OPEN ACCESS Freely available online

PLoS one

Modeling Skewness in Human Transcriptomes

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

1 Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, Spain, **2** Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, Zaragoza, Spain

Abstract

Gene expression data are influenced by multiple biological and technological factors leading to a wide range of dispersion scenarios, although skewed patterns are not commonly addressed in microarray analyses. In this study, the distribution pattern of several human transcriptomes has been studied on free-access microarray gene expression data. Our results showed that, even in previously normalized gene expression data, probe and differential expression within probe effects suffer from substantial departures from the commonly assumed symmetric Gaussian distribution. We developed a flexible mixed model for non-competitive microarray data analysis that accounted for asymmetric and heavy-tailed (Student's *t* distribution) dispersion processes. Random effects for gene expression data were modeled under asymmetric Student's *t* distributions where the asymmetry parameter (λ) took values from perfect symmetry ($\lambda = 0$) to right- ($\lambda > 0$) or left-side ($\lambda > 0$) over-expression patterns. This approach was applied to four free-access human data sets and revealed clearly better model performance when comparing with standard approaches accounting for traditional symmetric Gaussian distribution patterns. Our analyses on human gene expression data revealed a substantial degree of right-hand asymmetry for probe effects, whereas differential gene expression addressed both symmetric and left-hand asymmetric patterns. Although these results cannot be extrapolated to all microarray experiments, they highlighted the incidence of skew dispersion patterns in human transcriptome; moreover, we provided a new analytical approach to appropriately address this biological phenomenon. The source code of the program accommodating these analytical developments and additional information about practical aspects on running the program are freely available by request to the corresponding author of this article.

Citation: Casellas J, Varona L (2012) Modeling Skewness in Human Transcriptomes. PLoS ONE 7(6): e38919. doi:10.1371/journal.pone.0038919

Editor: Timothy Ravasi, King Abdullah University of Science and Technology, Saudi Arabia

Received August 19, 2011; **Accepted** May 16, 2012; **Published** June 11, 2012

Funding: The research contract of JC was partially funded by Spain's Ministerio de Ciencia e Innovación (reference, RYC-2009-04049). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding was received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: joaquim.casellas@uab.cat

Introduction

Mixed models have been advocated in gene expression analyses due to their superiority in partitioning sources of variation and their flexibility for accommodating various experimental designs [1]; furthermore, they can be used for joint analysis of all loci [2], appropriately accounting for variability both across and within microarray probes [3]. Microarray data sets are characterized by high dimensionality in the sense of a small number of replicates (i.e. microarray slides) and a large number of probes per replicate. Mixed models account for these peculiarities of the microarray gene expression data, the sources of variation being preferentially treated as random effects [4] to appropriately address large numbers of levels with scarce amounts of information per level. A typical assumption for the distribution of random effects in mixed model analyses is the Gaussian density function [4], which is systematically applied in standard gene expression analyses [3,5]. Although this parametric assumption could be viewed as a reasonable compromise between mathematical convenience and biological plausibility, its suitability in gene-expression analyses has been questioned in recent studies [6–12].

Taking the probe-specific differential expression effect between two treatments, the Gaussian distribution forces a symmetrical pattern between the two treatments, whereas a wide range of skewed distributions and treatment-related over-expressions may seem more reasonable. Moreover, the Gaussian assumption suffers substantial misadjusts in the presence of outliers [13], which are common in microarray data [14]. Given the inconsistencies of the Gaussian distribution for random effects in the gene expression data, recent researches have proposed parametric alternatives for modeling gene expression data, assuming heavy-tailed processes like Cauchy [8] and Student's *t* distributions [7] or asymmetric distributions like Pareto [15], Gamma [6] and skew Laplace [16,17]. Although these studies have reported substantial improvements in terms of model fit to experimental data, none of them allowed joint, flexible modeling of gene expression data under variable incidence of outliers or asymmetry, or the incidence of both positive (right-hand tail over-expressed) and negative (left-hand tail over-expressed) asymmetric patterns.

