Top (2011) 19:121–129
DOI 10.1007/s11750-009-0111-6

ORIGINAL PAPER

Using game theory to detect genes involved in Autism Spectrum Disorder

Francisco J. Esteban · Dennis P. Wall

Received: 27 February 2009 / Accepted: 2 July 2009 / Published online: 24 July 2009

Abstract Microarray technology is a current approach for detecting alterations in the expression of thousands of genes simultaneously between two different biological conditions. Genes of interest are selected on the basis of an obtained p-value, and, thus, the list of candidates may vary depending on the data processing steps taken and statistical tests applied. Using standard approaches to the statistical analysis of microarray data from individuals with Autism Spectrum Disorder (ASD), several genes have been proposed as candidates. However, the lists of genes detected as differentially regulated in published mRNA expression analyses of Autism often do not overlap, owed at least in part to (i) the multifactorial nature of ASD, (ii) the high inter-individual variability of the gene expression in ASD cases, and (iii) differences in the statistical analysis approaches applied. Game theory recently has been proposed as a new method to detect the relevance of gene expression in different conditions. In this work, we test the ability of Game theory, specifically the Shapley value, to detect candidate ASD genes using a microarray experiment in which only a few genes can be detected as dysregulated using conventional statistical approaches. Our results showed that coalitional games significantly increased the power to iden-

F.J. Esteban was supported, as visiting Professor at the Center for Biomedical Informatics (Harvard Medical School), by the program *Salvador de Madariaga* (Ref: PR2007-0255), Ministerio de Educación y Ciencia (Spain).
This work was aided in part by funding from the National Science Foundation under Grant No. 0543480 awarded to D.P. Wall.

F.J. Esteban (✉)
Dept. Experimental Biology, University of Jaén, 23071 Jaén, Spain
e-mail: festeban@ujaen.es

F.J. Esteban
e-mail: francisco_esteban@hms.harvard.edu

D.P. Wall
Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
e-mail: dennis_wall@hms.harvard.edu

tify candidates. A further functional analysis demonstrated that groups of these genes were associated with biological functions and disorders previously shown to be related to ASD.

Keywords Cooperative games · Gene expression · Microarrays

Mathematics Subject Classification (2000) 91A12 · 91A80 · 92B15 · 92C40

1 Introduction

A microarray is a small analytical device that allows genomic analysis with speed and precision (Lee and Saeed 2007). The design and capacity of microarrays allows the analysis of the entire genome in a single experiment. The expression of tens of thousands of genes can be simultaneously explored, thus allowing a complete genome-wide analysis of the alteration of the gene expression taking place between two different biological conditions. This technology is widely used in biomedical research to detect altered gene expression of particular genes in a given disease when compared to healthy controls. Autism Spectrum Disorder (ASD) is a complex multigenic disorder that has a broad range of behaviors, the genetic component of which has proven particularly challenging to standard microarray-based analysis. Different genes, involved in diverse biological functions and molecular interactions, have been experimentally demonstrated as dysregulated in ASD cases. In the few microarray studies done to date, more than 900 genes have been detected as differentially expressed in ASD; however, only a few ($\sim$1%) have been replicated across studies replication (Abrahams and Geschwind 2008). Here we question the extent to which this lack of reproducibility is due to true biological difference or to difference in data processing approaches, which have been shown to cause large discrepancies in the results derived even by a single microarray experiment (Shi et al. 2008).

Various statistical tests have been used to measure differentially gene expression from microarrays experiments (Draghici 2003). In many of these approaches, genes are ranked according to their independent p-values, but p-values are not necessarily indicative of biological signal. Small p-values may actually be unrelated to the biological condition under study, and large, presumably insignificant p-values, may be tied genes that play important mechanistic roles. Statistical processing of p-values, such as multiple test correction procedures, are often applied to correct for false discovery and loss of signal, but in most cases these corrections are too conservative to detect biologically informative signal. Motivated by the need for more powerful approaches to signal detection, a method for gene expression analysis based on game theory has been recently proposed (Moretti et al. 2007). The main advantage of this approach is the computed numerical index, called the Shapley value, which represents the relevance of each gene under a certain condition while simultaneous accounting for the expression behaviors of the other genes under the same condition, a method that, in combination with statistics, has been demonstrated to be useful for differential gene expression data analysis. Thus, the aim of this work is to apply this methodology to determine the gene relevance on an ASD microarray experiment in which only a few genes could be detected using standard statistical approaches.

