Biostatistics (2012), **13**, 3, *pp.* 553–566
doi:10.1093/biostatistics/kxr042
Advance Access publication on November 14, 2011

ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments

MARIA J. NUEDA*

Departamento de Estadística e Investigación Operativa, Universidad de Alicante,
Apartado 03080, Alicante, Spain
mj.nueda@ua.es

ALBERTO FERRER

Departamento de Estadística e Investigación Operativa Aplicadasy Calidad, Universidad
Politécnica de Valencia, Apartado 46022, Valencia, Spain

ANA CONESA

Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe,
Avenida Autopista Saler 16, 46012 Valencia, Spain

Summary

Transcriptomic profiling experiments that aim to the identification of responsive genes in specific biological conditions are commonly set up under defined experimental designs that try to assess the effects of factors and their interactions on gene expression. Data from these controlled experiments, however, may also contain sources of unwanted noise that can distort the signal under study, affect the residuals of applied statistical models, and hamper data analysis. Commonly, normalization methods are applied to transcriptomics data to remove technical artifacts, but these are normally based on general assumptions of transcript distribution and greatly ignore both the characteristics of the experiment under consideration and the coordinative nature of gene expression. In this paper, we propose a novel methodology, ARSyN, for the preprocessing of microarray data that takes into account these 2 last aspects. By combining analysis of variance (ANOVA) modeling of gene expression values and multivariate analysis of estimated effects, the method identifies the nonstructured part of the signal associated to the experimental factors (the noise within the signal) and the structured variation of the ANOVA errors (the signal of the noise). By removing these noise fractions from the original data, we create a filtered data set that is rich in the information of interest and includes only the random noise required for inferential analysis. In this work, we focus on multifactorial time course microarray (MTCM) experiments with 2 factors: one quantitative such as time or dosage and the other qualitative, as tissue, strain, or treatment. However, the method can be used in other situations such as experiments with only one factor or more complex designs with more than 2 factors. The filtered data obtained after applying ARSyN can be further analyzed with the appropriate statistical technique to obtain the biological information required. To evaluate the performance of the filtering strategy, we have applied different statistical approaches for MTCM analysis to several real and simulated

[*] To whom correspondence should be addressed.

data sets, studying also the efficiency of these techniques. By comparing the results obtained with the original and ARSyN filtered data and also with other filtering techniques, we can conclude that the proposed method increases the statistical power to detect biological signals, especially in cases where there are high levels of structural noise. Software for ARSyN is freely available at http://www.ua.es/personal/mj.nueda.

Keywords: Analysis of variance; Batch effect; Microarrays; Principal components analysis; Systematic noise.

1. Introduction

Time course microarray (TCM) experiments analyze time-dependent transcriptional changes along one or more series of data. The TCM design is employed when the dynamics of gene expression changes are to be studied as a response to a drug treatment, for its association with a genetic background or simply as a consequence of development or aging. If a second factor, such as diversity of treatments, strains, or environment, is present in the study, we are dealing with a multifactorial time course microarray (MTCM) experiment. Examples of such controlled multifactorial experiments can be found in the fields of toxicology (Heijne *and others*, 2003), agronomy (Brumós *and others*, 2009), biomedicine (Agudo *and others*, 2008), and ecology (Svendsen *and others*, 2008), to cite just a few. Although recent advances in sequencing technologies have created alternatives to microarrays for transcriptome profiling, the relatively high costs of sequencing platforms rule out their use in complex transcriptomics experiments such as the MTCM in which a large number of conditions and samples are required. In these circumstances, microarrays continue to be the preferred option to address genome-wide gene expression analysis. Typically, in MTCM designs, time constitutes one factor, a variable of quantitative nature, while the other factors are either quantitative or qualitative (dosis/level, treatment, strain, etc.). Statistical analysis of this kind of data is more complicated than that of simple control–cases studies. In MTCM, not only significant changes at different factor levels and interactions are sought but also the identification of patterns of transcriptional regulation is frequently pursued. Several methodologies for the analysis of TCM have been proposed so far (Conesa *and others*, 2006; Tai and Speed, 2006; Storey *and others*, 2005) that apply different statistical strategies for the modeling of time-dependent gene expression and the identification of significant changes.

One of the aspects that has received most attention in methodological studies of microarray data analysis is the treatment of noise. Although microarrays have greatly improved technical quality and reproduction over the years, microarray data are still highly noise prone and are affected by random and systematic sources of error that obscure the transcriptional signal. The first step in the analysis of microarray data is usually normalization, whose aim is to adjust data from different arrays to a common baseline and distribution. This data treatment addresses sources of technical variation such as hybridization efficiency, starting messenger RNA concentrations, or different physical properties of labeling molecules. Normalization methods have been established over the last decade (Do and Choi, 2006). However, not all sources of technical noise are removed by normalization. This is due to the fact that most of the current normalization methods are designed to center and scale the data assuming general invariability for all observations and ignoring the particular sample hybridized in each array (Yang *and others*, 2002). When exploring normalized microarray data using common clustering techniques, it is still not infrequent to observe artifacts associated to identifiable factors such as the array type, the lab, or the date of execution generally referred to as "batch effects." Moreover, other types of systematic biases that are not as traceable as the batch effects might also be embedded in the data. All these elements represent sources of structured noise that reduce statistical power when assessing differential expression.

