BMC Genomics

BioMed Central

Methodology article

Open Access

Between-groups within-gene heterogeneity of residual variances in microarray gene expression data

Joaquim Casellas*[1] and Luis Varona[1,2]

Address: [1]Genètica i Millora Animal, IRTA-Lleida, 25198 Lleida, Spain and [2]Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, 50013 Zaragoza, Spain

Email: Joaquim Casellas* - joaquim.casellas@irta.es; Luis Varona - lvarona@unizar.es

* Corresponding author

Published: 4 July 2008

BMC Genomics 2008, **9**:319 doi:10.1186/1471-2164-9-319

Received: 4 March 2008

Accepted: 4 July 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/319

Abstract

Background: The analysis of microarray gene expression data typically tries to identify differential gene expression patterns in terms of differences of the mathematical expectation between groups of arrays (e.g. treatments or biological conditions). Nevertheless, the differential expression pattern could also be characterized by group-specific dispersion patterns, although little is known about this phenomenon in microarray data. Commonly, a homogeneous gene-specific residual variance is assumed in hierarchical mixed models for gene expression data, although it could result in substantial biases if this assumption is not true.

Results: In this manuscript, a hierarchical mixed model with within-gene heterogeneous residual variances is proposed to analyze gene expression data from non-competitive hybridized microarrays. Moreover, a straightforward Bayes factor is adapted to easily check within-gene (between groups) heterogeneity of residual variances when samples are grouped in two different treatments. This Bayes factor only requires the analysis of the complex model (hierarchical mixed model with between-groups heterogeneous residual variances for all analyzed genes) and gene-specific Bayes factors are provided from the output of a simple Markov chain Monte Carlo sampling.

Conclusion: This statistical development opens new research possibilities within the gene expression framework, where heterogeneity in residual variability could be viewed as an alternative and plausible characterization of differential expression patterns.

Background

Gene expression measured by microarray chips is an emerging and cost-effective tool to assess the expression of thousands of genes in different tissues and organisms [1]. This technology has been intensively used to monitor changes in gene expression between tissues, treatments or time points in order to detect genes, or even metabolic pathways, involved in differential expression patterns [2]. As was highlighted by Wolfinger et al. [3], inference in microarray gene expression analyses is typically focused on gene-specific differences between mathematical expectations of two (or more) groups of biological conditions. However, discrepancies in gene expression could also be characterized by other statistics of interest like dispersion parameters [4,5].

Heterogeneity of residual variances is a topic of main concern in biological studies where residual variance changes

under alternative treatments [6,7]. In gene expression analyses, heterogeneity of gene-specific residual dispersion has been addressed recently [8,9], where hierarchical mixed models with gene-specific residual variances substantially reduced the rate of false positives and allowed for a more realistic fit of gene expression data [10]. Nevertheless, a common within-gene residual variance was assumed in these analyses, although within-gene discrepancies in the dispersion parameters could also be feasible. To our best knowledge, discrepancies in terms of gene-specific residual variance across different biological conditions (groups or arrays) have never been considered in the microarray literature. Besides a plausible scale effect on the residual variance due to changes in mathematical expectation under different groups of microarrays, within-gene heterogeneity of the residual variance could suggest a group-specific pattern of variability at the transcription level. Variability could be just due to within tissue variability in cell type composition, but may or may not be related to any meaningful difference in transcription.

The aim of this research is to propose a hierarchical mixed model analysis of microarray gene expression data assuming within-gene heterogeneous residual variances. In addition, a straightforward Bayes factor approach to test differences between two within-gene residual variances is developed, taking Verdinelli and Wasserman [11] and Varona et al. [12] as starting point. This methodology could open a new research field in gene expression analysis where differential gene expression will be characterized in terms of variability of the transcription process.

Methods

Hierarchical mixed model with within-gene heterogeneous residual variances

Assume as starting point n replicates of non-competitive hybridization microarray data with m genes (or probes; each probe is a fragment of complementary nucleic acid covering genomic or inter-genomic annotated regions) per array. Under the simplest design, these replicates are distributed in two different groups of treatments (e.g. tissues, species or time points) with r and s replicates per treatment, respectively ($r + s = n$). This gene expression data can be analyzed under the following hierarchical mixed model [13],

$$\mathbf{y} = \mathbf{Xa} + \mathbf{Z}_1\mathbf{p}_1 + \mathbf{Z}_2\mathbf{p}_2 + \mathbf{e},$$

where $\mathbf{y}$ is the $(nm) \times 1$ column vector of intensity scores sorted by array within treatment within gene and $\mathbf{e}$ is the $(nm) \times 1$ column vector of residuals. Effects in model were array ($\mathbf{a}$; dimension $n \times 1$) and probe ($\mathbf{p}_1$ and $\mathbf{p}_2$; dimension $m \times 1$) linked to $\mathbf{y}$ by appropriate incidence matrices ($\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively). Vector $\mathbf{e}$ is assumed to be normally distributed [14],

$$\mathbf{e} \sim N(0, \mathbf{R}),$$

$\mathbf{R}$ being the matrix of residual (co)variances. Assuming null residual (co)variances [8,9,13] and heterogeneous gene-specific residual variances between treatments, $\mathbf{R}$ can be stated as

$$\mathbf{R} = \bigoplus_{i=1}^{m} \begin{bmatrix} \mathbf{I}_1\sigma^2_{e(i1)} & 0 \\ 0' & \mathbf{I}_2\sigma^2_{e(i2)} \end{bmatrix},$$