2 Methods

2.1 Data source and processing

We have analyzed a subset from experiment GSE6575 downloaded from Gene Expression Omnibus (GEO) (Gregg et al. 2008; http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6575). This subset consisted of 17 samples of Autistic patients without regression and 12 healthy children from the general population. Statistical analysis was performed with Bioconductor (http://www.bioconductor.org/) in R (http://www.r-project.org/). In order to remove all the possible sources of variation of a nonbiological origin between arrays, data were normalized using the RMA normalization function implemented in the Bioconductor affy package (http://www.bioconductor.org/packages/release/bioc/html/affy.html). From the total number of probes ($n = 54675$), we then selected those probes with gene symbol identity ($n = 45590$), and the probes corresponding to the same gene were averaged (final number of probes 20680). Genes showing significant differences between groups were identified using a Student's t-test (raw p-value), and resultant p-values were corrected for multiple hypothesis testing using the False Discovery Rate (FDR) as implemented in the multtest package (http://www.bioconductor.org/packages/release/bioc/html/multtest.html). Even though a raw p-value < 0.01 was associated to 496 genes, only two of them survived to the multiple hypothesis correction (FDR value ≤ 0.05).

2.2 Game theory method

To obtain the gene relevance on an ASD microarray data-set, we used the coalitional game method as previously described (Moretti et al. 2008). Briefly, a *coalitional* game is a pair (N, v), where N denotes a finite set of players, and $v : 2^N \to \mathbb{R}$ is the *characteristic function* with $v(\emptyset) = 0$. A group of players $T \subseteq N$ is called a *coalition* and the real value $v(T)$ is the worth of coalition T. A *solution* for a class of coalitional games is a function ψ that assigns a vector number $\psi(v) \in \mathbb{R}^N$ to each coalitional game in the class; a well-known solution for coalitional games is the Shapley value, which assigns to each player his average marginal contribution over all the possible orderings, i.e., permutations of players. Formally, given a coalitional game (N, v), for each player $i \in N$ the Shapley value $\phi(v)$ is defined by

$$\phi_i = \frac{1}{n!} \sum_{\pi} v\big(P(\pi, i) \cup \{i\}\big) - v\big(P(\pi, i)\big), \tag{1}$$

where π is a permutation of players, $P(\pi, i)$ is the set of players that precede player i in the permutation π, and n is the cardinality of N.

Moretti et al. (2007) introduced the definition of microarray game as a coalitional game (N, w) with the objective to stress the relevance ("sufficiency") of groups of genes in relation to a specific condition. Let $N = \{1, \ldots, n\}$ be a set of genes. On a single microarray experiment on N, a sufficient requirement to realize in a coalition $S \subseteq N$ the association between a condition and an expression property is that all the genes showing that expression property belong to coalition S (*sufficiency principle*). Different expression properties for genes might be considered like, e.g., under- or over-expression, strong variation, abnormal expression, etc. A group of genes $S \subseteq N$ which realizes the association between the expression property and the condition on a single array is called a *winning coalition* for that array.

We refer to a Boolean matrix $\boldsymbol{B} \in \{0, 1\}^{n \times k}$, where $k \geq 1$ is the number of arrays, and where the Boolean values 0–1 represent two complementary expression properties, for example, the normal expression (coded by 0) and the over-expression (coded by 1). Let $\boldsymbol{B}_{.j}$ be the jth column of $\boldsymbol{B}$, we define the *support* of $\boldsymbol{B}_{.j}$, denoted by $sp(\boldsymbol{B}_{.j})$, as the set $sp(\boldsymbol{B}_{.j}) = \{i \in \{1, \ldots, n\}$ such that $\boldsymbol{B}_{ij} = 1\}$. The microarray game corresponding to $\boldsymbol{B}$ is defined as the coalitional game (N, w), where $w : 2^N \rightarrow \mathbb{R}^+$ is such that $w(T)$ is the rate of occurrences of coalition T as a winning coalition, i.e., as a superset of the supports in the Boolean matrix $\boldsymbol{B}$; in formula, $w(T)$, for each $T \in 2^N \backslash \{\emptyset\}$, is defined as the value

$$w(T) = \frac{c(\Theta(T))}{k} \tag{2}$$

where $c(\Theta(T))$ is the cardinality of the set $\Theta(T) = \{j \in K$ such that $sp(\boldsymbol{B}_{.j}) \subseteq T$, $sp(\boldsymbol{B}_{.j}) \neq \emptyset\}$, with the set of arrays $K = \{1, \ldots, k\}$ and $v(\emptyset) = 0$. Since it is computationally too expensive to calculate the Shapley value $\phi(w)$ of game (N, w) according to relation (1), Moretti et al. (2008) introduced an easy way to calculate $\phi(w)$ for whatever microarray game (N, w). We have adapted the script from these authors (Moretti et al. 2008) run under R (http://www.r-project.org/).

