Hierarchical Bayesian modelling of multiple array experiments

Annibale Biggeri[1], Emanuela Dreassi[1], Corrado Lagazio[2], Simona Toti[1,3], Carlotta De Filippo[3], Duccio Cavalieri[3,4]

[1] "G. Parenti" Department of Statistics, University of Florence, Italy
[2] Department of Statistical Sciences, University of Udine, Italy
[3] Department of Pharmacology, University of Florence, Italy
[4] Bauer Center for Genomics Research, Harvard University, USA

Corresponding Author:
Annibale Biggeri
"G. Parenti" Department of Statistics
University of Florence, Italy
Viale Morgagni 59, 50134 Florence, Italy
E-mail: abiggeri@ds.unifi.it

Summary

Objectives. A promising application of DNA microarrays relies on detection of changes in copy number following genomic DNA hybridization (array CGH). An open issue in array CGH experiments is to assess the extent of signal variation related to changes in copy number and that is due to differences in the sequence of the experimental sample. This is particularly true of samples whose genomic sequence is not necessarily identical to that used to design the probes printed on the array. A second, related issue regards the consequences of such phenomenon on the inference on relative intensity levels.

Material and methods. A microarray experiment was specifically designed on two genetically different strains of yeast (*Saccharomyces cerevisiae*) varying the hybridization temperatures (50°, 55° and 60°C). Data were analyzed with a hierarchical Bayesian model that jointly takes into account all the slides. The model allowed us to normalize the data and adjust for temperature-related differential hybridization and cross-hybridization.

Results. Bayesian analysis on all the arrays identified 29 spots as significantly less represented in one strain vs the other. Bayesian curve clustering showed two distinct temperature dependent patterns.

Conclusions. The method that we propose is potentially relevant for studies using array CGH in medicine and in population genetics.

KEY WORDS: *CGH microarray, yeast, Hierarchical Bayesian model.*

Introduction

Population genetic studies using array CGH seek to distinguish signal variation due to changes in gene dosage from that due to sequence differences among individuals (genetic polymorphism). Yeasts have provided the first test bed for array CGH studies. cDNA microarrays have been used to detect changes in gene copy number and expression (1). Allelic variation among different strains can be studied using Affymetrix™ (2-4). cDNA arrays were used to assess relatedness of strains from different species of the Saccharomyces sensu stricto complex (5). The authors discussed extensively the problems in interpreting hybridization results using DNA from strains of different yeast species. Finally, open reading frame (ORF) arrays in CGH experiments were proposed to assess the differences in industrial wine strains versus the lab strain (6). Interpretation of data from genomic DNA hybridization on microarrays containing the full set of ORFs (C-DNA arrays) is more complex than interpretation of results from the analysis of RNA expression data or from allelic variation scanning on oligonucleotide arrays.

Our aim was to develop a model to quantify genetic variation using cDNA arrays.

DNA from two different yeast strains, S288c (hereinafter referred to as L) and M28, a wine strain isolated from Montalcino grapes (hereinafter, M), was allowed to hybridize to a microarray representing all the yeast coding regions (7). L is the laboratory strain of yeast whose sequence has been determined according to the *Saccharomyces cerevisiae* genome sequencing project (8). Three different hybridization temperatures (50°, 55°, 60°C) were used. The extent of signal variation between two strains can depend on hybridization temperature, reflecting three distinct issues: differences in gene dosage, differences in the sequence of the experimental sample with respect to that present on the array, and cross-hybridization among highly homologous sequences. Families of duplicated paralogous genes can share homology up to 98%; changes in the copy number of a given gene can be hidden by cross-hybridization of a paralogous gene proportionally to their sequence similarity. Lowering the hybridization temperature increases the effect of cross-hybridization, and we expect this effect to be multiplicative with the sequence divergence effect.

In the present paper we propose a hierarchical Bayesian model which incorporates data normalization and adjusts for temperature-related differential hybridization. Twenty-nine genes were identified as significantly less represented in the natural M strain versus the L strain. Bayesian hierarchical clustering of curves (9) identified two distinct patterns by temperature.