where $\mathbf{I}_1$ is a $r \times r$ identity matrix, $\mathbf{I}_2$ is a $s \times s$ identity matrix, $\mathbf{0}$ is a $r \times s$ matrix of zeros, and $\sigma^2_{e(ij)}$ is the residual variance for the ith gene and jth treatment. Under a standard Bayesian development, the joint posterior probability of all unknowns in model is proportional to

$$p\left(\mathbf{a}, \mathbf{p}_1, \mathbf{p}_2, \mathbf{R}, \sigma^2_{p1}, \sigma^2_{p2}, |\mathbf{y}\right) \propto p\left(\mathbf{y}|\mathbf{a}, \mathbf{p}, \mathbf{d}(\mathbf{p}), \mathbf{R}\right) p(\mathbf{a}) p\left(\mathbf{p}_1|\sigma^2_{p1}\right) p\left(\sigma^2_{p1}\right) \times p\left(\mathbf{p}_2|\sigma^2_{p2}\right) p\left(\sigma^2_{p2}\right) p(\mathbf{R}),$$

with a flat prior for $\mathbf{a}$ and multivariate normal *a priori* distributions for $\mathbf{y}$, $\mathbf{p}_1$ and $\mathbf{p}_2$ [13],

$$p\left(\mathbf{y}|\mathbf{a}, \mathbf{p}_1, \mathbf{p}_2, \mathbf{R}\right) \sim N(\mathbf{Xa} + \mathbf{Z}_1\mathbf{p}_1 + \mathbf{Z}_2\mathbf{p}_2, \mathbf{R}),$$
$$p\left(\mathbf{p}_1|\sigma^2_{p1}\right) \sim N\left(0, \mathbf{I}_m\sigma^2_{p1}\right),$$

and

$$p\left(\mathbf{p}_2|\sigma^2_{p2}\right) \sim N\left(0, \mathbf{I}_m\sigma^2_{p2}\right),$$

where $\mathbf{I}_m$ is an $m \times m$ identity matrix, and σ^2_{p1} and σ^2_{p2} are the variance components for $\mathbf{p}_1$ and $\mathbf{p}_2$, respectively. Additionally, inverted χ^2 priors with hyperparameters S^2 and ν are assumed for variance components,

$$p\left(\sigma^2_{p1}\right) \sim \chi^{-2}_{S^2_{p1}, \nu_{p1}},$$
$$p\left(\sigma^2_{p2}\right) \sim \chi^{-2}_{S^2_{p2}, \nu_{p2}},$$

and

$$p(\mathbf{R}) \sim \prod_{i=1}^{m}\prod_{j=1}^{2} \chi^{-2}_{S^2_{e(ij)}, \nu_{e(ij)}}.$$

All the unknowns in model can be sampled under a Markov chain Monte Carlo framework by standard Gibbs sampling [15].

Bayes factor to test within-gene heterogeneous residual variances

When gene expression data is grouped in two different treatments or groups the Verdinelli and Wasserman's [11] approach to Bayes factor can be easily adapted. In order to allow for a straightforward comparison between $\sigma^2_{e(i1)} = \sigma^2_{e(i2)}$ and $s^2_{e(i1)} \neq s^2_{e(i2)}$ hypothesis and without loss of generality, $\mathtt{R} = \bigoplus_{i=1}^{m} \begin{bmatrix} \mathtt{I}_1\sigma^2_{e(i1)} & 0 \\ 0' & \mathtt{I}_2\sigma^2_{e(i2)} \end{bmatrix}$ can be redefined as

$$\mathtt{R}^* = \bigoplus_{i=1}^{m} \begin{bmatrix} \mathtt{I}_1\sigma^2_{e(i)}\pi_i & 0 \\ 0' & \mathtt{I}_2\sigma^2_{e(i)}(1-\pi_i) \end{bmatrix},$$

and consequently, $\sigma^2_{e(i)} = \sigma^2_{e(i1)} + \sigma^2_{e(i2)}$ and $\pi_i = \sigma^2_{e(i1)} \big/ \sigma^2_{e(i)}$. Note π_i can be viewed as a variance heterogeneity parameter where π_i = 0.5 accounts for equal residual variances between treatments and $\pi_i \neq 0.5$ suggests within-gene (between treatments) heterogeneity of residual variances. Assuming $\boldsymbol{\sigma}' = \begin{bmatrix} \sigma^2_{e(1)} & \sigma^2_{e(2)} & \cdots & \sigma^2_{e(m)} \end{bmatrix}$ and σ' = [π_1 π_2 ... π_m], this reparameterization can also be developed within a Bayesian frame, with the following joint posterior probability,

$$p(\mathtt{a}, \mathtt{p}_1, \mathtt{p}_2, \boldsymbol{\sigma}, \boldsymbol{\pi}, \sigma^2_{p1}, \sigma^2_{p2}, |\mathtt{y}) \propto p(\mathtt{y}|\mathtt{a}, \mathtt{p}_1, \mathtt{p}_2, \mathtt{R}^*)p(\mathtt{a})p(\mathtt{p}_1|\sigma^2_{p1})p(\sigma^2_{p1}) \times p(\mathtt{p}_2|\sigma^2_{p2})p(\sigma^2_{p2})p(\boldsymbol{\sigma})p(\boldsymbol{\pi}),$$

and Bayesian likelihood,

$$p(\mathbf{y}|\mathbf{a}, \mathbf{p}_1, \mathbf{p}_2, \mathbf{R}^*) \sim N(\mathbf{Xa} + \mathbf{Z}_1\mathbf{p}_1 + \mathbf{Z}_2\mathbf{p}_2, \mathbf{R}^*).$$