Batch effects are present in many data sets, and this can seriously hinder statistical analysis. This technical problem has been recently reviewed within the framework of the MAQC-II Project that studied the quality of microarray data for their application as a molecular prediction tool (MAQC-Consortium, 2010). This project resulted in an extensive evaluation of the batch effect and of the existing batch-removal strategies (Luo *and others*, 2010). Some methodologies for removing batch effects require large batch sizes, such as singular value decomposition (Alter *and others*, 2000) and distance weighted discrimination (Benito *and others*, 2004). Empirical Bayes methods have been claimed to be more flexible and robust to outliers since the batch bias is considered common across all genes in each batch (Johnson *and others*, 2007). A requirement for the application of all these strategies is the previous identification of the batches, generally understood as the group of samples affected by the same noise level, and this is not always possible. When systematic noise is associated with an array or spatial effects, the experiment design may be the key for correcting this noise (Leek and Storey, 2007). Moreover, the co-regulation mechanism that underlies gene expression implies that transcriptomics data have an inherent correlation structure. Taking this covariance structure into account is, likewise, an effective way to enhance data analysis.

In this paper, we propose a novel strategy named ARSyN (ASCA [ANOVA simultaneous component analysis] removal of systematic noise). ARSyN is based on the ASCA model developed by Smilde *and others* (2005) to remove structural noise from microarray data sets. ASCA combines analysis of variance (ANOVA) and principal components analysis (PCA) to analyze multifactorial omics data sets. So far, ASCA has been used for exploratory analysis (Jansen *and others*, 2005; Brumós *and others*, 2009) and for the identification of responsive genes in transcriptomics (Nueda *and others*, 2007). In the present work, we take advantage of the data decomposition provided by the ASCA model to develop a novel statistical framework for the preprocessing of microarray data. In brief, ARSyN uses the PCAs of the ANOVA parameters and residuals in the ASCA model to identify and separate noise from signal in microarray data. After this decomposition, the data elements of interest are joined back together to reconstruct a filtered gene expression matrix which is free of structural biases. The filtered matrix has 2 main advantages:

1. Extracts the relevant gene expression variation related to the controlled variables in the experimental design. This is obtained from the main principal components (PCs) of the ANOVA parameters.
2. Is free of structural noise that can be associated to batch effects or to other nontraceable sources of variation. This is identified in the main PCs of the residuals of the ANOVA model.

Although ARSyN relies on the ASCA model, it is a different methodology in scope and statistical realization. While ASCA has been used for descriptive analysis and for the identification of differentially expressed genes and focuses on the analysis of the ANOVA parameters, ARSyN is a preprocessing strategy that renders a noise-reduced expression matrix. The processed data can then be submitted to statistical analysis with any dedicated methodology for (M)TCM.

We have analyzed how ARSyN improves the performance of 3 time course methods: maSigPro (Conesa *and others*, 2006), *timecourse* (Tai and Speed, 2006, 2009), and EDGE (Storey *and others*, 2005). We have employed synthetic data to investigate the effects of the proposed methodology on different types of noise and relationships between samples. Our results demonstrate that ARSyN effectively removes structural (but not random) noise in both independent and longitudinal multifactorial data sets. Finally, we assess the usability of the filtering approach from a biological point of view through the application to 2 experimental scenarios. Furthermore, we compare ARSyN with current batch-removal methods: ComBat (Johnson *and others*, 2007) and surrogate variable analysis (SVA) (Leek and Storey, 2007).

2. Material and methods

2.1 *The ASCA model*

Since the ARSyN approach relies on the ASCA framework, it is pertinent to present this methodology on the first place. We will describe the general case of a multiseries TCM experiment where the experimental design is defined by 2 factors: the time component and the experimental groups for which temporal gene expression differences are studied. More complex experimental designs are equally amenable to ASCA and ARSyN analysis by taking appropriate ANOVA models. Let us consider I time points $(i = 1, \ldots, I)$, J experimental groups $(j = 1, \ldots, J)$, R_{ij} replications, $(r = 1, \ldots, R_{ij})$ for each case ij, and N genes $(n = 1, \ldots, N)$. For each gene, we will denote by x_{ijr} the gene expression measure at the time i, under condition j and for replicate r. The analysis of this experiment using the ASCA approach (Smilde *and others*, 2005) implies the definition of the ANOVA model for each gene by

$$x_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijr}, \tag{2.1}$$

where μ is an offset term, α_i is the model parameter for factor time on level i, β_j measures the jth group effect, $(\alpha\beta)_{ij}$ represents the interaction effect between the ith time and jth group, and the individual variation is indicated by $(\alpha\beta\gamma)_{ijr}$ instead of ϵ_{ijr} to avoid confusion with the error term in the subsequently derived ASCA model.

We consider a microarray experiment with N genes and $M = \sum_{i,j} R_{ij}$ samples; a matrix $\mathbf{X}$ of dimensions $M \times N$ can be defined containing the entire gene expression data set. Similarly, the estimates of the ANOVA parameters on the right-hand side of (2.1) can be obtained for all genes and collected into matrices where rows represent samples and columns represent genes. This gives the expression

$$\mathbf{X} = \mathbf{1}\mathbf{m}^t + \mathbf{X}_a + \mathbf{X}_b + \mathbf{X}_{ab} + \mathbf{X}_{abg}, \tag{2.2}$$

where $\mathbf{1}$ is a size M column vector of ones, $\mathbf{m}^t$ is a size N row vector containing estimates of μ for each gene, matrices $\mathbf{X}_a$, $\mathbf{X}_b$, and $\mathbf{X}_{ab}$ contain the estimates of parameters α_i, β_j, and $(\alpha\beta)_{ij}$, respectively, and $\mathbf{X}_{abg}$ contains the residuals named $(\alpha\beta\gamma)_{ijr}$.