Biological Rationale

ORF array construction

A set of clones containing 6129 verified ORFs of the yeast genome were obtained from Research Genetics (Huntsville, AL) and amplified to the levels required for the preparation of DNA microarrays using PCR (10). Several longer ORFs were amplified using the Gibco BRL Amplification Kit, (Life Technologies, Rockville, MD). Each amplified product was confirmed by agarose gel electrophoresis. Ninety-eight per cent of the ORFs yielded bands of appropriate length. The amplified DNA was precipitated with isopropanol, washed in 70% ethanol, and resuspended in 3x SSC. DNA from two different strains, BY4743, a S288c derivative (L, the laboratory strain whose sequence has been determined according to the *Saccharomyces cerevisiae* genome sequencing project) and a wine strain isolated from Montalcino grapes (M), was extracted in accordance with the protocol of Winzeler et al. (2). The DNA was fragmented in accordance with the protocol of Dunn et al. (6) and allowed to hybridize to a microarray representing all the yeast coding regions, in accordance with the protocol described by Giuntini et al. (11). The overnight incubation was performed at three different temperatures (50°, 55°, 60°C). We chose BY4743 as a reference because its sequence is identical to that of L, and its genome contains a number of well-characterized genetic markers that can be useful in interpreting hybridization results (12). BY4743 contains the alleles ura3delta, leu2delta, his3delta (i.e., an extensive deletion of the URA3 gene, a complete deletion of the LEU2 gene, and a partial deletion of the HIS3 gene that eliminates 25% of the sequence, 187 bp out of 662, respectively). BY4743 is also heterozygous for lys2delta and met15delta. The signal for the URA3 gene or for the LEU2 gene can serve as an indication of absence of a gene from the array: in theory an absent gene should result in no signal in one channel, with an infinite asymptotic log-fold change value. In practice, cross-hybridization reduces this difference, and since the reference strain is diploid, a signal exceeding the 2-fold threshold for unique genes is, in theory, sufficient to indicate a gene deletion, on ORF arrays.

Experimental design

Replicates are produced for each cell of the 2 × 3 (two strains, L and M, and three temperatures, 50°, 55°, 60°) factorial design (13). We first used fluorescent red dye (Cy5) for the M strain and green dye (Cy3) for L strain, and subsequently reversed the colours (dye swap).

The experiment considered 6129 genes, six arrays each printed with 16 different pins (96 different pins per experiment). The extent of signal variation between L and M at different temperatures can reflect three distinct phenomena: changes in copy number (gene dosage), differences between the sequence of the experimental sample and that present on the array, and cross-hybridization among highly homologous sequences.

We expect that the third effect is modulated by hybridization temperature. This information is extremely important in the use of microarrays in medicine and in population genetics.

Indeed, since we considered two strains that show differences in the genomic sequences, differential hybridization was deemed likely. Differential hybridization can result from 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 hybridization 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.

Fluorescent cDNA bound to the microarray was detected with a GENEPIX 4000B microarray scanner (Axon Instruments, Foster City, CA), using the GENEPIX 4000 software package to quantitate fluorescence of the microarray. Fluorescence intensity values were adjusted by subtracting surrounding background from spots. We optimized the array production and scanning to keep this signal as low as possible, usually between 40 and 100 PMTs. To eliminate signals that are most prone to estimation error, a spot was excluded from analysis if both the Cy3 and Cy5 mean fluorescence signals were within two standard deviations of the mean background signals for that spot (14).

Hierarchical Bayesian model

We specified a hierarchical Bayesian model to decompose the different sources of variability and to adjust for temperature-related differential hybridization and cross-hybridization.

The background-adjusted intensities for each gene of the two strains were assumed to be gamma distributed (with common coefficient of variation $1/\sqrt{a}$; see 15). We indicated with L_{ijkp} and M_{ijkp} the intensity for the i-th gene (i = 1,...,6129), the j-th hybridization temperature (j = 1,2,3 corresponding respectively to 50°, 55° and 60°C), the k-th dye (k = 1,2), and the p-th pin (p = 1,...,96), for L and M respectively. Each array is uniquely identified by the combination of temperature and dye indexes.