We assume the same prior distributions for **a**, $\mathbf{p}_1$, $\mathbf{p}_2$, σ^2_{p1} and σ^2_{p2} as in previous model, and a scaled inverted χ^2 prior for elements in σ

$$p(\boldsymbol{\sigma}) \sim \prod_{i=1}^{m} \chi^{-2}_{S^2_{e(i)}, \nu_{e(i)}}.$$

Note that this parameterization allows for a gene-specific definition of hyperparameters $S^2_{e(i)}$ and $\nu_{e(i)}$, modifying the shape of the inverted scaled χ^2 prior accordingly to our a priori knowledge about the dispersion patter of each gene. Nevertheless, if we lack of a priori information about gene-specific dispersion patterns, this prior could be reduced to a proper flat distribution with appropriate bound. The priori distribution for π is stated as flat between appropriate bounds,

$$p(\boldsymbol{\pi}) \sim \prod_{i=1}^{m} 1 \quad \text{if } \pi_i \in [0, 1] \text{ and 0 otherwise.}$$

Note that this prior distribution is the key point for the further calculation of the Bayes factor and covers all possible values taken by π_i with equal probability, following Verdinelli and Wasserman [11] and Varona et al. [12]. As in previous model parameterization, all unknowns can be updated by Gibbs sampling [15] with the exception of π_i that requires a Metropolis-Hastings step [16].

For a given gene, model comparison between $s^2_{e(i1)} \neq s^2_{e(i2)}$ and $\sigma^2_{e(i1)} = \sigma^2_{e(i2)}$ hypotheses simplifies to conditions $\pi_i \neq 0.5$ (within-gene heterogeneous residual variances for all genes; Model HE) and π_i = 0.5 (homogeneous residual variance for the *i*th gene, within-gene heterogeneous residual variances for the remaining genes; Model HO_i). Note that π_i is assumed known and fixed in Model HO_i and then, Model HO_i and Model HE are nested models that only differ in a bounded variable (π_i). It is important to highlight that this Bayes factor testes gene-by-gene dispersion patterns, although it does not inform us about the best analytical model for the joint inference of all genes. Following the methodology developed by Verdinelli and Wasserman [11], the Bayes factor between Model HE and Model HO_i (BF_{HE/HO_i}) can be easily calculated from the Markov chain Monte Carlo sampler output of Model HE, by averaging the full conditional densities of each cycle at π_i = 0.5 using the Rao-Blackwell argument [17]. Following García-Cortés et al. [18] and Varona et al. [12], the posterior density $p(\pi_i = 0.5|\mathbf{y})$ suffices to obtain BF_{HE/HO_i},

$$BF_{HE/HO_i} = \frac{p(\pi_i = 0.5)}{p(\pi_i = 0.5|\mathtt{y})} = \frac{1}{p(\pi_i = 0.5|\mathtt{y})},$$

because $p(\pi_i = 0.5)$ was previously defined with the *a priori* distribution of π_i. On the basis of GEAMM v.1.4 program [13], the Bayes factor developed above was implemented

with FORTRAN90 language. All the subsequent analyses were performed with this software.

Example 1. Simulated data

Our Bayes factor approach was tested on simulated data sets under three different scenarios in order to check its statistical performance. For each scenario, a total of 100 data sets were generated, each one including 40 arrays (unrelated individuals), 10,000 genes per array and two groups of treatments (A and B). More specifically, scenario 1 (S1) assigned 20 arrays to each treatment without missing data, scenario 2 (S2) assumed an unbalanced design with 10 and 30 arrays for treatments A and B, respectively (no missing data), and scenario 3 (S3) assumed two groups of 20 arrays with a 5% of randomly distributed missing data. Intensity scores (y_{ijk}) were simulated under the following model,

$$y_{ijk} = \mu + a_i + g_j = e_{ijk},$$

were μ was the overall mean arbitrarily fitted to 6, a_i was the effect of each array sampled from an uniform distribution between 0 and 1, g_j was the effect of the gene sampled from a Gaussian distribution with mean 0 and variance 1, and e_{ijk} was the residual term obtained from a Gaussian distribution with mean zero and variance $0.1 \times \pi_i$ (Group A) or $0.1 \times (1 - \pi_i)$ (Group B). We assumed a unique (and plausible) value for the overall residual variance in order to allow for a direct comparison and interpretations of the results. Genes were grouped in five levels with different values of π_i: 0.5 (6,000 genes), 0.4 (1,000 genes), 0.3 (1,000 genes), 0.2 (1,000 genes) and 0.1 (1,000 genes). Statistical performance of the developed BF to check differential expression in terms of dispersion pattern (residual variances) was compared with a well-known standard *F*-test. The effect of within-gene differential expression was not considered in order to allow for a straightforward comparison between Group A and Group B residual variances under *F*-test. Indeed, additional sources of variation were avoided in order to allow for a direct calculation of the *F*-test without requiring preliminary pre-correction of the data. Each data set was analyzed by the Bayes factor approach described above with the scaled χ^{-2} prior distribution for variances components generalized to proper flat priors ($S^2 = 0$ and $\nu = -2$) defined between > 0 and 1000. A unique Monte Carlo Markov chain with 100,000 elements was launched for each data set, after discarding the first 10,000 iterations as burn-in. Convergence was checked by the Raftery and Lewis [19] algorithm.