When experimental data contains numerous variables (genes) in which correlation relationships are present, as would normally be expected in transcriptomics, the data matrix $\mathbf{X}$ contains redundant information and, subsequently, so do the matrices $\mathbf{X}_a$, $\mathbf{X}_b$, $\mathbf{X}_{ab}$, and $\mathbf{X}_{abg}$. In this case, information can be summarized by applying multivariate projection techniques that reduce data dimensionality. Given the data decomposition obtained with the ANOVA model, it is possible that each source of variability has different principal directions. It is therefore convenient to apply dimension reduction separately to each one of matrices $\mathbf{X}_a$, $\mathbf{X}_b$, $\mathbf{X}_{ab}$, and $\mathbf{X}_{abg}$. Consequently, the ASCA model corresponding to (2.2) gives us

$$\mathbf{X} = \mathbf{1}\mathbf{m}^t + \overbrace{\overbrace{\mathbf{T}_a\mathbf{P}_a^t + \mathbf{E}_a}^{\mathbf{X}_a} + \overbrace{\mathbf{T}_b\mathbf{P}_b^t + \mathbf{E}_b}^{\mathbf{X}_b} + \overbrace{\mathbf{T}_{ab}\mathbf{P}_{ab}^t + \mathbf{E}_{ab}}^{\mathbf{X}_{ab}}}^{\text{PART I: Signal of interest}} + \overbrace{\overbrace{\mathbf{T}_{abg}\mathbf{P}_{abg}^t + \mathbf{E}_{abg}}^{\mathbf{X}_{abg}}}^{\text{PART II: Residuals}}, \tag{2.3}$$

where the component scores of each submodel are given by the matrices $\mathbf{T}_a$, $\mathbf{T}_b$, $\mathbf{T}_{ab}$, and $\mathbf{T}_{abg}$; the loadings are given by the matrices $\mathbf{P}_a$, $\mathbf{P}_b$, $\mathbf{P}_{ab}$, and $\mathbf{P}_{abg}$; and the residuals of each submodel are collected in $\mathbf{E}_a$, $\mathbf{E}_b$, $\mathbf{E}_{ab}$, and $\mathbf{E}_{abg}$. The analysis of TCM data is focused on the differences between experimental groups, which over time implies the study of this factor jointly with the interaction: $\mathbf{X}_{b+ab} = \mathbf{X}_b + \mathbf{X}_{ab}$. In this case, we will denote as $\mathbf{E}_{b+ab}$ the residuals of this submodel. For more details about the ASCA model, see Jansen *and others* (2005).

2.2 *ARSyN: the filtering strategy*

Equation (2.3) indicates that the ASCA model can be divided into 2 parts: one corresponding to the gene expression signals the experiment tries to reveal ($\mathbf{X}_a$, $\mathbf{X}_b$, $\mathbf{X}_{ab}$) and the other corresponding to the noise captured by the model residuals ($\mathbf{X}_{abg}$). The PCA on these matrices further separates the correlated structure of each element of $\mathbf{X}$ ($\mathbf{T}_x\mathbf{P}_x^t$ elements) from the unstructured variation ($\mathbf{E}_x$ elements). Considering that relevant transcriptomic signals are those shared by different genes as part of co-expression programs, it follows that the $\mathbf{T}_x\mathbf{P}_x^t$ elements of the time, group, and interaction submodels bear information of interest concerning the target experimental factors, while the $\mathbf{E}_x$ elements collect the random noise present within these factors. Hence, a first filtering strategy should consist of subtracting these noise elements from the $\mathbf{X}$ gene expression matrix, as shown in (2.4):

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{E}_a - \mathbf{E}_b - \mathbf{E}_{ab}. \tag{2.4}$$

In this model, the noise embedded in the gene expression value associated to the experimental factors is filtered out. $\tilde{\mathbf{X}}$ still contains the residuals of the ANOVA model (Part II or $\mathbf{X}_{abg}$ in (2.3)). This second component can be relatively large and will typically collect all sources of noise, random and systematic. Random noise is required to carry out an effective inferential analysis, while structured noise, which corresponds to batch and other systematic errors, is an unwanted feature that distorts statistical analysis. Formally, this systematic noise can be modeled as the latent structures present in the $\mathbf{X}_{abg}$ component of the ASCA model, which are collected by the $\mathbf{T}_{abg}\mathbf{P}_{abg}^t$ element of the SCA of this matrix. Therefore, by subtracting the $\mathbf{T}_{abg}\mathbf{P}_{abg}^t$ element in (2.4), we obtain (2.5) that represents the ARSyN filtering. In this formulation, we generate a modified data matrix $\tilde{\tilde{\mathbf{X}}}$ from which both the noise of the signal ($\mathbf{E}$ matrices of submodels a, b, and ab) and the signal of the noise ($\mathbf{T}_{abg}\mathbf{P}_{abg}^t$ element of the ANOVA error) are supposedly removed. Consequently, we suggest that $\tilde{\tilde{\mathbf{X}}}$ is used instead of $\mathbf{X}$ as signal-enriched data for further analysis by any statistical methodology for TCM.

$$\tilde{\tilde{\mathbf{X}}} = \mathbf{X} - \underbrace{\mathbf{E}_a - \mathbf{E}_b - \mathbf{E}_{ab}}_{\text{Noise of the signal}} - \underbrace{\mathbf{T}_{abg}\mathbf{P}_{abg}^t}_{\text{Signal of the noise}}. \tag{2.5}$$