2.3 Data processing for game theory analysis

A final matrix including the expression levels of 496 genes (raw p-value < 0.01) and 29 samples (17 autistics without regression -*AnR*-, 12 controls -*C*-) was generated from the original data as stated above (see Sect. 2.1). Then, in order to discriminate over-regulated levels of gene expression with respect to expressions measured in control children, each continuous value in the vector $\mathbf{X}_{i.} = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{i29})$ which was equal to or greater than Mean$[\mathbf{X}_{i.}^C]$ + Stdev$[\mathbf{X}_{i.}^C]$ was coded as 1, or as 0 otherwise. Consequently, a Boolean matrix $\mathbf{B}^+$ with 496 rows and 29 columns and with values $\{0, 1\}$ was generated from $\mathbf{X}$. Separately, a procedure aimed to discriminate under-regulated levels of gene expression with respect to expressions measured also in control children was applied. Each continuous value in the vector $\mathbf{X}_{i.} = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{i29})$ which was equal to or smaller than Mean$[\mathbf{X}_{i.}^C]$ − Stdev$[\mathbf{X}_{i.}^C]$ was coded as 1, or as 0 otherwise. Consequently, a Boolean matrix $\mathbf{B}^-$ with 496 rows and 29 columns with values $\{0, 1\}$ was also generated from $\mathbf{X}$. According to the distinction between control and *AnR* samples, the Boolean matrix $\mathbf{B}^+$ was split into two different Boolean matrices $\mathbf{B}^{C+}$ and $\mathbf{B}^{AnR+}$, and the Boolean matrix $\mathbf{B}^-$ was split into two other Boolean matrices $\mathbf{B}^{C-}$ and $\mathbf{B}^{AnR-}$. By relation (2), from the Boolean matrix $\mathbf{B}^{AnR+}$ the microarray game w^{AnR+} is defined, and, in a similar way, the microarray game w^{AnR-} from the Boolean matrix $\mathbf{B}^{AnR-}$ is defined; the corresponding Shapley values also were calculated. In order to remove genes with high Shapley values that could be attributed to chance, we ran a resampling procedure over our observed Shapley values (Comparative Analysis of Shapley value; shortly, CASh), similar to the procedure used in Moretti et al. (2008). Specifically, we generated 1000 bootstrap matrices and for each calculated the corresponding unadjusted p-values.

As an additional filtration step in each microarray game, genes showing both a p-value ≤ 0.05 from the Boostrap resampling and the absolute difference of the Shapley value $\triangle\phi_i^+ = |\phi_i(w^{C+}) - \phi_i(w^{AnR+})|$ (or $\triangle\phi_i^- = |\phi_i(w^{C-}) - \phi_i(w^{AnR-})|$) greater than the mean plus the standard deviation for each group, were selected for functional analysis.

2.4 Functional analysis

We detected enrichment of biological function and overrepresentation of disease(s) among gene terms using Ingenuity Pathways Analysis (IPA; http://www.ingenuity.com/). IPA functional analysis is a powerful tool enabling to associate biological functions and diseases to experimental results, including differentially expressed genes selected from microarray experiments. IPA works by leveraging the complex biological interactions that are stored in the manually curated Ingenuity's knowledge base, and it is designed to organize the biological information allowing one to gain different level overviews of the biology that is associated with the data.

3 Results

When conventional processing and statistical approaches were used to analyze the Autistic expression experiment studied here, only two genes survived multiple hypothesis test correction (FDR ≤ 0.05): C21orf23 and MAPK10 (both down-regulated in the *AnR* group). Because 496 genes showed a statistical significant difference (raw p-value < 0.01) between control and *AnR* groups, we selected these 496 genes as the most probable candidates to be mechanistically involved in ASD, and thus most likely to reveal biologically meaningful signal when analyzed using a microarray game approach.