The Student's *t* distribution has been proposed as a useful assumption for attenuating the impact of outliers in mixed models [7] and, furthermore, asymmetry can be easily accommodated in the Student's *t* density [18]. Within this context, the aim of this research was to check for asymmetric and heavy-tailed patterns in random effects of gene expression data, developing a new analytical approach to appropriately accommodate both sources of departure from the standard symmetric Gaussian assumption.

Results

Model Comparison

Four independent microarray gene expression data sets from human tissues (Table 1) were analyzed under three different

hierarchical mixed linear models. These analyses accounted for the systematic effect of each microarray slide and two random sources of variation, probe and treatment within probe (with two levels in each data set). Models differed in the a priori distributions of these random effects, they being multivariate normal densities (Model SG) [5], symmetric Student's t densities (Model ST), or asymmetric Student's t densities (Model AS) following Sahu et al. [18]. The deviance information criterion (DIC) [19] assessed model performance under these three different prior distributions for random effects, revealing a huge penalization for model SG in all cases (Table 2). Note that models with a smaller DIC were favored as this indicated a better fit and lower degree of model complexity [19]. In all four comparisons, DIC sequentially and drastically reduced with models ST and AT (Table 2), where Sahu et al. [18] asymmetric Student's t priors for probe and differential expression within probe effects were clearly preferred (Table 2). Note that differences larger that 3 to 5 DIC units are assumed as statistically relevant [19] and Model AT showed the lowest DIC with 4,678 (dataset 1) to 60,335 (dataset 3) less DIC units than Model ST. These large DIC departures ruled out any possible controversy concerning the most preferable model. It is important to note that the number of differentially expressed genes reduced with Model AT, also suggesting a more conservative behavior for this parameterization (Table 2).

Estimates for Asymmetry and non-Gaussian Patterns

Under Model AT, the non-Gaussian distributions of probe and differential expression within probe effects were characterized in terms of heavy tails (Student's t) and asymmetric dispersion patterns by means of v (degrees of freedom of the Student's t distribution) and λ (asymmetry parameter). Probe effects revealed both heavy tails and positive asymmetry with a substantial over-expression of the right tail of the distribution. The modal estimates of the degrees of freedom fluctuated between 5.62 (data set 2) and 8.95 (data set 1), with the highest posterior density region at 95% roughly ranged between 4 and 30. The right-hand asymmetry was clearly demonstrated in all datasets with positive modal estimates of λ, their HPD95 excluding the null or negative values (Figure 1a). The differential expression within probe effect showed a similar pattern with small v, although significant asymmetry was only revealed in data set 3 ($\lambda = -1.88$; HPD95: -1.96 to -1.81; Figure 1b).

Discussion

The asymmetric and non-Gaussian distribution of the human transcriptome has been revealed in four independent human data sets from different microarray platforms and technologies. Although our results cannot be completely extrapolated to all microarray data, they show that deviations from the standard Gaussian prior for random effects should be accurately considered in current gene expression studies. Normalization of gene expression data has been a topic of main interest during the last decade [20,21], but our results suggested that non-Gaussian patterns must be considered as an inherent property of gene expression data, and this phenomenon should be appropriately accounted for in analytical models in order to avoid biases on final estimates (Table 2). Note that the Student's t density converges to a Gaussian density when v tends to infinity, although both densities are assumed roughly similar for v values larger than 30 [13]. In our case, small modal estimates (<10) were obtained for the degrees of freedom of the Student's t distribution, suggesting a relevant departure from the standard Gaussian distribution as corroborated by the DIC statistic. Our small values for v reported a substantial incidence of outlier gene expressions as was previously sugested by Gottardo et al. [7] and Khondoker et al. [8] in alternative microarray data sets. Moreover, Model AT was preferred, highlighting the usefulness of the hierarchical mixed model with asymmetric Student's t prior distributions for random sources of variation.

