority of genes are unaffected (>95%)
2. Affected genes tends to show a two-fold intensity increase, or multiple (due to polyploidia)

These information helped in tuning appropriately the choice of hyperparameters values

Conclusions

- Temperature of hybridization has an effect on the level of absolute intensities of the two strains and on $\log_2$ratios.
- Strain-Gene specific temperature effects are likely for differentially represented genes.
- The observation of a clear temperature effect for the Montalcino strain suggests that microarray technology and our Bayesian approach can be used to assess sequence divergence in individuals of the same species or closely related species.

Acknowledgments

This work has been made possible by CRF-Florence under the project “Center for Genomic Research” and by Italian Ministry of Scientific Research.

References

Kerr, M.K. e Churchill, G.A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2: 183-201.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Computional Biology*, 8(1): 37-52.

Parmigiani, G., Garrett, E.S., Anbazhagan, R. e Gabrielson, E (2002). A statistical framework for expression based molecular classification in cancer. *J. R. Statist. Soc. B.*

Yang, Y.H., Dudoit, S., Luu, P. e Speed, T.P. (2001). Normalization for cDNA microarray data. Berkeley Stat. Tech. Rep. 589.

Chen, Y., Dougherty, E.R., Bittner M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA Microarrays images. *J. Biomed. Optics*, 2(4): 364-374.