Example 2. Free-access gene expression data

To illustrate the methodology described above, we applied the model to free-access fibroblast gene expression data from 10 chimps and 11 gorillas (available at Gene Expression Omnibus [20], accession number GDS340). As described Karaman et al. [21], hybridization was performed in the Human Genome U95 Set platform (Affymetrix, Santa Clara, CA). A rough normalization was performed on the original data set by calculating multiplicative scaling factors on the basis of the median intensity of the 60th and 95th quintile of gene-expression scores [21]. All gene-expression scores below 100 were set to 100 in the original data set (untransformed scale) and. Within this context, all genes with one or more scores equal to 100 were removed from the final analysis. After editing, data set included gene expression scores of 3,700 genes, transformed by a base-2-logarithm as suggested Yeung et al. [22]. The analytical process followed the same specifications as for simulated data sets.

Results

Example 1. Simulated data

As can be seen in Table 1, differences between simulated and predicted values of π were small, suggesting a reasonable model adjustment to gene expression data. Indeed, the average posterior mean for the residual variance was 0.102 (the empirical standard error across-genes and replicates was 0.002; S1) and agreed with the value used in simulations (0.1). When gene expression data was generated under equal residual variances across groups (π = 0.5), the Bayes factor (BF_{HE/HO_i}) discarded heterogeneous variances in the greater part of the cases (S1: 88% to 98%; S2: 90% to 98%; S3: 90% to 97%). Under S1 and following Jeffreys' [23] scale of evidence, between 1% and 12% of genes reached vague evidences of heterogeneous variances and only between 1% and 3% of genes showed substantial evidences of heterogeneous variances (Table 1). Results under S2 (unbalanced design) and S3 (missing data) provided a similar trend with and expectable power loss (Table 1). Although a small percentage of genes supported the existence of heterogeneous variances, these results do not invalidate our Bayes factor approach, given that a substantial increase in false positives is expected when the number of replicates (arrays) per analyses is small, a typical phenomenon in microarray data sets. Moreover, these results agreed with the ones obtained by a standard *F*-test, where a 1–12% of genes (across data sets and simulation scenarios) reached *p*-values lower than 0.05.

As was expected, BF_{HE/HO_i} showed an overall increase when π values used in the simulation process decreased (Table 1). The percentage of $BF_{HE/HO_i} < 1$ decreased with π, it ranging between 60% and 84% (π = 0.4), between

Table 1: Simulated (π) and predicted ($\tilde{\pi}$; average of the posterior mean across-genes and replicates) heterogeneity and percentage of genes falling within each category of the Bayes Factor for the three simulation scenarios

π	$\tilde{\pi}$ [1]	BF_{HE/HO_i} < 1	1 ≥ BF_{HE/HO_i} < 3.16	3.16 ≥ BF_{HE/HO_i} < 10	10 ≥ BF_{HE/HO_i} < 31.62	31.62 ≥ BF_{HE/HO_i} < 100	BF_{HE/HO_i} ≥ 100
Simulation scenario 1							
0.5	0.498	88 to 98	1 to 12	1 to 3	0	0	0
0.4	0.412	60 to 84	2 to 28	1 to 16	0 to 8	0 to 4	0 to 4
0.3	0.301	20 to 48	16 to 36	12 to 28	4 to 12	0 to 12	0 to 16
0.2	0.196	4 to 16	8 to 32	4 to 28	4 to 24	8 to 28	16 to 36
0.1	0.121	0	0 to 4	0 to 4	0 to 8	0 to 8	84 to 100
Simulation scenario 2							
0.5	0.492	90 to 98	2 to 11	0 to 2	0 to 1	0	0
0.4	0.407	64 to 89	1 to 22	0 to 11	0 to 7	0 to 2	0 to 1
0.3	0.296	23 to 51	11 to 27	9 to 22	2 to 9	0 to 8	0 to 7
0.2	0.201	10 to 21	7 to 30	2 to 25	2 to 25	3 to 27	9 to 31
0.1	0.115	0 to 1	0 to 8	0 to 10	0 to 12	0 to 17	76 to 100
Simulation scenario 3							
0.5	0.508	90 to 97	1 to 11	1 to 4	0	0	0
0.4	0.409	62 to 87	1 to 26	0 to 14	0 to 6	0 to 4	0 to 3
0.3	0.308	23 to 51	15 to 34	10 to 25	3 to 10	0 to 9	0 to 10
0.2	0.211	5 to 17	7 to 31	4 to 26	4 to 23	7 to 26	14 to 31
0.1	0.119	0 to 1	0 to 8	0 to 8	0 to 10	0 to 11	77 to 100

[1]Empirical standard errors were smaller than 0.01 in all cases.

20% and 48% (π = 0.3), between 4% and 16% (π = 0.2) and 0% (π = 0.1). Additionally, evidences favoring the presence of heterogeneous variances increased when gene expression data were simulated under small π, 84% to 100% of the genes reaching $BF_{HE/HO_i} \geq 100$ for π = 0.1 simulated genes (decisive evidence according to Jeffreys's [23] scale). This increase in BF_{HE/HO_i} when the bounded variable (π) departed from the "null hypothesis" (π = 0.5) agrees with previous results published by García-Cortés et al. [18] and Casellas et al. [24] with the same Bayes factor approach although adapted to test heritability of linear and threshold traits.