All data sets agreed with right-hand over-distributed probe effects, whereas left-hand over-expression was revealed for differential expression within probe estimates in data set 3 (Figure 1b). This right-hand asymmetry in human transcriptome must be linked to the fact that lowly expressed probes are roughly grouped in the left tail of the scanning spectrum due to technological limitations of the microarray technique, whereas a substantial incidence of high or extremely-high gene expression intensities can be anticipated [22]. Note that this phenomenon is not commonly accounted for in gene expression analyses worldwide, whereas the mixed model parameterization developed in this manuscript provides a highly flexible statistical tool accounting for the non-Gaussian properties of human (and even non-human) transcriptome. As shown in Figure 1a, departures from the standard Model SG do not reduce to the symmetry pattern only, but also rely on the average mathematical expectation for probe effects. Sahu's et al. [18] method can

Table 1. Summary of the free-access data sets analyzed.

	Platform[a]	Tissue	Groups of comparison (number of samples per group)	Reference	GEO[b]
Dataset 1	Affymetrix GeneChip Human Full Length Array HuGeneFL	Mononuclear cell layer	Non-pulmonary arterial hypertension (6) *vs.* pulmonary arterial hypertension (14)	Bull et al. [26]	GSE703
Dataset 2	Affymetrix GeneChip Human Full Length Array HuGeneFL	Bronchoalveolar lavage cells	Non-smoker (5) *vs.* smoker (5)	Heguy et al. [27]	GSE3212
Dataset 3	Affymetrix GeneChip Human Genome U133 Plus 2.0 Array	Spermatozoa	Normal (12) *vs.* teratozoospermic individuals (8)	Platts et al. [28]	GSE6969
Dataset 4	Illumina humanRef-8 v2.0 expression beadchip	Carotid endarterectomy samples	Carotid artery stenosis treated with mycophenolate (9) *vs.* placebo (11)	Unpublished	GSE13922

[a]The approximate number of interrogated transcripts were 5,000, 47,000 and 16,000 for Affymetrix GeneChip Human Full Length Array HuGeneFL (Affymetrix, Inc., Santa Clara, CA), Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (Affymetrix, Inc., Santa Clara, CA) and Illumina human Ref-8 v2.0 expression beadchip (Illumina, Inc., San Diego, CA), respectively.
[b]Gene Expression Omnibus accession number (http://www.ncbi.nlm.nih.gov/geo/).
doi:10.1371/journal.pone.0038919.t001

Table 2. Model comparison and characterization of the dispersion patter of probe and differential expression within probe under Model AT.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
DIC[a] (and number of probes with significant differential expression[b])				
Model SG[c]	284,161 (189)	231,581 (5)	3,692,344 (702)	1,756,122 (12)
Model ST[d]	247,509 (31)	224,741 (1)	3,614,823 (692)	1,734,053 (4)
Model AT[e]	242,831 (2)	188,835 (0)	3,554,488 (639)	1,724,667 (2)
Parameters[f] under Model AT. Mode (and highest posterior density region at 95%)				
v_p	8.95 (4.21 to 26.61)	5.62 (4.16 to 11.05)	6.77 (4.15 to 16.0)	8.87 (4.40 to 30.09)
λ_p	0.38 (0.04 to 0.66)	0.13 (0.01 to 0.32)	1.84 (1.61 to 1.93)	2.03 (1.98 to 2.09)
v_d	7.36 (4.18 to 18.05)	6.90 (4.15 to 19.76)	5.99 (4.38 to 11.51)	8.48 (4.66 to 23.90)
λ_d	0.01 (−0.04 to 0.06)	−0.00 (−0.04 to 0.04)	−1.88 (−1.96 to −1.81)	−0.00 (−0.01 to 0.01)

[a]Deviance information criterion.
[b]Differentially expressed genes after Bonferroni [29]-like correction ($\alpha = 0.05$). The adjusted significance threshold for posterior probabilities was calculated as α/π, were π was the number of probes included in each analysis.
Random effects **g** and **d**(**g**) were assumed as symmetric Gaussian[c], symmetric Student's t[d] or asymmetric Student's t[e] distributed following Sahu et al. [18].
[f]Degrees of freedom (v) and asymmetry parameter (λ) for probe (p) and differential expression within probe (d) effects.
doi:10.1371/journal.pone.0038919.t002