The model was

$$L_{ijkp} \sim \text{Gamma}\left(\theta^{L}_{ijkp}, a\right)$$

and

$$M_{ijkp} \sim \text{Gamma}\left(\theta^{M}_{ijkp}, a\right)$$

with means

$$E(L_{ijkp}) = \theta^{L}_{ijkp} \quad \text{and} \quad E(M_{ijkp}) = \theta^{M}_{ijkp}$$

and variances

$$\text{Var}(L_{ijkp}) = (\theta^{L}_{ijkp})^2 / a \quad \text{and} \quad \text{Var}(M_{ijkp}) = (\theta^{M}_{ijkp})^2 / a.$$

We defined a log-linear model on the expected intensity values

$$\log_2\left(\theta^{L}_{ijkp}\right) = \mu^{gene}_{i} + \mu^{temp}_{j} + \mu^{dye}_{k} + \mu^{pin}_{(jk)p}$$

and

$$\log_2\left(\theta^{M}_{ijkp}\right) = \mu^{gene}_{i} + \mu^{temp}_{j} + \mu^{dye}_{k} + \mu^{pin}_{(jk)p} + \beta_i \qquad [1]$$

The parameter μ^{gene}_{i} represents the "normalized" mean intensity level for the i-th gene, i.e. the expected value corresponding to the reference (first) category of temperature, dye and pin. The terms μ_j^{temp}, μ_k^{dye} and $\mu^{pin}_{(jk)p}$ represent the effects of temperature, dye and print tip (the notation $\mu^{pin}_{(jk)p}$ denotes that pin effects are nested within jk-th array). The parameter β_i can then be interpreted as "fold-change":

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

Non-informative prior distributions were specified for the parameters of the model: flat normals for the μ terms; a Gamma distribution for a. For the gene strain effects, β_i, we assumed a mixture of normal distributions, with components representing over, under or non-differential expression. Formally, indicating with μ_{β_i} the expected value of β_i, we assumed that

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

where (π^-, π^+, $1–\pi^+$ π^-) ~ Dirichlet (ν_1, v_2, v_3), *TN* indicates the truncated Normal distribution and τ_β^{-1} ~ Gamma (0.001,0.001).

We then extended model [1] adding a new set of parameters to describe strain-specific temperature effects:

$$\log_2\left(\theta^M_{ijkp}\right) = \mu_i^{gene} + \mu_j^{temp} + \mu_k^{dye} + \mu_{(jk)p}^{pin} + \beta_i + \beta_j^{temp} \; [2]$$

Flat normal distributions were specified for β_j^{temp}.

Strain gene-specific temperature effects were investigated defining a third model

$$\log_2\left(\theta^M_{ijkp}\right) = \mu_i^{gene} + \mu_j^{temp} + \mu_k^{dye} + \mu_{(jk)p}^{pin} + \beta_j^{temp} + \beta_{ij}^{gene.temp} \; [3]$$

A mixture prior distribution was used for the expected value $\mu_{\beta_{ij}}$ of the strain gene-specific temperature parameters:

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

considering $\mu_{rij} = \alpha_r + \gamma_{ij}^{temp}$. For μ_{1ij} and μ_{2ij} a non-informative truncated normal distribution was specified. γ_{ij}^{temp} ~ Normal ($0,\tau_\beta$), where τ_β^{-1} ~ Gamma (0.001,0.001).

Inference was based on the full posterior distributions approximated by Monte Carlo Markov Chain (MCMC) simulations (using the WinBUGS 1.4 package, 16). The convergence of the algorithm was evaluated using the Gelman-Rubin test (17) for multiple chains for a subset of the monitored parameters. Convergence was achieved after 10000 simulations and a further 10000 iterations were used for estimation.

With model [3], we identified over- and under-expressed genes and, moreover, estimated temperature expression profiles specific for each gene.

In order to identify genes with similar temperature patterns we further analyzed the estimated profiles using the model-based Bayesian hierarchical clustering algorithm proposed by Heard et al. (9). This method, originally developed to identify co-regulated genes in time-course experiments, is based on non-linear regression splines that describe the signal variation in each cluster. The number of clusters is not fixed a priori, but is assumed as part of the estimation procedure.

Convergence to the posterior distribution was assessed using the Gelman and Rubin test.

Results

A large majority of genes (more than 95%) did not vary between the L and M strains. Variable genes tended to show an intensity increase factor of between 2 and 15.

Bayesian analyses identified 30 spots, 29 diverged or deleted in the M strain and one deleted from the By4743 strain. The latter, as expected, was YEL021W, URA3. Surprisingly, the model did not detect LEU2, YCL018W, also deleted in By4743. This might be explained by the fact that LEU2 overlaps almost completely with a TY2 element as well as with a tRNAleu (transfer RNA for Leucine). The probe for LEU2 printed on the array likely cross-hybridized with the other copies of TY2 of the By4743 genome, which share a sequence identity of approximately 95% (18). The model behaved conservatively. Indeed it failed to detect the partial deletion of the HIS3 gene that is still present at 75%, and the heterozygosity in LYS2 and MET15.