As can be seen in Figure 1, our Bayes factor and the standard *F*-test performed similarly, in contrast to the noticeable computational instability of previous approximations to the Bayes factor [25]. Nevertheless, the approximation adapted in this manuscript has been previously compared with other statistics of reference like likelihood ratio test [12] or the deviance information criterion [13] developed by Spiegelhalter et al. [26], and showed a very similar performance without detecting remarkable deviations. The strong similarity between the proposed Bayes factor and the standard *F*-test could the viewed as a critical advantage for the *F*-test under a very simple microarray design with two different treatments. When additional sources of variation are included in model, the proposed Bayes factor takes advantage of the joint analysis for all the parameters in the model and simultaneous testing for discrepancies between residual variances of interest. Within this scenario, the *F*-test requires a previous pre-correction for additional sources of variation in the model and therefore, implies a two-steps analysis.

Example 2. Free-access gene expression data

Results are shown in Table 2, where 67.9% of genes did not reveal evidence of within-gene heterogeneous residual variances, and 20.1% of genes suggested almost discernable deviations from Model HO_i ($1 \geq BF_{HE/HO_i} < 3.16$). It is graphically illustrated in Figure 2 where most of the estimated π values were accumulated around 0.5, the value characterizing within-gene homogeneous residual variances. Nevertheless, substantial (8.2% of genes), strong (2.2%), very strong (1.1%) and decisive evidences (0.6%) of within-gene heterogeneous residual variances following Jeffreys' [23] scale were detected (Table 2; Figure 3) in this free-access data set.

As was expected, the across-genes average π values (transformed to 1 - π when π was greater than 0.5) was maxi-

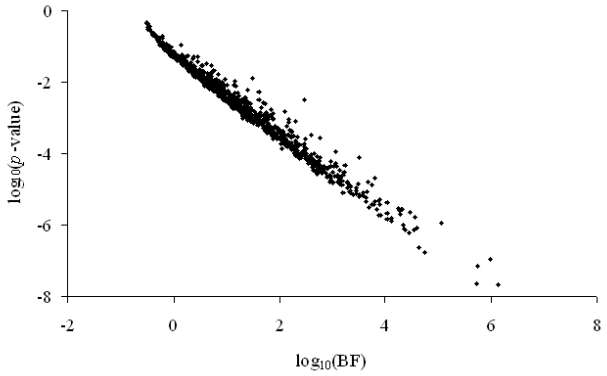

Figure 1
Plot of $\log_{10}(p\text{-value})$ against $\log_{10}(BF)$ for residual variance comparison in the first simulated data set of S1.

mum for genes with $BF_{HE/HO_i} < 1$ (0.431), whereas this parameter reduced to around 0.25 when BF_{HE/HO_i} increased (Table 2). Note that in our analyses, extreme values of π values (< 0.1 or > 0.9) were scarce (Figure 2). For the joint residual variance ($\sigma^2_{e(i)}$), averages ranged between 0.241 and 0.744 (Table 2).

Discussion

Within-gene heterogeneity of the residual dispersion patter in real data

Results obtained in the comparison between chimp and gorilla gene expression data suggested a substantial incidence of within-gene heterogeneity, which is not typically accounted for in standard gene expression analyses. Moreover, detection of relevant (or significant) genes was substantially affected by the analytical model, as is illustrated in Table 3. Model HE showed a more conservative pattern and, when BF_{HE/HO_i} took greater-than-one values, this phenomenon suggested that the rate of false positives increased if within-gene heterogeneity of residual variances was not accounted for in the model [10,27]. A moderate percentage of genes with heterogeneous residual variances did not show differences in terms of mathematical expectation (Table 3), therefore discarding a scale effect. Although these results can not be directly extrapolated to all microarray data sets, these results suggests that heterogeneous residual patterns could be a biological phenomenon of special interest in further analysis of gene expression data. Variability could be just due to within tissue variability in cell type composition, but may or may not be related to any meaningful difference in transcription.

Bayes factor to compare dispersion patterns in microarray studies

Although gene expression analyses have been typically focused on the comparison between mathematical expectations of two or more (within-gene) groups of arrays, the analytical approach developed in the present paper allow for an alternative characterization of differential expression patterns. Moreover, it allows for an appropriate data modeling when within-gene heterogeneity of residual variances exists. This approach could be viewed as statistically inefficient for those genes with homogeneity of residual variances [28,29]. Nevertheless, the aim of this

Table 2: Distribution of genes according to $BF_{HE/Hoi}$, and across-genes average estimates (and empirical standard error across average estimates) for the heterogeneity parameter and residual variance

BF_{HE/HO_i}	Genes	π^{*1}	$\sigma^2_{e(i)}$
$BF_{HE/HO_i} < 1$	2,511[2] (67.9)[3]	0.431 (0.001)	0.244 (0.005)
$1 \geq BF_{HE/HO_i} < 3.16$	743 (20.1)	0.316 (0.001)	0.243 (0.008)
$3.16 \geq BF_{HE/HO_i} < 10$	302 (8.2)	0.249 (0.001)	0.241 (0.013)
$10 \geq BF_{HE/HO_i} < 31.62$	80 (2.2)	0.262 (0.022)	0.725 (0.021)
$31.62 \geq BF_{HE/HO_i} < 100$	41 (1.1)	0.275 (0.030)	0.665 (0.050)
$BF_{HE/HO_i} \geq 100$	23 (0.6)	0.278 (0.042)	0.744 (0.061)

[1]Values were transformed to 1 - π when π was grater than 0.5.
[2]Number of genes.
[3]Percentage related to the overall number of genes analyzed.