The ASCA approach requires the selection of a given number of PCs to obtain the different PCA submodels. The number of components chosen affects the distribution of variability between signal and noise and, subsequently, the goodness of fit of the solution. Moreover, the magnitude of the filtering applied depends on the correctness of this selection. There are several common methods for PC selection, for example, analysis of the scree-plots, cross-validation, and choosing a predefined threshold of variability. As the goal here is to select the variation of interest described by models $\mathbf{X}_a$ and $\mathbf{X}_{b+ab}$ and to remove the possible structural noise included in model $\mathbf{X}_{abg}$, different criteria are required for each part. To retain the signal, we adopt the number of components that explain a high quantity of variation of $\mathbf{X}_a$ and $\mathbf{X}_{b+ab}$ submodels, fixed at more than 75% of the variation in each case. To remove structural noise, the approach must be different. In this case, we intend to eliminate only structural noise, whereas the previous strategy potentially also removes random noise. If structural noise is present, this will be captured by the PCA of $\mathbf{X}_{abg}$ and there will be a number of eigenvalues of the covariance matrix of $\mathbf{X}_{abg}$ that are noticeably higher than the rest. Note that in the case of $\mathbf{X}_{abg}$ is only random, all these eigenvalues would be approximately equal. Therefore, the criterion in this case will be the selection of components with noticeably high eigenvalues. These eigenvalues can be identified as those that satisfy (2.6),

$$\lambda_k \geqslant \beta \frac{\sum_{k=1}^{\mathrm{rank}(X_{abg})} \lambda_k}{\mathrm{rank}(X_{abg})}, \quad \beta > 1. \tag{2.6}$$

In this work, we have taken $\beta = 2$ (see supplementary material, available at *Biostatistics* online, for a formal justification of the criterion and relevance of the number of components selection).

2.3 *Data sets*

2.3.1 *Simulated data.* ARSyN has been evaluated in 11 different scenarios in order to resemble situations with different types and magnitude of noise. These scenarios are simulated as independent TCM experiments (6 cases) and also longitudinal TCM experiments (5 cases). A detailed description of these simulated data has been included in the supplementary material, available at *Biostatistics* online, and a brief description in Table 1.

2.3.2 *Experimental data.* Two real transcriptomics examples were chosen to evaluate the biological consistency of the proposed method. The first was the toxicogenomic study by Heijne *and others* (2003), which investigates the effect of the hepatotoxicant bromobenzene in rats. This data set consists of 3 time points (6, 12, and 48 h after administration of the drug), 5 experimental groups (1 untreated group; 1 placebo, corn oil; and 3 different doses of bromobenzene: low, medium, and high), and 2665 genes. The second was a stress study in plants which investigates the transcriptional response to 3 different abiotic stressors (salt, cold, and heat) in the potato using the National Science Foundation (NSF) 10k potato array (Rensink *and others*, 2005). This data set has 4 series (1 control and 3 types of stress: heat, salt, and cold), 3 time points, 3 replicates per experimental condition, and 9993 genes.

2.4 *The evaluation approach*

The general strategy for evaluating the performance of ARSyN was to apply a statistical method for TCM data (maSigPro, EDGE, and *timecourse*) to the different data sets before and after ARSyN filtering and to compare results in terms of feature selection. The maSigPro approach (Conesa *and others*, 2006) is a regression-based method that uses a polynomial model to fit gene expression dynamics and dummy variables to differentiate between experimental groups or series. *timecourse* is based on the empirical Bayes procedure to study one- and two-sample longitudinal series (Tai and Speed, 2006), and recently, the method has been adapted to multiple conditions (Tai and Speed, 2009). Finally, EDGE (Storey *and others*, 2005) uses B-splines–based models to analyze both independent and longitudinal data. These methods are described in the supplementary material, available at *Biostatistics* online. In the case of simulated data, we have used sensitivity (true positives detected/real true positives) and specificity (true negatives

Table 1. *Description of simulated data sets*

	(a) Independent data		(b) Longitudinal data
Time points	3		5
Experimental groups	3		3
Total number of genes	10 000		10 000
Changing genes	410 classified in 5 patterns		500
Scenarios	6 with different quantity of structural and random noise. Always a dye effect		5 with different types of structural noise
	Structural noise	Random noise	Type of structural noise
Scenario 1	None	Low	None
Scenario 2	None	High	Horizontal effect: genes
Scenario 3	Low	Low	Vertical effect: arrays
Scenario 4	Low	High	Several arrays altered partially
Scenario 5	High	Low	Dye effect
Scenario 6	High	High	—

detected/real true negatives) as measures of quality. A good selection of genes is obtained when both measures are close to 1. In the case of experimental data, as the truly differentially expressed genes are unknown, these metrics cannot be used. Instead, we have applied a functional enrichment (FE) analysis (Al-Shahrour *and others*, 2007) to evaluate the biological consistency of the results. FE assesses whether specific cellular functions are overrepresented within a set of significant genes and is a well-established methodology for interpreting and evaluating transcriptomic data. Additionally, we have compared our results to those obtained by current batch-removal methods. We have chosen ComBat (Johnson *and others*, 2007) and SVA (Leek and Storey, 2007) which were recently recommended by Luo *and others* (2010) and Leek *and others* (2010), respectively. Both methods were applied to the simulated studies, and ComBat was also applied to the toxicogenomic experimental data. ComBat could not be applied to NSF potato stress data because it has not a defined batch effect. SVA was not applied to real data sets to simplify results as the M(TCM) methods used in this paper cannot be directly applied with SVA.

3. Results

3.1 *Simulation studies*

Several data sets were generated for each one of the analysis scenarios designed in each simulation study. In order to highlight the balance between signal and noise introduced in each analysis scenario, we show in Table 2 the amount of variation simulated and explained in each ASCA submodel. In general, we observe that residual variation increased as higher noise was modeled. The percentage of explained variance in $\mathbf{X}_a$ and $\mathbf{X}_{b+ab}$ submodels was more or less constant across scenarios, whereas the explained variance in the $\mathbf{X}_{abg}$ submodel was strongly associated to the presence of structural noise. This result confirms the ability of the $\mathbf{X}_{abg}$ submodel in capturing the systematic noise embedded in the data. Next, we simulated 50 data sets for each scenario, obtained filtered data by ARSyN, and applied maSigPro and *timecourse* to all the data sets. Only 10 simulations were run with EDGE as this software is only accessible from a graphical user interface and could not be integrated in high-performing scripting pipelines. However, the stability of the results in all cases made this simplification acceptable.