Using the microarray game and the CASh approach, 85 over- and 84 down-regulated genes provided a p-value ≤ 0.05 (after rounding the final p-value to two decimal places). On the other hand, 109 and 72 genes showed $\triangle\phi_i$ greater than the mean plus the standard deviation for the microarray games w^{AnR+} and w^{AnR-}, respectively. Taken together, the intersection of these gene filtering methods gave rise to 78 (65 over- and 13 down-regulated) candidates genes related to ASD (Table 1). Figure 1 shows the plots of the CASh p-values versus $\triangle\phi_i$ corresponding to the 496

Table 1 Relevant genes in ASD we detected by a coalitional microarray game

Gene symbols
ADCY7, AK3L1, APEH, ARIH2, ATP2A2, ATP5A1, ATP6AP2, BYSL, C21orf23, C9orf114, CD3Z, CHAF1B, CHST2, CHTF18, CKLFSF3, CLIC3, CREB1, CX3CR1, CXorf12, DCTN5, DDX56, DST, ECRG4, FLJ23441, FLJ38984, FXN, GFM1, GLE1L, GPSN2, GTF3C4, GUSB, GZMB, HERPUD1, HNRPA1, IL18R1, IL2RB, ITGB2, KBTBD4, KCNIP4, KIR3DL1, KIR3DL3, KSP37, LARS2, LECT2, LHFP, LOC129138, LOC143381, LOC284409, LOC285749, LOC339448, LOC440350, LRP1B, MCM3, MORF4L1, NCALD, NKG7, NMT2, NONO, PCDH17, PDPR, PMM1, PRF1, RFC2, RNASEH2A, SAMD3, SH2D1B, SMYD3, SPON2, TRA@, UNC5A, UTP11L, VMD2L2, WDR70, YIPF5, YWHAE, ZNF330, ZNF618, ZW10.

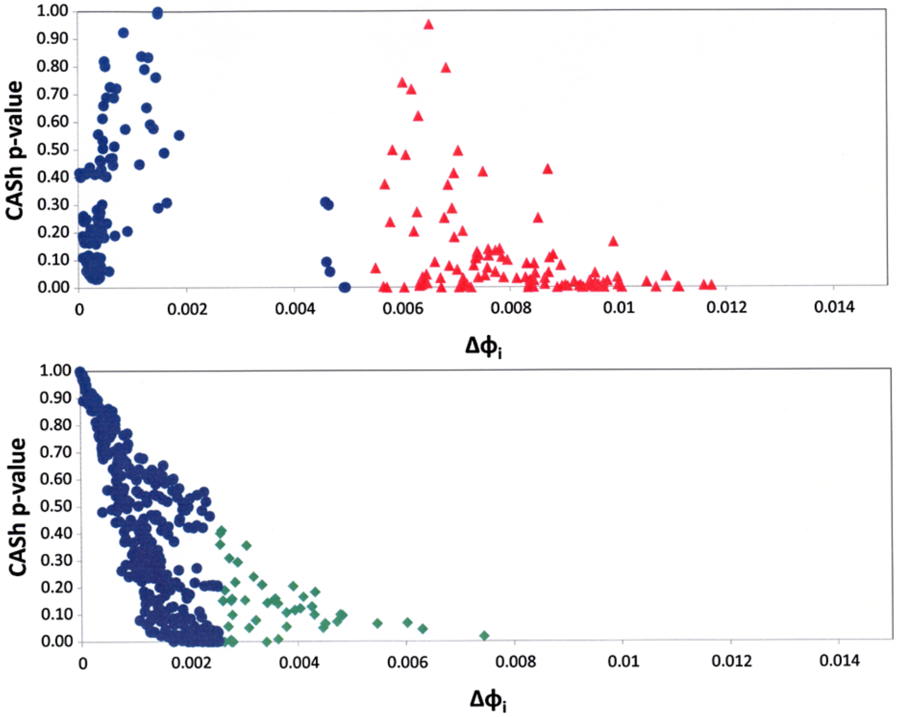

Fig. 1 Plot of the p-values obtained using the CASh approach versus the absolute difference of the Shapley values $\triangle\phi_i^+ = |\phi_i(w^{C+}) - \phi_i(w^{AnR+})|$ (*top*) and $\triangle\phi_i^- = |\phi_i(w^{C-}) - \phi_i(w^{AnR-})|$) (*bottom*). *Triangles* represent the 109 genes selected in game w^{AnR+}; *diamonds* represent the 72 genes selected in game w^{AnR-}. *Dashed line*: cutoff at a p-value of 0.05

initial candidates; as can be observed, a larger number of over-regulated genes is selected for a given p-value when compared to down-regulated genes. Taking into account only the 78 final candidates selected (Fig. 2), a stable Shapley value for increasing CASh p-values can be detected. Finally, and as shown in Table 1, C21orf23, one of the two genes satisfying the multiple hypothesis test correction in the conventional statistical approach, was also detected by the game theory strategy.