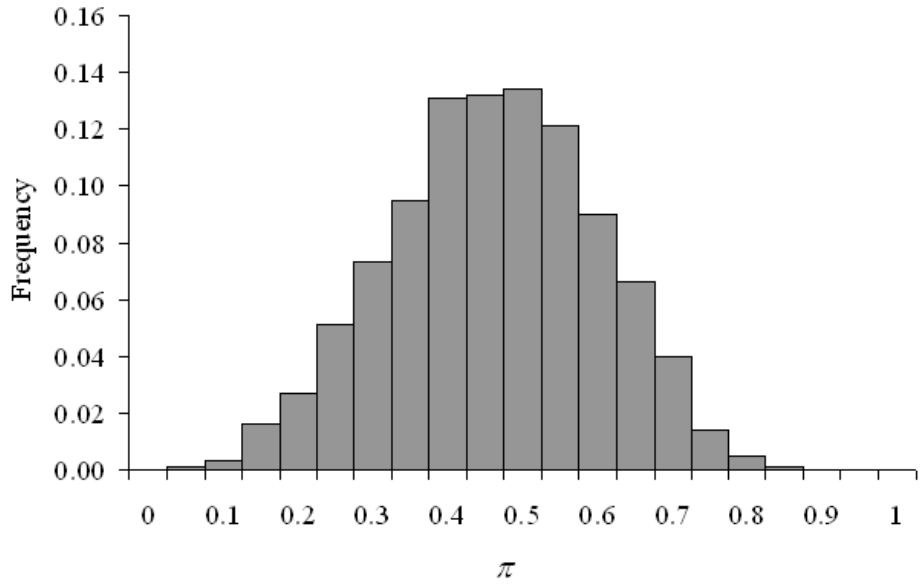

Figure 2
Distribution of π values for gene expression analysis of fibroblast data between chimps and gorillas.

research was to provide an accurate method to compare dispersion patterns, whereas differences between the mathematical expectances of groups of treatments are not of interest in this case. This test could also be applied to experiments with less replicates per group although its results must be taken with caution, given the inherent loss of robustness under small data sets,. As is shown in Figure 1, our Bayes factor performed similarly to the standard *F*-test, with a stable a coherent behavior under moderate

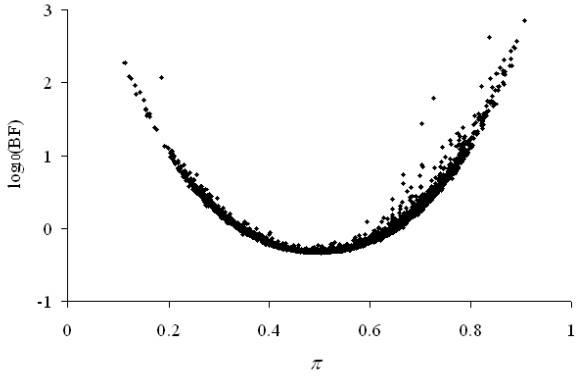

Figure 3
Plot of $\log_{10}$(BF) against π (posterior mean) for the analysis of gene expression data between chimps and gorillas.

sample sizes (number or arrays per group). Although the Bayes factor approach has been described under a simple scenario (simulated datasets), this can be easily generalized to more complex frameworks without additional requirements. Within this context, across-genes shrinkage of residual variances is a topic of main interest in microarray research [29,30] which can be easily adapted to the hierarchical mixed model applied above. Indeed, several Bayesian methods proposed for residual variances shrinkage [31,32] can be applied to both residual variances and heterogeneity parameters, and the calculation of the Bayes factor does not substantially change within-gene or within a group of genes. In a similar way, other approaches can also be jointly implemented with the developed Bayes factor such as mixtures of distributions [33-35]. If several sources of variation are expected on the residual term, the mixed model with within-gene heterogeneous residual variances could be viewed as a useful tool to characterize their distribution pattern, the Bayes factor being a straightforward way to check their statistical relevance. Within this context, our Bayes factor procedure could provide preliminary results required for the application of more complex and computational demanding approaches like the mixed model with log-transformed hierarchical residual variances developed by Jaffrezic et al. [36].

Changes in residual dispersion patterns could be due to a scale effect when mathematical expectations of two (or

Table 3: Distribution of relevant genes according to $BF_{HE/Hoi}$ and under two different analytical models

	Posterior probability[1] < 0.05		Posterior probability[1] < 10^{-5}	
BF_{HE/HO_i}	Model HE	Model HO[2]	Model HE	Model HO[2]
$BF_{HE/HO_i} < 1$	995	1058	210	240
$1 \geq BF_{HE/HO_i} < 3.16$	337	349	73	79
$3.16 \geq BF_{HE/HO_i} < 10$	161	168	31	33
$10 \geq BF_{HE/HO_i} < 31.62$	45	46	10	10
$31.62 \geq BF_{HE/HO_i} < 100$	28	32	2	4
$BF_{HE/HO_i} \geq 100$	8	9	3	4

[1]Posterior probability above (greater-than-zero average difference) of below zero (lower-than-zero average difference) of the difference between chip *vs*. gorilla mathematical expectation.
[2]All π_i were arbitrarily fitted to 0.5.

more) groups of arrays are different. Nevertheless, this scale-related hypothesis was only attributable to a small percentage of genes with heterogeneous residual variances (Tables 3), whereas more than 75% of differential dispersion patters must be related to other unknown causes in the analyzed free-access microarray data. These changes in the dispersion pattern were previously suggested in genes involved in cancer pathogenesis [4,37] and other diseases [38], although within-gene residual heterogeneity is not commonly considered in gene expression analyses [8,9]. Moreover, heterogeneity in gene expression increases with age [5] and therefore, our approach could be of special interest in time-series analyses where individuals at different ages are compared. As a whole, the hierarchical mixed model with within-gene heterogeneous residual variances allows for a new and more accurate modeling of gene expression data with appealing perspectives, and the Bayes factor developed is an easy way to check differences between within-gene residual variances.