Among these affected genes, it is worth noting the deletion of a region of Chr XII (confirmed by PCR), the reduction of the number of transposable elements, and of the regions proximal to these repetitive elements (Table 1).

Hybridization temperature modifies signal intensities. The temperature effect was more evident for the Montalcino strain (in grey) than for the laboratory strain (in black) (Figure 1). Figure 2 shows $\log_2$ ratio (M/L) at 60° *vs* $\log_2$ ratio (M/L) at 50°; red-labelled Montalcino strain (+); green-labeled Montalcino strain (∘).

Figure 3 shows posterior distributions of strain-specific temperature effect for $j = 60$ by model [2]. The $\log_2$ ratios appeared weakly affected, as shown by the posterior density of β_j^{temp} (Figure 4, for $j = 60$).

Table 1. List of affected genes.

	Gene	log FC_{50}	log FC_{55}	log FC_{60}	Eff55	Eff60	Cluster
	YOL163W	-2.4628	-2.9631	-2.5318	-0.4999	-0.0690	3
	YOL162W	-2.4631	-2.6373	-2.5449	-0.1745	-0.0820	3
+	YMR051C	-2.4643	-3.1378	-3.4305	-0.6748	-0.9672	3
	YMR046C	-2.4631	-2.4509	-2.2924	0.0120	0.1708	3
+	YML045W	-2.4643	-3.2911	-3.5378	-0.8279	-1.0745	3
+	YML040W	-2.4630	-2.7041	-2.5969	-0.2409	-0.1340	3
*	YLR161W	-2.4631	-2.9328	-3.0168	-0.4698	-0.5535	3
*	YLR160C	-2.4648	-3.9512	-3.7200	-1.4880	-1.2567	2
*	YLR159W	-2.4628	-2.7530	-2.7929	-0.2898	-0.3300	3
*	YLR158C	-2.4646	-3.9544	-3.5457	-1.4913	-1.0825	2
*	YLR157C	-2.4647	-3.9632	-3.7753	-1.4998	-1.3122	2
*	YLR156W	-2.4639	-3.6203	-2.5883	-1.1570	-0.1252	3
*	YLR155C	-2.4652	-4.6292	-4.3848	-2.1659	-1.9216	2
	YJR153W	-2.4650	-3.9596	-4.0596	-1.4964	-1.5966	2
+	YJR029W	-2.4606	-2.1289	-2.6585	0.3342	-0.1955	3
+	YJR028W	-2.4637	-3.1226	-2.8960	-0.6595	-0.4328	3
+	YJR026W	-2.4633	-2.7952	-2.9568	-0.3323	-0.4937	3
	YJL218W	-2.4649	-4.0343	-3.7919	-1.5711	-1.3289	2
	YJL217W	-2.4622	-2.5780	-2.4434	-0.1151	0.0197	3
+	YJL114W	-2.4647	-3.8520	-3.6607	-1.3888	-1.1976	2
	YIR042C	-2.4636	-2.9806	-2.8582	-0.5177	-0.3951	3
+	YIL082W	-2.4625	-2.4654	-2.2299	-0.0025	0.2330	3
+	YIL080W	-2.4594	-1.9416	-1.8865	0.5215	0.5762	3
	YIL015C-A	-2.4608	-1.9993	-2.5472	0.4636	-0.0840	3
+	YHR054C	-2.4604	-1.9813	-1.9195	0.4815	0.5435	3
+	YGL053W	-2.4590	-1.7825	-1.9639	0.6802	0.4989	3
+	YBR012W-A	-2.4613	-2.1502	-2.0769	0.3126	0.3857	3
+	YAR031W	-2.4591	-1.9468	-2.2244	0.5161	0.2385	3
+	YAR010C	-2.4628	-2.3400	-2.4173	0.1228	0.0456	3
	YEL021W	2.5028	2.6452	2.7413	0.1424	0.2384	1

Symbols: * = deletion Chr XII (Oligo-PCR); + = transposomes and neighbouring regions.