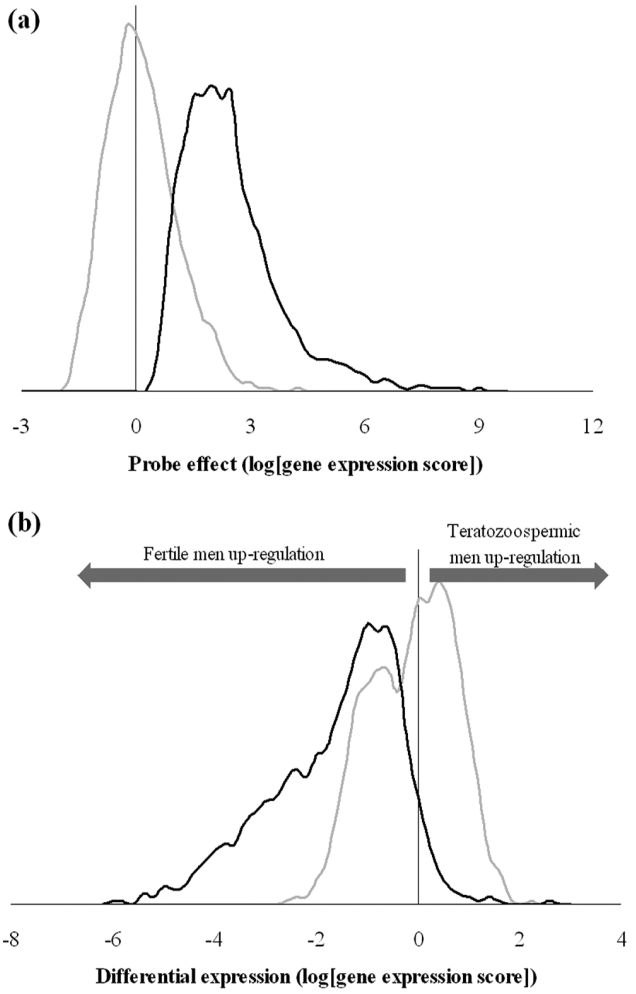

Figure 1. Distribution of mean estimates for probe (**a**) and differential expression within-probe (**b**) effects under Model SG (grey line) and Model AT (black line) for data set 3.
doi:10.1371/journal.pone.0038919.g001

accommodate distributions with non-zero modal estimates (Figure 1a). Focusing on data set 3, the modal estimate for differential expression effects was placed around 2 (Model AT) and linked to larger estimates for the array effect when comparing with Model SG. It implied a moderate relocation of systematic and random sources of variation for gene expression data in the output of the mixed model.

Results from differential expression within probe effects highlighted the remarkable flexibility of Sahu's et al. [18] method to accommodate any kind of asymmetry pattern. This peculiarity is of special relevance for differential gene expression given that a wide range of asymmetric patterns could be find in gene expression studies. Although the standard symmetric Gaussian distributions may be valid sometimes, a wide range of left- and right-tail over-expressions could be addressed with Model AT. Indeed, data set 3 (all datasets) showed that asymmetric (heavy-tailed) patterns are not unusual and they must be considered in gene expression analyses. The relevance of a proper modeling of random effects is clearly highlighted in Figure 1b where the symmetric Gaussian prior distribution for the differential expression produces a bimodal artifact in the posterior distribution of the estimates, clearly differing from the expected drawn under the a priori assumption.