Under the Bayesian framework, model comparison is usually made by calculating Bayes factors [39], the ratio between the marginal probabilities of the data given the tested models and after integrating out all parameters in the models. The Bayes factor developed by Verdinelli and Wasserman [11] from generalization of the Savage-Dickey density ratio, and adapted to the animal breeding context by García-Cortés et al. [18] and Varona et al. [12], has been easily applied to check heterogeneous residual variances in gene expression analyses when two groups of treatments are compared. It provides a rigorous and clear framework to compare competing models, avoiding the calculation of significance levels and without depending upon asymptotic properties of frequentist estimators [40], Bayes factor behaves well even when the bounded variable to be tested is either close to the boundary of the parametric space [18]. In addition, Bayes factor provides a ratio of probabilities between models, without any requirement to define the null or the alternative hypothesis, without trying to discard the null hypothesis in favor of an alternative hypothesis, and without referring to the asymptotic properties of the frequentist estimators [12].

Although other Bayes factor approaches could be used, the Verdinelli and Wasserman's [11] approach allows for a simplified calculation, where only the analysis of the complex model is necessary. Moreover, a unique analysis is required to calculate all the gene-specific Bayes factors, and chain storage is not needed because only the (within-gene) average of the full conditional densities at $\pi_i = 0.5$ is used during calculation. Under alternative Bayes factor approaches [39], an additional model with $\pi = 0.5$ for the gene to be tested (and sampling π for the remaining genes) must be analyzed for each gene, in order to obtain the gene-specific Bayes factor comparing heterogeneous *versus* homogeneous residual variances.

Given the null a priori information about the expected distribution of π, we assumed a flat prior distribution between 0 and 1 in order to give the same a priori probability to all plausible values. This is a standard assumption for the Verdinelli and Wasserman's Bayes factor [11,12], although other prior distributions could also become reasonable. It is important to note that $p(\pi)$ equally influences both $p(\pi_i = 0.5)$ and $p(\pi_i = 0.5|\mathbf{y})$ terms and therefore, the Bayes factor must be relatively robust to prior modifications. In the light of the results obtained from the analysis of great ape gene expression data, a pri-

ori distributions favoring values close to 0.5 and with decreasing probability in their tails could be realistic. Within this context, Gaussian, Laplace and Student's *t* distributions with mean 0.5 and truncated at 0 and 1 could be useful a priori distributions, among others. Nevertheless, further studies are required to confirm this pattern in real gene expression data.

As was recently demonstrated at the gene-specific level [10], an accurate modeling of residual dispersion allows for a more realistic fit of gene expression data. Moreover, it has a relevant impact on the rate of false positives when gene expression is characterized in terms of mathematical expectations or their differences [8-10]. In this manuscript, we have adapted Lo and Gottardo [10] mixed model to account for within gene heterogeneity of residual variances, where a relevant incidence of within-gene heterogeneity has been revealed in real gene expression data. Moreover, this heterogeneity can be easily checked gene-by-gene by applying a straightforward Bayes factor with a minimal increase in computational requirements. Note that differences between average gene expression without assuming equal residual variances is a typical example of the Behrens-Fisher problem [41], which could be easily by-passed in microarray analyses by appropriately adapting Welch's [42]*t*-test. Nevertheless, our approach seeks a novel point of view were, not only differences between mathematical means are tested but differences between residual dispersion patterns must also be checked and characterized. In addition, our Bayes factor allows to detect those genes with heterogeneous residual variances where Behrens-Fisher problem [41] holds.

Conclusion

Accounting for within-gene between-groups heterogeneous residual variances in mixed model analyses of non-competitive microarray gene expression data (or even competitive microarray gene expression data after suitable data editing) allows to characterize differential expression patterns in terms of gene expression variability. The Bayes factor approach here presented provides a straightforward comparison between within-gene group-specific residual variances with minimal computing requirements. This methodology is freely available in GEAMM v.1.7 software [43].

Authors' contributions

The methodological approach described in this manuscript was developed by both JC and LV. In addition, JC implemented the analytical model and simulated data sets in FORTRAN90 language, evaluated both real and simulated data sets and drafted the manuscript. LV and contributed to the interpretation of the analyses results and manuscript preparation. Both authors read and approved the final manuscript.

Acknowledgements

The research contract of J. Casellas was partially financed by Spain's *Ministerio de Educación y Ciencia* (program *Juan de la Cierva*) within the context of project AGL2004-08368-CO3.