Table 2. *Percentage of variation simulated, and explained with ASCA, in each submodel for different scenarios from one of the simulated independent (a) and one of the longitudinal (b) data sets*

	% Variation			Number of	% Explained		
Scenario	$\mathbf{X}_a$	$\mathbf{X}_{b+ab}$	$\mathbf{X}_{abg}$	components	$\mathbf{X}_a$	$\mathbf{X}_{b+ab}$	$\mathbf{X}_{abg}$
(a) Independent data							
1	33.6	39.8	26.6	1, 2, 0	89.7	89	0
2	19.4	28.2	52.3	1, 3, 0	83.5	71.7	0
3	12.3	18.9	68.8	1, 2, 1	89.7	89.9	79.8
4	10.5	18.3	71.2	1, 3, 1	83.7	79.5	57.1
5	3	7.8	89.2	1, 2, 1	89.6	83.5	92.6
6	3.3	8.6	88.1	1, 2, 1	83.7	75	86
(b) Longitudinal data							
1	10.6	26.1	63.3	3, 8, 0	75.7	82.9	0
2	11	26	63	3, 8, 0	75.6	82.6	0
3	8.8	21.9	69.3	3, 8, 1	76	82.7	27
4	9.6	23.2	67.2	3, 8, 1	76	82.6	18
5	5.8	14.2	80	3, 8, 1	75.6	82.4	57.2

Figure 1 shows the sensitivity average with the original and filtered data for each method and type of time course data. The details of this analysis are shown in the supplementary material, available at *Biostatistics* online, in terms of false positives, false negatives, sensitivity, specificity averages, and their correspondent confidence intervals. Performance analysis indicated that specificity was high and similar in all cases and that differences were revealed by the sensitivity indicator. We explain these differences in detail below.

***maSigPro*.** Performance indicators showed that, in all scenarios, the selection of genes by applying maSigPro to the ARSyN filtered data was equal or better than that obtained by applying maSigPro to the original data. In scenarios where no systematic noise was introduced (Scenarios 1 and 2 of independent and longitudinal data studies), maSigPro was efficient with respect to both the original and the filtered data, and ARSyN did not affect the good performance of the statistical method. On the other hand, in scenarios with high structural noise, ARSyN clearly improved sensitivity, while specificity was unaffected.

***timecourse*.** The analysis of the simulated independent data sets revealed that, in scenarios without structural noise, performance indicators were similar with and without ARSyN filtering. In Scenarios 4 and 5, a slight improvement in sensitivity was observed when ARSyN was applied, while Scenarios 3 and 6 clearly showed the higher sensitivity of the filtered data. In contrast, no significant performance differences between original and ARSyN data were observed when longitudinal data were analyzed by *timecourse*.

***EDGE*.** The study of the independent data with the EDGE methodology showed that the number of false negatives was 100 in many cases. These were largely genes simulated with a pattern of parallel gene-expression profiles among series, which are hard to detect by this method. In general, we observed that

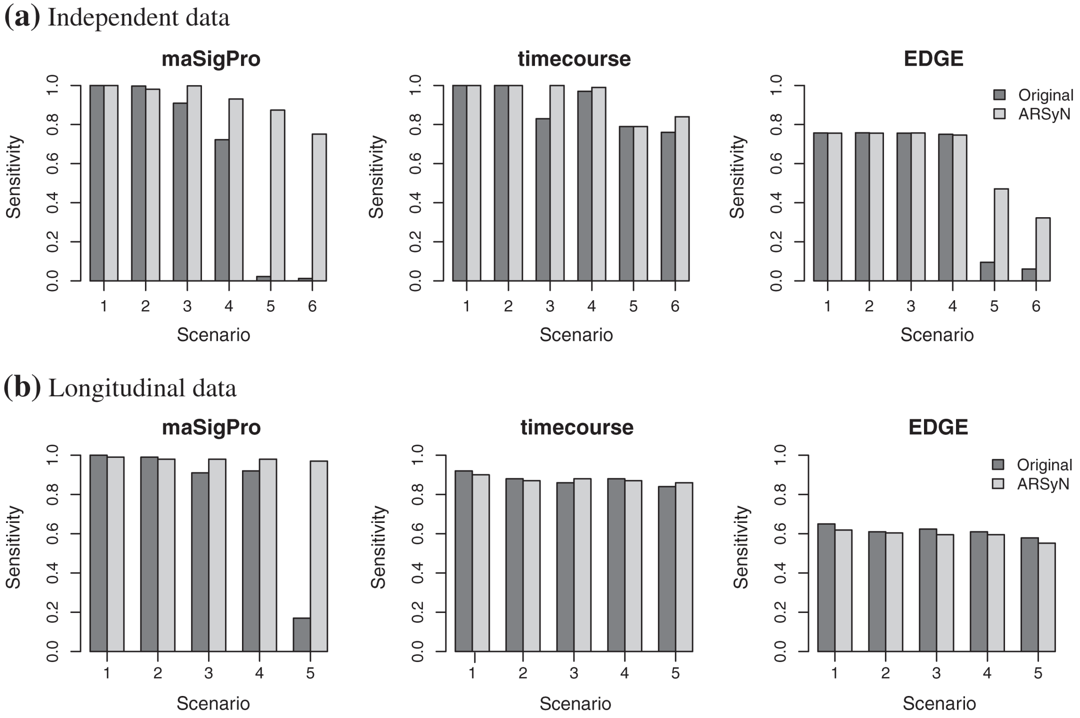

Fig. 1. Sensitivity plot. The height of the bars represents the average sensitivity obtained in 50 simulations (10 for EDGE) with the original and filtered data of (a) independent simulated data and (b) longitudinal simulated data.

sensitivity of EDGE was lower than in the other 2 methods. Preprocessing of data by ARSyN improves detection capacity in some scenarios, although sensitivity values continued to be low.