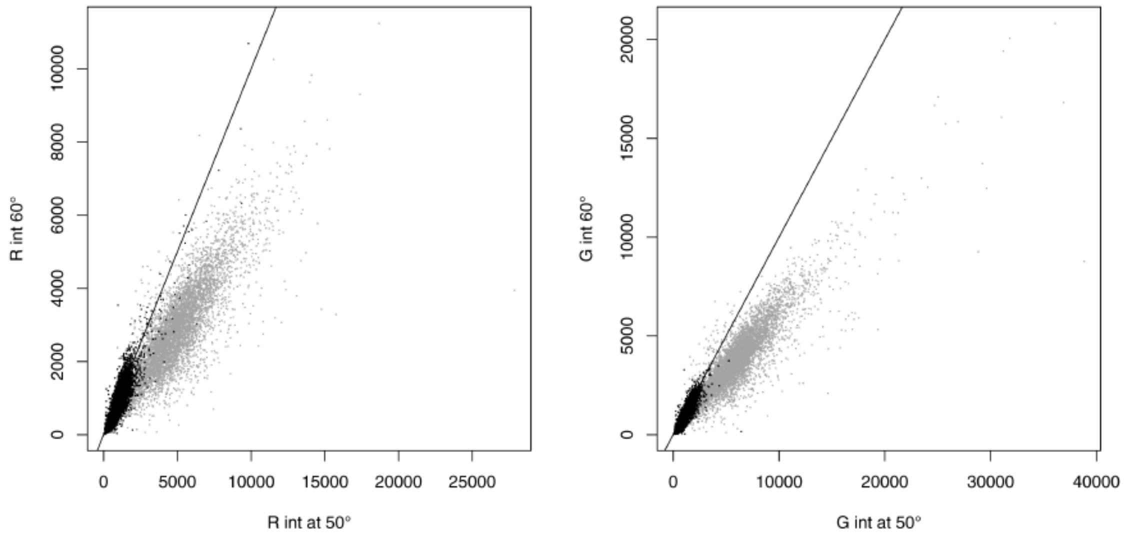

Figure 1. Intensity at 50° *vs* 60°C; M strain (in grey) L strain (in black); Red dye (left), Green dye (right).

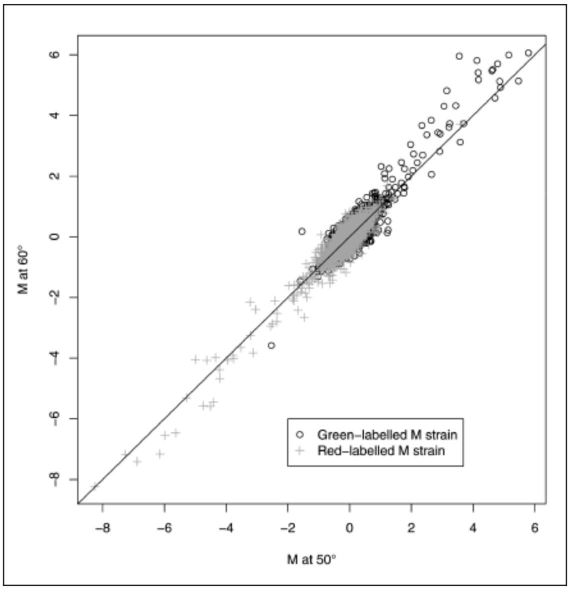

Figure 2. Y-axis: $\log_2$ ratio (M/L) at 60° *vs* X-axis: $\log_2$ ratio (M/L) at 50°.

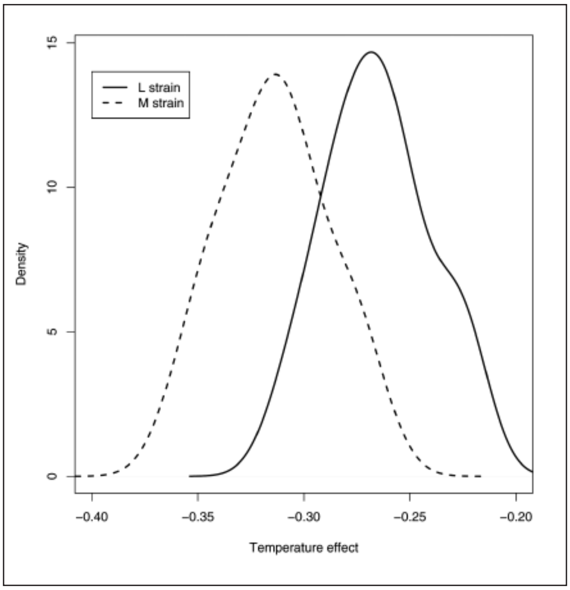

Figure 3. Posterior distribution of strain-specific temperature effect for $j = 60$.