IPA analysis detected 55 and 16 non overlapping functional categories with significant FDR-adjusted p-values ≤ 0.05 and 0.01 (after rounding the final p-value to two decimal places), respectively (Table 2). Furthermore, IPA analysis identified 19 genes in our set of 78 candidates as significantly overrepresented in the general category of Neurological Disease.

4 Discussion

Using a classical approach for microarray data processing and statistical analysis, we were able to detect only two genes as differentially regulated in individuals with ASD when compared with controls. In an effort to boost the biologically meaningful

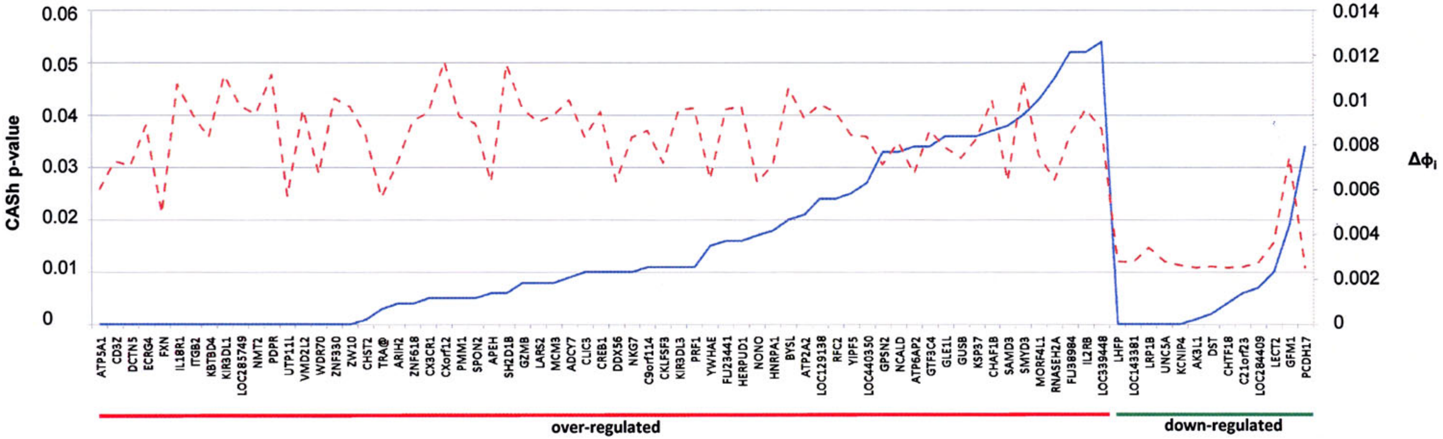

Fig. 2 Plot of the p-values obtained using the CASh approach (*solid line*) and the absolute difference of the Shapley values (*dashed line*) for the 78 candidate ASD genes

Table 2 Functional annotations with an FDR ≤ 0.01

Category	FDR	Gene symbols
Cell death	0.00201	CD247, CX3CR1, GZMB, ITGB2, KIR3DL1, PRF1
Genetic disorder	0.00201	ADCY7, ARIH2, ATP2A2, ATP5A1, ATP6AP2, BYSL, CD247, CHAF1B, CLIC3, CREB1, CX3CR1, DST, FXN, GFM1, GLE1, GUSB, GZMB, HNRNPA1, IL18R1, IL2RB, ITGB2, KIR3DL1, LARS2, MCM3, NKG7, NONO, PRF1, RNASEH2A, TRA@, YWHAE
Antimicrobial response	0.00454	GZMB, PRF1
Cell-mediated immune response	0.00718	CD247, GZMB, IL2RB, ITGB2, KIR3DL1, PRF1
Cell-to-cell signaling and interaction	0.00718	CD247, GZMB, IL2RB, ITGB2, KIR3DL1, PRF1
Hematological system development and function	0.00718	IL2RB, ITGB2, KIR3DL1, PRF1
Immune cell trafficking	0.00718	IL2RB, ITGB2, KIR3DL1, PRF1
Neurological disease	0.00718	ADCY7, ARIH2, ATP5A1, ATP6AP2, CHAF1B, CLIC3, CREB1, DST, FXN, GFM1, GLE1, IL18R1, IL2RB, ITGB2, NKG7, PRF1, RNASEH2A, TRA@, YWHAE
Cellular assembly and organization	0.00754	GZMB, PRF1
Cellular compromise	0.00754	GZMB, PRF1
Cellular development	0.00976	CD247, CREB1, IL2RB, ITGB2, TRA@
Hematopoiesis	0.00976	CD247, CREB1, IL2RB, ITGB2, TRA@
Lymphoid tissue structure and development	0.00976	CD247, CREB1, IL2RB, ITGB2, TRA@
Hematological disease	0.0129	GZMB, PRF1
Immunological disease	0.0129	GZMB, PRF1
Cell morphology	0.0135	GZMB, PRF1