Considered as a whole, the simulation study revealed that ARSyN is an efficient preprocessing technique for improving the detection of differentially expressed genes in scenarios with high structural noise but has no effect when noise is low. We have also demonstrated that the combination of ARSyN and maSigPro is the analysis strategy with the best overall performance.

Comparison with other filtering techniques. ARSyN was compared to 2 other noise-removal methodologies: ComBat and SVA. ComBat outputs, as ARSyN, a filtered data set that can be further analyzed by TCM methods. However, ComBat cannot be applied in situations where the batch is not identified, which we consider a limitation with respect to ARSyN. When the batch is known (Scenarios 3–6), ComBat rendered a higher number of false positives, in comparison of ARSyN, whereas the number of false negatives slightly decreased, except in Scenarios 5 and 6 of independent data set (Supplementary Tables 4 and 5 available at *Biostatistics* online). In contrast, SVA performed poorer on independent data whereas produced similar results as ComBat and ARSyN in the longitudinal study (Supplementary Table 6 available at *Biostatistics* online). However, a major disadvantage of SVA is that it does not give a filtered data matrix, so TCM methods cannot be directly applied. Altogether these results point to ARSyN as a more robust and versatile solution for noise removal than other approaches.

3.2 *Toxicogenomics data set*

ASCA analysis of this data set decomposed data into 3 submodels: “time,” “treatment + treatment × time,” and “residuals.” After component selection by ARSyN, 1, 5, and 2 components, respectively, were retained for each submodel. This component selection explained 75% of time variation, 78% of treatment plus interaction variation, and 48% of residual variation.

Exploratory analysis of the 2 first PCs of the original data set revealed a considerable batch effect (Figure 2(a)) that was removed with ARSyN (Figure 2(b)). The origin of this structural bias was identified as a dye effect since the experiment had a dye-swap design and the dye used in each array was known. This effect can also be treated by simply centering genes with the corresponding dye average (Figure 2(c)) as in Conesa *and others* (2006), where maSigPro was applied to the analysis of this data set (note that this basic centering preprocessing is the comparing scenario in this toxicogenomics example). ComBat filtering was also effective in removing the dye bias (Figure 2(d)). Interestingly, ComBat preprocessing resulted in very similar PC plots as dye centering (Figure 2(c)). From this analysis, we concluded that ARSyN filtering and also other batch-removal approaches removed the dye bias from the data and revealed the differences between the high doses of bromobenzene and the remaining doses. Furthermore, ARSyN preprocessing resulted in an increase of the number of genes that obtained low p-values in maSigPro and EDGE analysis (Figure 3), which is consistent with a general removal of noise from the data. Gene selection obtained with the different methods and comparisons are shown in the Supplementary Figure 3 available at *Biostatistics* online.

Finally, we investigated the gain in biological interpretability of the filtered data by analyzing the number and types of enriched gene ontology (GO) terms in the selected genes in comparison to those obtained from unfiltered data. In general, the number of enriched GO terms and the size of the term within the pool of selected genes were greater in ARSyN filtered data than without filter (Supplementary Table 7 available at *Biostatistics* online), indicating that the noise-removal procedure enhanced the detection of coordinated gene sets. Furthermore, GO functions revealed by the filtered data were related to processes of the cellular detoxification response (Heijne *and others*, 2003). For example, “glutathione transferase activity” (found in maSigPro–ARSyN and EDGE–ARSyN analysis) is the major cellular activity that targets bromobenzene for degradation, whereas “heme binding” (maSigPro–ARSyN results) refers to redox enzymes involved in this process. Similarly, “nitric oxide signal transduction” (enriched in

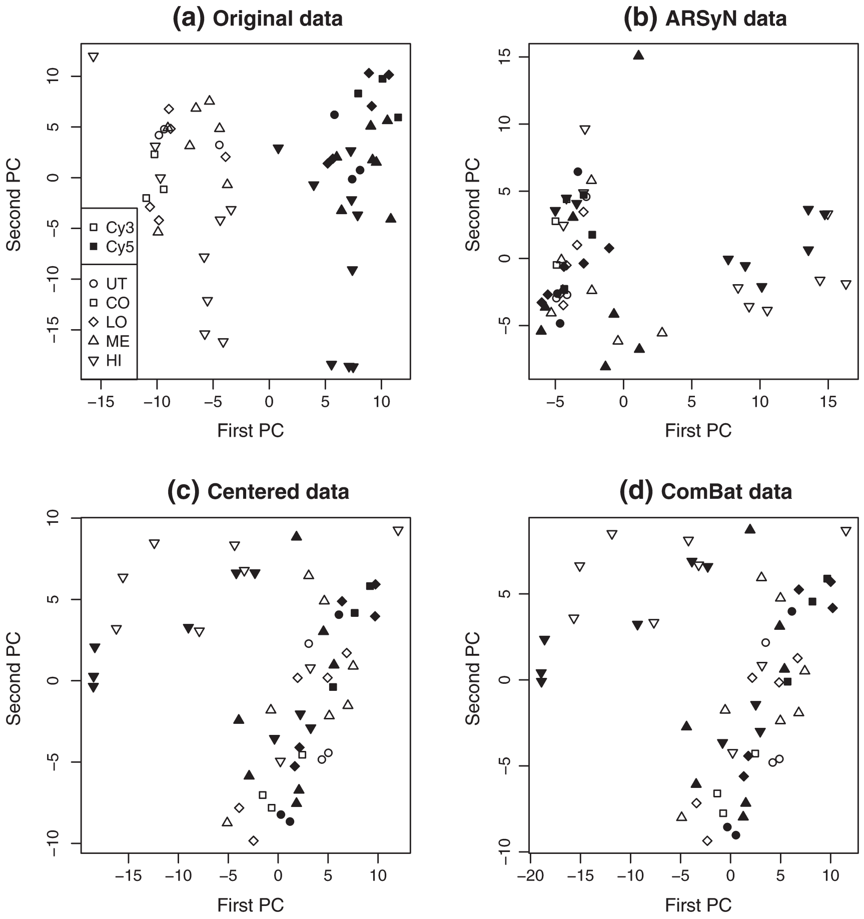

Fig. 2. PCA of (a) original data, (b) ARSyN filtered data, (c) centered data by dye, and (d) ComBat filtered data. Cy3 and Cy5 are green and red dyes. Experimental groups: untreated (UT), corn oil (CO), low (LO), medium (ME), and high (HI) doses of bromobenzene.