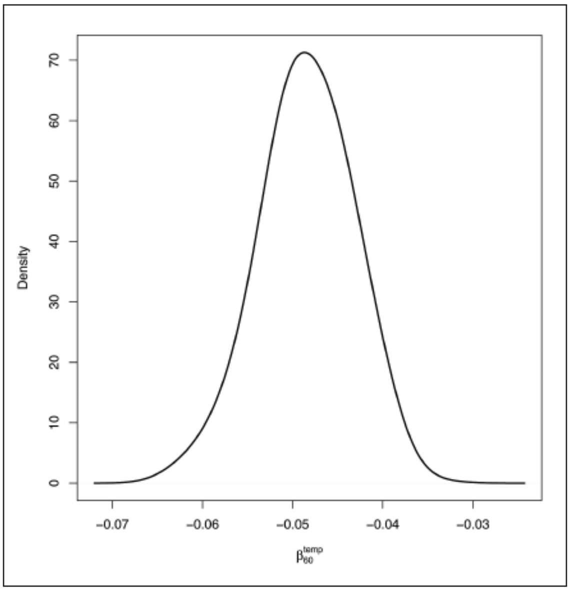

Figure 4. Posterior distribution of β_{60}^{temp}.

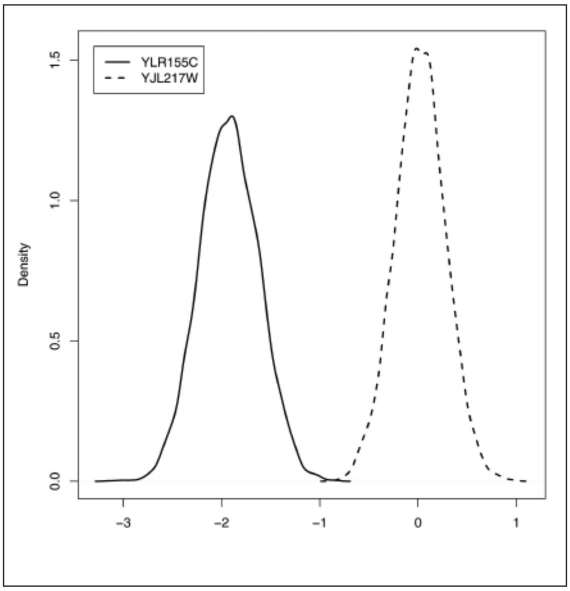

Figure 5. The interaction effect varies among affected genes: posterior densities of temperature effects for gene YLR155C (more affected by temperature changes) and gene YJL217W (less affected).

Strain gene-specific temperature effects estimated by model [3] suggested the presence of interactions for some of the affected genes. For example, in Figure 5, we report the posterior densities of temperature effects for gene YLR155C (more affected by temperature changes) and gene YJL217W (less affected).

Figure 6 shows strain gene effect profiles by temperatures for the 29 diverged or deleted genes of the M strain.

Clustering of the gene effect profiles by temperature clearly identified two patterns (Table 1 and Figure 7). Seven genes showed a strong fold-change increment with increasing temperature. Four of these were the ASP3 cluster of genes on the Crick strand of chro-

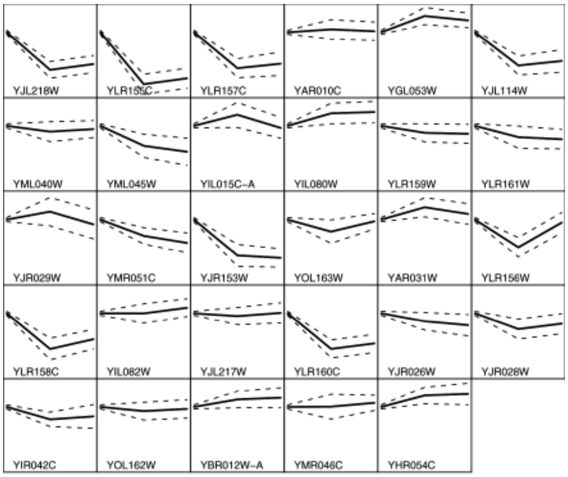

Figure 6. Strain gene effect profiles by temperature.