signal, we have applied game theory to the same dataset in order to ascertain whether this methodology could be useful to reveal group of genes candidates to be involved in ASD. To be restricted to the most suitable candidates, the microarray game was carried out only on those genes having a raw p-value < 0.01 using the Student's t-test. The final results showed a relevant combination of 78 genes, including one of the two genes that survived multiple test correction using standard analysis of microarray data.

As stated in Moretti et al. (2008), the different results obtained depending on the approach are due to the different underlying criteria for selecting genes. The parametric t-test simply tests whether or not two independent populations (control and *AnR*) have different mean values of their expression, and yields a p-value that indicates how likely the result is a chance event. The Shapley value that we have used here measures the contribution of genes together with the contributions of all other

Table 3 Genes related to ASD in our microarray game and in Gregg et al. (2008)

Gene symbols
CLIC3, CX3CR1, DCTN5, GFM1, GZMB, IL18R1, IL2RB, ITGB2, KIR3DL3, KSP37, NCALD, NKG7, NMT2, PRF1, SAMD3, SH2D1B, SPON2, WDR70, YIPF5, ZNF330

genes producing signal in the experiment. Using this global approach together with the CASh method for measuring significant gene contributions, only those genes with the highest relevance—calculated as the average marginal contribution over all possible permutations, not only analyzing single differences in expression—are detected. As previously stated (Moretti et al. 2008), when comparing CASh p-values versus $\triangle\phi_i$, the number of genes selected for a given p-value in the microarray game w^{AnR+} is larger than those selected in the microarray game w^{AnR-}, and the comparison of these parameters only for the 78 candidates selected shows a stable Shapley value for increasing CASh p-values. Similar to previous applications of game theory on whole genomic gene expression experiments (Moretti et al. 2007, 2008), we have demonstrated here that the approach can significantly boost biologically informative signal.

The biological significance of the relevant candidates, and thus the convenience of the microarray game approach, can be established taking into account the functional results we have obtained, where 19 genes were significantly related to neurological diseases. Interestingly, and according to this functional criteria for selecting candidates genes involved in ASD, we have recently provided a novel picture of autism from the perspective of related neurological disorders (Wall et al. 2009). In addition, 20 out of the 78 candidates (Table 3) have also been previously related to ASD when the same data-set was analyzed, with a different processing method, by others (Gregg et al. 2008).

In conclusion, the joint expression behavior of genes reflected by the Shapley value and the CASh method in a microarray game is a useful approach to obtain biological knowledge in the complex molecular basis of ASD.

References

Abrahams BS, Geschwind DH (2008) Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet 9:341–355

Draghici S (2003) Data analysis tools for DNA microarrays. Chapman&Hall/CRC Press, UK

Gregg JP, Lit L, Baron CA, Hertz-Picciotto I, Walker W, Davis RA, Croen LA, Ozonoff S, Hansen R, Pessah IN, Sharp FR (2008) Gene expression changes in children with autism. Genomics 91:22–29

Lee NH, Saeed AI (2007) Microarrays: an overview. Methods Mol Biol 353:265–300

Moretti S, Patrone F, Bonassi S (2007) The class of Microarray games and the relevance index for genes. TOP 15:265–280

Moretti S, van Leeuwen D, Gmuender H, Bonassi S, van Delft J, Kleinjans J, Patrone F, Merlo DF (2008) Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. BMC Bioinf 9:361

Shi L, Perkins RG, Fang H, Tong W (2008) Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol 19:10–18

Wall DP, Esteban FJ, Deluca TF, Huyck M, Monaghan T, Velez de Mendizabal N, Goñí J, Kohane IS (2009) Comparative analysis of neurological disorders focuses genome-wide search for autism genes. Genomics 93:120–129