timecourse–ARSyN) points to a detoxification mechanism associated to the response to xenobiotic compounds (Morán *and others*, 2010; Farina *and others*, 2011). These results show the biological relevance of the new genes uncovered by the filtering procedure. ComBat preprocessing did not add new relevant functional conclusions to the analysis of these data.

3.3 *NSF potato stress data set*

The ARSyN analysis for this data set resulted in a model with 1, 3, and 2 components for submodel time, treatment + treatment × time, and residuals, respectively. This component selection explains 100% of time variation, 80% of treatment plus interaction variation, and 28% of residual variation. Gene selection obtained with the different methods and comparisons are shown in the Supplementary Figure 3, available at *Biostatistics* online.

When considering the functional analysis (Supplementary Table 8 available at *Biostatistics* online), again, the number of enriched GO terms obtained by maSigPro and timecourse analysis was higher when

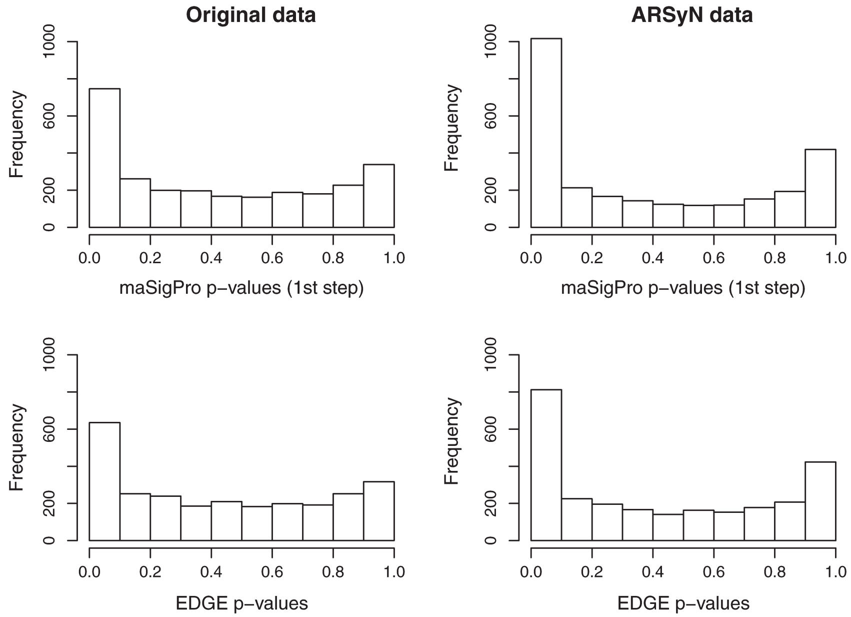

Fig. 3. Distribution of p-values obtained by maSigPro (first step) and EDGE on the toxicogenomics data set before and after ARSyN filtering. ARSyN filtering increases the number of genes with low p-values which is consistent with a decrease in noise levels.

data were preprocessed by ARSyN and also the specificity of the identified functions which included hormone "signalling cascades," "diverse enzymatic binding activities," and "defined metabolic functions." Notably, EDGE analysis on these data did not result in a relevant number of significant results, regardless of the filtering option.

4. Discussion and conclusions

This paper describes the methodology ARSyN that uses a model-based multivariate projection technique such as ASCA for the removal of systematic biases in microarray data. The rational of the methodology is the extraction of the relevant shared behavior and the identification and removal of structured noise that cannot be associated with the experimental factors included in the design of the transcriptomics study. This structural noise is habitually referred to as the "batch" effect. It is a result of dye, lab, experimentalist, etc., factors and can affect the data of the related arrays both globally and locally. ASCA uses ANOVA to identify signals associated with experimental factors and PCA to separate structured and random variation in these signals. By removing the nonstructured part of the experimental factor signals (the noise within the signal) and the structured variation of the ANOVA errors (the signal of the noise) from the original data set, we create a filtered data set that is enriched in the information of interest and retains only the random noise needed for inferential analysis. This procedure offers the advantage of not requiring previous knowledge of the nature of the "batch effect." Any possible structural noise is identified in the signal of the residuals of the ASCA model.

The efficacy of this filter was analyzed in 2 simulation studies in which independent and longitudinal data, respectively, were mimicked. The proposed ARSyN method targets the systematic noise in gene expression data sets, different types—systematic and random—and magnitude—high or low—of noise were introduced into the synthetic data. Additionally, we assessed whether or not the filter was generally valid, irrespective of the inference methodology used to identify differentially expressed genes. Therefore, we tested the filter with 3 available methods for the analysis of TCM data that follow very different statistical strategies: maSigPro applies polynomial regression, timecourse is based on empirical Bayes, and EDGE uses B-splines to model the dynamics of gene expression.