Note that all model comparisons were made on the basis of the DIC statistic [19], a widely used statistical criterion to assess model complexity and fit. Indeed, DIC measures posterior predictive error by penalizing the fit of the model (i.e., deviance) by its complexity, determined by the effective number of parameters as defined by Spiegelhalter et al. [19]. Within this context, model AT must be clearly viewed as the most parsimonious and reliable parameterization, at least among the alternatives we are considering in this study. DIC evidenced that the incidence of asymmetry and heavy-tailed patterns in human gene expression data must be out of any doubt and, as consequence, model AT characterized a quasi-optimum approach to analyze this kind of microarray data. Nevertheless, DIC does not provide specific information about testing properties of current models when evaluating differentially expressed genes, although better model fit must be linked to better testing properties. Model AT reported the smallest number of differentially expressed probes in all data sets

(Table 2) and all those probes were previously identified as differentially expressed by models SG and ST. The same patterns were obtained in preliminary analyses of simulated gene expression data (results not shown) and suggested that the better fit and more restrictive testing behavior of model AT could be linked to false positives under models SG and ST.

In conclusion, the incidence of asymmetric random effects has been highlighted in non-competitive gene expression data from human tissues; the new model proposed below provides a better adjustment of gene expression data and even a more conservative testing pattern has been suggested. Although this manuscript has focused on non-competitive hybridization microarrays, models can be easily adapted to two channel microarrays following Purdom and Holmes [16].

Materials and Methods

Mixed Model for Non-competitive Microarray Data

We assume as a starting point non-competitive hybridization microarray data from n unrelated individuals appropriately grouped in two different treatments (e.g. normal versus tumor cells) and m probes. These data ($\mathbf{y}$) can be analyzed by the mixed model:

$$\mathbf{y}=\mathbf{X}\mathbf{a}+\mathbf{Z}_1\mathbf{p}+\mathbf{Z}_2\mathbf{d}(\mathbf{p})+\mathbf{e}$$

where $\mathbf{X}$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ are incidence matrices for array ($\mathbf{a}$), probe ($\mathbf{p}$) and differential expression (between treatments) within probe ($\mathbf{d}(\mathbf{p})$) effects, and $\mathbf{e}$ is the vector of residuals. Following a standard Bayesian development, the joint posterior distribution of all parameters in the model conditional to the data is proportional to the Bayesian likelihood,

$$p(\mathbf{y}|\mathbf{a},\mathbf{p},\mathbf{d}(\mathbf{p}),\mathbf{R})\propto \exp\left[-\frac{1}{2}\left(\mathbf{X}\mathbf{a}+\mathbf{Z}_1\mathbf{p}+\mathbf{Z}_2\mathbf{d}(\mathbf{p})'\mathbf{R}^{-1}(\mathbf{X}\mathbf{a}+\mathbf{Z}_1\mathbf{p}+\mathbf{Z}_2\mathbf{d}(\mathbf{p}))\right)\right]$$

multiplied by the *a priori* distribution of each parameter in the model. Note that this equation describes a heteroskedastic normal density [5] with gene-specific residual variances and with null residual covariance between genes ($\mathbf{R}$). *A priori* distributions for $\mathbf{p}$ and $\mathbf{d}(\mathbf{p})$ could be described as

$$p\left(\mathbf{p}|\sigma_p^2\right)\propto \prod_{i=1}^{r}\exp\left(-\frac{p_i^2}{2\sigma_p^2}\right)$$

and

$$p\left(\mathbf{d}(\mathbf{p})|\sigma_d^2\right)\propto \prod_{j=1}^{s}\exp\left(-\frac{d_j^2}{2\sigma_d^2}\right)$$

they being independent Gaussian densities with a mean of zero and variances equal to σ_p^2 and σ_d^2, respectively (Model SG). Note that i was the number of elements in $\mathbf{p}$ and j was the number of elements in $\mathbf{d}(\mathbf{p})$. Nevertheless, robustness must be gained under a skew-Student's t prior. This prior can be parameterized as a skewed-normal density,

$$p\left(\boldsymbol{\alpha}|\sigma_\alpha^2,\,\lambda_\alpha,\,s_k\right)\propto \prod_{k=1}^{t}\frac{2}{\sqrt{\left(\frac{\sigma_\alpha^2}{s_k^2}\right)+\lambda_\alpha^2}}\phi\left(\frac{\alpha_k}{\sqrt{\left(\frac{\sigma_\alpha^2}{s_k^2}\right)+\lambda_\alpha^2}}\right)\Phi\left(\frac{\lambda_\alpha}{\left(\frac{\sigma_\alpha}{s_k}\right)}\frac{\alpha_k}{\sqrt{\left(\frac{\sigma_\alpha^2}{s_k^2}\right)+\lambda_\alpha^2}}\right),$$