References

1. Brown P, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21:**33-37.
2. Lander ES: **Array of hope.** *Nat Genet* 1999, **21(suppl):**3-4.
3. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models.** *J Comput Biol* 2001, **8:**625-637.
4. Bachtiary B, Boutros PC, Pintilie M, Bastianutto C, Li J-H, Schwock J, Zhang W, Penn LZ, Jurisica I, Fyles A, Liu F-F: **Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity.** *Hum Cancer Res* 2006, **12:**5632-5640.
5. Somel M, Khaitovich P, Bahn S, Pääbo S, Lachmann M: **Gene expression becomes heterogeneous with age.** *Current Biol* 2006, **16:**R360.
6. Gill JL: **Analyses of data with heterogeneous variance: a review.** *J Dairy Sci* 1971, **54:**369-373.
7. Weigel KA, Gianola D: **Estimation of heterogeneous within-herd variance components using empirical Bayes methods: a simulation study.** *J Dairy Sci* 1992, **75:**2824-2833.
8. Kendziorski CM, Newton MA, Lan H, Gould MN: **On parametric empirical Bayes methods for comparing multiple groups using replicate gene expression profiles.** *Stat Med* 2003, **22:**3899-3914.
9. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametrical hierarchical mixture method.** *Biostatistics* 2004, **5:**155-176.
10. Lo K, Gottardo R: **Flexible empirical Bayes methods for differential gene expression.** *Bioinformatics* 2007, **23:**328-335.
11. Verdinelli I, Wasserman L: **Computing Bayes factors using a generalization of the Savage-Dickey density ratio.** *J Am Stat Assoc* 1995, **90:**614-618.
12. Varona L, García-Cortés LA, Pérez-Enciso M: **Bayes factors for detection of Quantitative Trait Loci.** *Genet Sel Evol* 2001, **33:**133-152.
13. Casellas J, Ibáñez-Escriche N, Martínez-Giner M, Varona L: **GEAMM v.1.4.: a versatile program for mixed model analysis of gene expression data.** *Anim Genet* 2008, **39:**89-90.
14. Cui X, Churchill CA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4:**210.
15. Gelfand A, Smith AFM: **Sampling based approaches to calculating marginal densities.** *J Am Stat Assoc* 1990, **85:**398-409.
16. Hastings WK: **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika* 1970, **57:**97-109.
17. Wang CS, Rutledge JJ, Gianola D: **Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs.** *Genet Sel Evol* 1994, **26:**91-115.
18. García-Cortés LA, Cabrillo C, Moreno C, Varona L: **Hypothesis testing for the genetic background of quantitative traits.** *Genet Sel Evol* 2001, **33:**3-16.
19. Raftery AE, Lewis SM: **How many iterations in the Gibbs sampler?** In *Bayesian Statistics IV* Edited by: Bernardo JM, Berger JO, Dawid AP, Smith AFM. Oxford, Oxford University Press; 1992:763-773.
20. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
21. Karaman MW, Houck ML, Chemnick LG, Nagpal S, Chawannakul D, Sudano D, Pike BL, Ho VV, Ryder OA, Hacia JG: **Comparative analysis of gene-expression patterns in human and African great ape.** *Genome Res* 2003, **13:**1619-1630.
22. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformation for gene expression data.** *Bioinformatics* 2001, **17:**977-987.
23. Jeffreys H: *Theory of probability* Oxford, Clarendon Press; 1961.
24. Casellas J, Piedrafita J: **Bayes factor for testing the genetic background of quantitative threshold traits.** *J Anim Breed Genet* 2006, **123:**301-306.
25. Newton MA, Raftery AE: **Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion).** *J Royal Stat Soc B* 1994, **56:**3-48.

26. Spiegelhalter DJ, Best NG, Carlin BP, Linde A van der: **Bayesian measures of model complexity and fit (with discussion).** *J Royal Stat Soc B* 2002, **64:**583-639.
27. Searle SR, Casella G, McCulloch CE: *Variance components* New York, John Wiley & Sons; 1992.
28. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17:**509-519.
29. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates.** *Biostatisitics* 2005, **6:**59-75.
30. Lonnstedt I, Speed T: **Replicated microarray data.** *Statistica Sinica* 2002, **12:**31-46.
31. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**. article 3
32. Feng S, Wolfinger RD, Chu TM, Gibson GC, McGraw LA: **Empirical Bayes analysis of variance component models for microarray data.** *J Agric Biol Env Stat* 2006, **11:**197-209.
33. Lee MLT, Kuo FC, Whitemore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97:**9834-9838.
34. Ibrahim JG, Chen M-H, Gray RJ: **Bayesian models for gene expression with DNA microarray data.** *J Am Stat Assoc* 2002, **97:**88-99.
35. McLachlan GJ, Bean RW, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18:**413-422.
36. Jaffrezic F, Marot G, Degrelle S, Hue I, Foulley JL: **A structural mixed model for variances in differential gene expression studies.** *Genet Res* 2007, **89:**19-25.
37. Ho SB, Niehans GA, Lyftogt C, Yan PS, Cherwitz DL, Gum ET, Dahiya R, Kim YS: **Heterogeneity of mucin gene expression in normal and neoplastic tissues.** *Cancer Res* 1993, **53:**641-651.
38. Galligan CL, Baig E, Bykerk V, Keystone EC, Fish EN: **Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity.** *Genes Immunity* 2007, **8:**480-491.
39. Kass RE, Raftery AE: **Bayes factors.** *J Am Stat Assoc* 1995, **90:**773-795.
40. Stram DO, Lee JW: **Variance components testing in longitudinal mixed effects model.** *Biometrics* 1994, **50:**1171-1177.
41. Fisher RA: **The fiducial argument in statistical inference.** *Ann Eugenics* 1935, **6:**391-398.
42. Welch BL: **The significance of the difference between two means when the population variances are unequal.** *Biometrika* 1938, **29:**350-362.
43. **GEAMM v.1.7** [http://www.bdporc.irta.es/Publicacions/GEAMM.zip]