The results showed that ARSyN significantly improves gene selection when a high quantity of structural noise is present and has no effect when only random noise affects the expression signals. Although this pattern was observed with each of the 3 statistical methodologies employed, maSigPro was clearly the method on which ARSyN had the greatest impact and which yielded the best end results. Sensitivity improvement with timecourse and EDGE was not as pronounced as with maSigPro, and the amount of differential expression detected when these 2 methodologies were applied to ARSyN data never reached the sensitivity levels obtained by the maSigPro analysis. This result can be explained by the nature of the maSigPro method, a univariate gene-by-gene regression that considers a normal distribution of the error. Given that the ARSyN filter exploits the co-expression of genes through the PCA on the estimates of the ANOVA parameters, the synergy with the inferential approach is likely to be maximal. However, both timecourse and EDGE use empirical methods to determine the statistical significance of statistics, which implies the consideration of possible structural noise in all data. On the other hand, timecourse employs shrinking covariance estimates and therefore takes into account the relationships within expression values. In this way, these methods consider aspects that are also considered by the proposed filter, and therefore the effect obtained is expected be of a lower magnitude than that observed with maSigPro. Even so, the ARSyN filter improves the sensitivity of timecourse and EDGE in some scenarios. We hypothesize that this is related to the more refined treatment of variation by ASCA as it imposes an ANOVA model prior to component analysis. This decomposition allows for an experimental factor–focused analysis of covariance that is more efficient than the design-blind analysis of correlation structures that characterize timecourse and EDGE methods. Finally, it should be mentioned that EDGE is based on B-splines models, which work well with series of more than 10 time points; the present study was restricted to short series of 3–5 time points, which may be the reason for the poor results obtained with this technique.

When applied to experimental data sets, preprocessing by ARSyN improved the significance of the statistical tests, the identification of the transcriptionally regulated biological processes, and the number of significant genes contained in selected functional categories. The better performance of the ARSyN data in the FE analysis could not be simply the consequence of an increase in the number of genes declared significant as this occurred with maSigPro but not with timecourse. We argue that ARSyN preprocessing, which modifies gene expression according to the correlation structure of the data set, helps to reveal the coordinated regulation of genes in the same functional class, thereby improving the detection of enriched functions. Additionally, ARSyN would eliminate noisy or poorly correlated genes that reduce statistical power of FE analysis.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

Funding

Spanish MICINN Project (BIO2008-04368-E and DPI2008-06880-C03-03/DPI).

References

Agudo, M., Perez-Marin, M. C., Lonngren, U., Sobrado, P., Conesa, A., Cánovas, I., Salinas-Navarro, M., Miralles-Imperial, J., Hallbk, F. and Vidal-Sanz, M. (2008). Time course profiling of the retinal transcriptome after optic nerve transection and optic nerve crush source. *Molecular Vision* **14**, 1050–1063.

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D. and Dopazo, J. (2007). FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Research* **35**, W91–W96.

Alter, O., Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10101–10106.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M. and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–114.

Brumós, J., Colmenero-Flores, J. M., Conesa, A., Izquierdo, P., Sánchez, G., Iglesias, D. J., López-Climent, M. F., Gómez-Cadenas, A. and Talón, M. (2009). Membrane transporters and carbon metabolism implicated in chloride homeostasis differentiate salt stress responses in tolerant and sensitive citrus rootstocks. *Functional and Integrative Genomics* **9**, 293–309.

Conesa, A., Nueda, M. J., Ferrer, A. and Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray. *Bioinformatics* **22**, 1096–1102.

Do, J. H. and Choi, D. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. *Molecules and Cells* **22**, 254–261.

Farina, M., Aschner, M. and Rocha, J. B. (2011). Oxidative stress in MeHg-induced neurotoxicity. *Toxicology and Applied Pharmacology,* doi:10.1016/j.taap.2011.05.001.

Heijne, W. H. M., Stierum, R., Slijper, M., van P. J. Bladeren and van B. Ommen. (2003). Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach. *Biochemical Pharmacology* **65**, 857–875.

Jansen, J. J., Hoefsloot, H. C. J., Timmerman, M. E., van der Greef, J., Westerhuis, J. and Smilde, A. K. (2005). ASCA: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics* **19**, 469–481.

Johnson, W. E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127.

Leek, J. T., Scharpf, R. B., Corrada, H., Simcha, D., Langmead, B., Johnson, E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733–739.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics* **3**, 1724–1735.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H. *and others* (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The Pharmacogenomics Journal* **10**, 278–291.

MAQC-Consortium. (2010). The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28**, 827–838.

Morán, J. M., Ortiz-Ortiz, M. A.,Ruiz-Mesa L. M. and Fuentes J. M. (2010). Nitric oxide in paraquat-mediated toxicity: a review. *Journal of Biochemical and Molecular Toxicology* **24**, 402–409.

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C. J., Smilde, A. K., Talón, M. and Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics* **23**, 1792–1800.

Rensink, W. A., Iobst, S., Hart, A., Stegalkina, S., Liu, J. and Buell, C. R. (2005). Gene expression profiling of potato responses to cold heat and salt stress. *Functional and Integrative Genomics* **5**, 201–207.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R. J. A. N., van der Greef, J. and Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **21**, 3043–3048.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12837–12842.

Svendsen, C., Owen, J., Kille, P., Wren, J., Jonker, M. J., Headley, B. A., Morgan, A. J., Blaxter, M., Sturzenbaum, S. R., Hankard, P. K. *and others* (2008). Comparative transcriptomic responses to chronic cadmium fluoranthene and atrazine exposure in lumbricus rubellus. *Environmental Science and Technology* **42**, 4208–4214.

Tai, Y. C. and Speed, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics* **34**, 2387–2412.

Tai, Y. C. and Speed, T. P. (2009). On gene ranking using replicated microarray time course data. *Biometrics* **65**, 40–51.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

[*Received December 16, 2010; revised July 12, 2011; accepted for publication October 11, 2011*]