multiplied by the conditional distribution of the mixing parameter (s_k), this being a Gamma prior,

$$p\left(s_k^2|v_\alpha\right)=\frac{\left(\frac{1}{2v_\alpha}\right)^{\frac{1}{2v_\alpha}}}{\Gamma\left(\frac{1}{2v_\alpha}\right)}\left(s_k^2\right)^{\left(\frac{1}{2v_\alpha}-1\right)}\exp\left(-\frac{s_k^2}{2v_\alpha}\right)$$

Note that σ_α^2 was the scale parameter, v_α were the degrees of freedom, and λ_α was the asymmetry parameter modelled following Sahu et al. [18] (Model AT). Moreover, ϕ and Φ denoted the density function and cumulative distribution function of a standard normal distribution with kernel as defined between parentheses, respectively, and Γ was the standard gamma function with argument as defined within parentheses. Note that $\lambda_\alpha=0$ describes perfect symmetry, whereas right (or left) tail proportionally increases for positive (or negative) values of λ_α. An uniform prior distribution was defined for λ_α, as previously suggested by Varona et al. [23]. Symmetric Student's t priors (Model ST) can be easily defined if λ_α is appropriately fixed to 0. *A priori* distributions for degrees of freedom were defined as exponential [24] and flat priors were assumed for the remaining parameters. Note that the Student's t density converges to the Gaussian one when degrees of freedom tend to infinity, whereas few degrees of freedom account for heavy-tailed densities [13]. All the unknown factors in the model can be easily sampled from their joint posterior distribution by Markov chain Monte Carlo methods [25].

Example with Free-access Human Gene Expression Data

To illustrate the asymmetric pattern of the human transcriptome, we applied the models to four free-access human microarray datasets (http://www.ncbi.nlm.nih.gov/geo/; accession numbers GSE703, GSE3212, GSE6969 and GSE 13922). Note that all data sets are MIAME compliant and they were previously deposited in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/). These datasets were representative of two different trademarks and hybridization technologies, evaluated in diverse human tissues (see Table 1). All of them focused on the comparison between two groups, non-pulmonary arterial hypertension *versus* pulmonary arterial hypertension (Dataset 1; [26]), non-smoker *versus* smoker (Dataset 2; [27]), normal *versus* teratozoospermic individuals (Dataset 3; [28]) and carotid artery stenosis treated with mycophenolate *versus* placebo (Dataset 4; unpublished). A base 2 logarithm was applied to normalize gene-expression scores.

Note that the four human data sets were selected at random to evaluate the three mixed model parameterizations on different human tissues and microarray platforms. Of course, both tissue and data quality could have some impact on the distribution pattern, although this escaped from the objectives of this research. Different preprocessing approaches would have different impacts on further analyses of gene expression data and even skewed or heavy-tailed patterns could be partially addressed by preliminary data editing

methodologies such as normalization of background correction. Nevertheless, we focused on the development, implementation and evaluation of a reliable parameterization to account for non-Gaussian patterns in gene expression data, assuming that all preliminary data editing processes where properly satisfied.

For each dataset, the three different models were analyzed (models SG, ST and AT). Each model was solved through Bayesian inference with a single Monte Carlo Markov chain of 500,000 elements after discarding the first 50,000 as burn-in. Models were compared with the DIC [19].

Author Contributions

Conceived and designed the experiments: JC LV. Performed the experiments: JC. Analyzed the data: JC. Contributed reagents/materials/analysis tools: JC LV. Wrote the paper: JC.