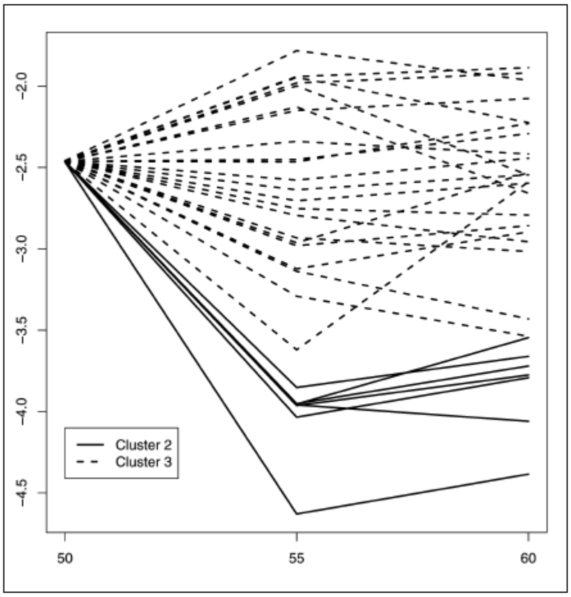

Figure 7. Temperature expression profiles for the under-expressed genes. Continuous and dashed lines indicated different profiles.

mosome XII. The other three elements were a transposable element, TyGAG, a subtelomeric Y element, and PGU1, a polygalacturonase in the ChrX subtelomeric region.

Discussion and conclusions

We proposed a full Bayesian approach to model the effect of temperature on the specificity of the hybridization reaction. Our approach used all the information collected to make inferences about the set of affected genes, borrowing strength from other genes. Several sources of variability were considered, avoiding arbitrary pre-processing of data and including the normalization phase in the modelling phase. Multiple testing was performed adding a third layer to the Bayesian model.

However, this model is computationally heavy and requires careful tuning.

Efforts to improve the model will seek to relax the exchangeability assumptions and the assumptions of homogeneity of variance of gene effects. Generally speaking (although less relevant to our data-set), incorporating background intensities or non-linear effects could improve the flexibility of the analysis.

The choice of the priors for the gene strain effects is crucial. We suggested a finite mixture model (19). This made it possible to estimate the posterior probability of being differentially represented; weakly informative priors could produce a strong shrinking effect.

A large majority of genes were unaffected (> 95%). Affected genes tended to show a two-fold intensity increase, or multiple increase (due to polyploidy). The signal in M28 with respect to By4743 can yield a log ratio of between 2.502 and 2.741 for YEL021W, the URA3 gene (Table 1); this result suggests that the total gene deletion in the absence of a similar gene does not show any cross-reaction, and this log ratio value thus has to be considered a clean indicator of gene deletion. To further test our assumptions, we compared our results with those reported by Winzeler et al. (3) on the comparison between M and L using affy arrays. Compared to the S288c Affymetrix array, 33 of the 35 ORFs that we described as altered had at least two polymorphic sites.

These findings helped us to tune appropriately the choice of hyperparameter values. Hybridization temperature had an effect on the level of absolute intensities of the two strains and on log ratios. Strain gene-specific temperature effects are likely for differentially represented genes.

Prior distributions for precision parameters are obtained from predictive distributions from a calibration experiment (20).

In problems of this kind, sensitivity analysis of prior assumptions is difficult. Indeed, prior distributions are usually tuned by preliminary data analysis and a careful choice of initial values is also important to speed up MCMC algorithm convergence (19, page 12).