References

1. Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 4: 210.
2. Hoeschele I, Li H (2005) A note on joint versus gene-specific mixed model analysis of microarray gene expression data. Biostatistics 6: 183–186.
3. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, et al. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8: 625–637.
4. Searle SR (1982) Matrix Algebra Useful for Statistics. John Wiley & Sons, New York, NY.
5. Casellas J, Ibáñez-Escriche N, Martínez-Giner M, Varona L (2008) GEAMM v1.4.: a versatile program for mixed model analysis of gene expression data. Anim Genet 39: 89–90.
6. Kendziorski CM, Newton MA, Lan H, Gould MN (2003) On parametric empirical Bayes methods for comparing multiple groups using replicate gene expression profiles. Stat Med 22: 3899–3914.
7. Gottardo R, Raftery AE, Yeung KY, Bumgarner RE (2006) Bayesian robust inference for differential gene expression in microarrays with multiple samples. Biometrics 62: 10–18.
8. Khondoker MR, Glasbey CA, Worton BJ (2006) Statistical estimation of gene expression using multiple laser scans of microarrays. Bioinformatics 22: 215–219.
9. Angelini C, Cutillo L, De Canditiis D, Mutarelli M, Pensky M (2008) BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. BMC Bioinformatics 9: 415.
10. Hardin J, Wilson J (2009) A note on oligonucleotide expression values not being normally distributed. Biostatistics 10: 446–450.
11. Salas-Gonzalez D, Kuruoglu EE, Ruiz DP (2009) A heavy-tailed empirical Bayes method for replicated microarray ata. Comput Stat Data Anal 53: 1535–1546.
12. Posekany A, Felsenstein K, Sykacek P (2011) Biological assessment of robust noise models in microarray data analysis. Bioinformatics 27: 807–814.
13. Lange KL, Little RJA, Taylor JMG (1989) Robust statistical modelling using the t distribution. J Am Stat Assoc 84: 881–896.
14. Model F, König T, Piepenbrock C, Adorján P (2002) Statistical process control for large scale microarray experiments. Bioinformatics 18: S155–S162.
15. Kuznetsov VA, Knott GD, Bonner RF (2002) General statistics of stochastic process of gene expression in eukaryotic cells. Genetics 161: 1321–1332.
16. Purdom E, Holmes SP (2005) Error distribution for gene expression data. Stat Appl Genet Mol Biol 4: 16.
17. Bhowmick D, Davison AC, Goldstein DR, Ruffieux Y (2006) A Laplace mixture model for identification of differential expressions in microarray experiments. Biostatistics 7: 630–641.
18. Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. Can J Stat 31: 129–150.
19. Spiegelhalter DJ, Best NG, Carlin BP, var der Linde A (2002) Bayesian measures of model complexity and fit. J Royal Statist Soc B 64: 583–639.
20. Beyene J, Hu P, Parkhomenko E, Tritchler D (2007) Impact of normalization and filtering on linkage analysis of gene expression data. BMC Proc 1: S150.
21. Smyth GK, Speed TP (2003) Normalization of cDNA microarray data. Methods 31: 265–273.
22. Chen D-T, Lin S-H, Soong S-J (2004) Gene selection for oligonucleotide array: an approach using PM probe level data. Bioinformatics 20: 854–862.
23. Varona L, Ibáñez-Escriche N, Quintanilla R, Noguera JL, Casellas J (2008) Bayesian analysis of quantitative traits using skewed distributions. Genet Res 90: 179–190.
24. Strandén I, Gianola D (1999) Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. Genet Sel Evol 31: 25–42.
25. Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall, New York, NY.
26. Bull TM, Coldren CD, Moore M, Sotto-Santiago SM, Pham DV, et al. (2004) Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. Am J Resp Crit Care Med 170: 911–919.
27. Heguy A, O'Connor TP, Luettich K, Worgall S, Cieciuch A, et al. (2006) Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsomkers reveals a previously unrecognized subset of genes modulated by cigarette smoking. J Mol Med 84: 318–328.
28. Platts AE, Dix DJ, Chemes HE, Thompson KE, Goodrich R, et al. (2007) Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. Hum Mol Genet 16: 763–773.
29. Bonferroni CE (1930) Elementi di Statistica Generale. Libreria Seber, Florence, Italy.