Our observation of a clear temperature effect for the Montalcino strain suggests that the absolute intensity decrease in M, but not in L, is linked to a non-perfect match between the probes and the M DNA. This finding demonstrates that microarray technology and our Bayesian approach can also be used to assess sequence divergence in individuals of the same species or of closely related species. In agreement with this explanation, increasing the temperature increases the contrast between the values. The clear division in two patterns resulting from the clustering of gene-temperature profiles indicates that our model allows a precise evaluation of the effects dependent on the presence of multiple copies with high similarity for one element of the genome. The ASP3 results clearly demonstrate what is going on. Our results demonstrate that the M strain is missing 4 ASP3 (YLR155, YLR157, YLR158, YLR160) ORFs on the Crick strand, and a further three small ORFs on the Watson strand (YLR156, YLR159, YLR161). All these ORFs are included in a large deletion of 21636 bp on ChrXII: from coordinates 468959 to 490595 in the M strain with respect to By4743. The ASP3 gene has 90% homology with the ASP1 gene, therefore on decreasing the hybridization temperature, ASP1 from M cross hybridizes with the ASP3 probe, indicating a temperature effect. The three sequences on the Watson strand are unique (YLR156, YLR159, YLR161), and therefore their ratio does not change at lower temperatures. The same is true for the transposable elements of the Ty class and some of the subtelomeric ORFs, which are known to have multiple copies in the genome (3). It is worth noting that our model failed to detect minor changes in copy number as well as heterozygosity. This is due to the decision to set conservative priors, since our main objective was to assess presence-absence of genes. Moreover, our model estimated that the ratios at 50 degrees for the differentially abundant genes are in the order of 2.4. This value is consistent with the a priori theoretical value of 2, as discussed in the introduction. (Setting less conservative priors might allow more subtle effects to be captured, but might also result in an unacceptable rate of false positives).

Acknowledgments

This work was made possible by: Cassa di Risparmio di Firenze, under the project “Center for Genomic Research”; the Italian Ministry of Scientific Research, PRIN 2003133820 and PRIN 2005134079; the European Nutrigenomics Organisation - NuGO Network of Excellence, the European Commission Sixth Framework Programme, grant to Duccio Cavalieri and Annibale Biggeri; DC-THERA, grant to Duccio Cavalieri.

References

1. Lashkari DA, DeRisi LJ, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997; 94: 13057-13062.
2. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW. Direct allelic variation scanning of the yeast genome. Science 1998; 281: 1194-1197.
3. Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, Hartl DL. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. Genetics 2003; 163: 79-89.
4. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE. The core meiotic transcriptome in budding yeasts. Nat Genet 2000; 26: 415-423.
5. Edwards-Ingram, LC, Gent ME, Hoyle DC, Hayes A, Stateva LI, Oliver SG. Comparative genomic hybridization provides new insights into the molecular taxonomy of the Saccharomyces sensu stricto complex. Genome Res 2004; 14: 1043-1051.
6. Dunn B, Levine RP, Sherlock G. Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC Genomics 2005; 6: 53.
7. Cavalieri D, Townsend JP, Hartl DL. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc Natl Acad Sci USA 2000; 97: 12369-12374.
8. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. Science 1996; 274: 546, 563-567.
9. Heard NA, Holmes CC, Stephens DA. A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. Journal of

the American Statistical Association 2006; 101: 18-29.

10. Hardwick JS, Kuruvilla FG, Tong JK, Shamji AF, Schreiber SL. Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. Proc Natl Acad Sci USA 1999; 96: 14866-14870.
11. Giuntini E, Mengoni A, De Filippo C, Cavalieri D, Aubin-Horth N, Landry CR, Becker A, Bazzicalupo M. Large-scale genetic variation of the symbiosis-required megaplasmid pSymA revealed by comparative genomic analysis of Sinorhizobium meliloti natural strains. BMC Genomics 2005; 6: 158.
12. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast 1998; 14: 115-132.
13. Kerr MK, Churchill GA. Experimental design for gene expression microarrays, Biostatistics 2001; 2: 183-201.
14. Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. Trends Genet 2003; 19: 649-659.
15. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA Microarrays images. J Biomed Opt 1997, 2: 364-374.
16. Spiegelhalter DJ, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 1.4. 2002 (On-line user manual, http://www.mrc-bsu.cam.ac.uk/bugs).
17. Gelman A, Rubin DR. Inference from iterative simulation using multiple sequences (with discussion). Stat Sci 1992; 7: 457-511.
18. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res 1998; 8: 464-478.
19. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression based molecular classification in cancer. J R Stat Soc B 2002; 64: 1-20.
20. Blangiardo M, Toti S, Giusti B, Abbate R, Magi A, Poggi F, Rossi L, Torricelli F, Biggeri A. Using a calibration experiment to assess gene-specific information: full Bayesian and empirical Bayesian models for two-channel microarray data. Bioinformatics 2006; 22: 50-57.