Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 326 (2003) 1067–1072
COMPTES RENDUS BIOLOGIES

Molecular biology and genetics

GeneNote: whole genome expression profiles in normal human tissues

Orit Shmueli [a,*], Shirley Horn-Saban [b], Vered Chalifa-Caspi [b], Michael Shmoish [a], Ron Ophir [b], Hila Benjamin-Rodrig [a,c], Marilyn Safran [b], Eytan Domany [c], Doron Lancet [a]

[a] *Departments of Molecular Genetics, The Weizmann Institute of Science, 76100 Rehovot, Israel*
[b] *Biological Services, The Weizmann Institute of Science, 76100 Rehovot, Israel*
[c] *Physics of Complex Systems, The Weizmann Institute of Science, 76100 Rehovot, Israel*

Received 16 September 2003; accepted 23 September 2003

Presented by François Gros

Abstract

A novel data set, GeneNote (**Gene No**rmal **T**issue **E**xpression), was produced to portray complete gene expression profiles in healthy human tissues using the Affymetrix GeneChip HG-U95 set, which includes 62 839 probe-sets. The hybridization intensities of two replicates were processed and analyzed to yield the complete transcriptome for twelve human tissues. Abundant novel information on tissue specificity provides a baseline for past and future expression studies related to diseases. The data is posted in GeneNote (http://genecards.weizmann.ac.il/genenote/), a widely used compendium of human genes (http://bioinfo.weizmann.ac.il/genecards). ***To cite this article: O. Shmueli et al., C. R. Biologies 326 (2003).***

Résumé

GeneNote : profils d'expression complets dans des tissus humains normaux. Un nouveau jeu de données, GeneNote **(Gene No**rmal **T**issue **E**xpression), a été produit pour décrire les profils d'expression complets des gènes dans les tissus humains sains, en utilisant les puces GeneChip HG-U95 d'Affymetrix, qui comprennent 62 839 jeux de sondes. Les intensités d'hybridation de deux réplicats ont été traités et analysés pour décrire le transcriptome complet pour douze tissus humains. Les données nouvelles abondantes sur la spécificité tissulaire fournissent une base de référence pour les études d'expression passées et futures en relation avec les maladies. Les données sont accessibles dans GeneNote (http://genecards.weizmann.ac.il/genenote/) une compilaton de gènes humains largement utilisée (http://bioinfo.weizmann.ac.il/genecards). ***Pour citer cet article : O. Shmueli et al., C. R. Biologies 326 (2003).***

Keywords: microarray; transcriptome; expression profile; normal human tissues; GeneCards

Mots-clés : microréseau ; transcriptome ; profil d'expression ; tissus humains normaux ; GeneCards

* Corresponding author.
E-mail address: orit.shmueli@weizmann.ac.il (O. Shmueli).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2003.09.012

1. Introduction

The human body orchestrates gene expression by co-regulating genes whose products function together. Many challenges are posed due to the vast amount of genomic and expression data that has to be divided into functional biological groups using biological, statistical and computational tools. Many past studies centered on comparisons of diseased to healthy tissues, and were limited to a subset of human genes.

High-density oligonucleotide arrays enable highly parallel and comprehensive studies of gene expression. The transcription patterns produced, known as the expression profile, depict the subset of mRNA synthesized in a certain cell or tissue. At its most fundamental level, the expression profile describes, in a quantitative way, which genes are expressed in a particular tissue. More sophisticated issues such as novel gene functional characterization, gene identification in biological pathways, genetic variation analysis or identification of drug targets could be surveyed by using bioinformatics tools such as cluster analysis ([1,2] and self organizing-maps [3,4]).

The construction of gene expression databases is a high priority of today's research community. Such databases, closely integrated with other types of genomic information, promise not only to enhance our understanding of many fundamental biological processes, but also to accelerate drug discovery and lead to customized diagnosis and treatment of disease [5,6]. An example generated by our own group is presented the recent releases of GeneCards, which contain expression data both from array experiments and from 'electronic Northern' analyses [7].

As a significant extension of previous relevant efforts, we describe a whole-genome repertoire of expression profiles in twelve normal human tissues. Previously, only studies that compare a single diseased tissue with a healthy one were preformed [8–10]. Other surveys that were done on healthy human tissues were limited to one array type (i.e. HG-U95A [11], Hu6800 [12]). Here, for the first time, we present the expression analysis of the full complement of more than 60 000 gene and EST representations in 12 normal human tissues. It is shown that the additional, less characterized genes harbor important information, and that a whole-genome expression analysis is essential for generating comprehensive transcription analyses.

2. Materials and methods

PolyA+ RNA samples from twelve normal human tissues were purchased from Clontech (Palo Alto, CA). This collection of major human tissues includes: Bone marrow (catalog number: 6573-1), brain (6516-1), heart (6533-1), kidney (6538-1), liver (6510-1), lung (6524-1), pancreas (6539-1), prostate (6546-1), skeletal muscle (6541-1), spinal cord (6593-1), spleen (6542-1) and thymus (6536-1).

Preparation and hybridization of cRNA were done according to the manufacture's instructions [13]. Sufficient hybridization cocktail solution for five independent hybridizations reactions (i.e. the full set HG-U95A-E) was prepared. The hybridization reactions of all the arrays in the set were carried out simultaneously. The mRNA from each tissue was reacted in duplicates against the full human Affymetrix arrays set (HG-U95A-E) to yield two sets of results. The 3′/5′ signal ratios of GAPDH were always below the value of 3 as expected for a fine labeling reaction.

Arrays were analyzed and expression value was calculated for each gene by using Microarray Suit (MAS) version 5.0 software (Affymetrix, Santa Clara, CA). Expression values for each gene, called *signal*, were calculated using the MAS 5.0 software with its default parameter settings. Scaling was not done via a MAS 5.0 option. Instead, we normalized our data as follows: the intensities of each array were $\log_{10}$ transformed and scaled to a constant reference value (global normalization). This reference value was the mean of all log intensities in all of the tissues. Present calls percentages are presented in Fig. 2 for all samples according to the array type (see also Results and Discussion).

3. Results and discussion

Whole-genome gene expression profiles were generated from the normalized signal values. The set of 12 tissue expression values for a given gene was defined as its *tissue vector*. Normalization was used to allow a meaningful comparison among different tissues.

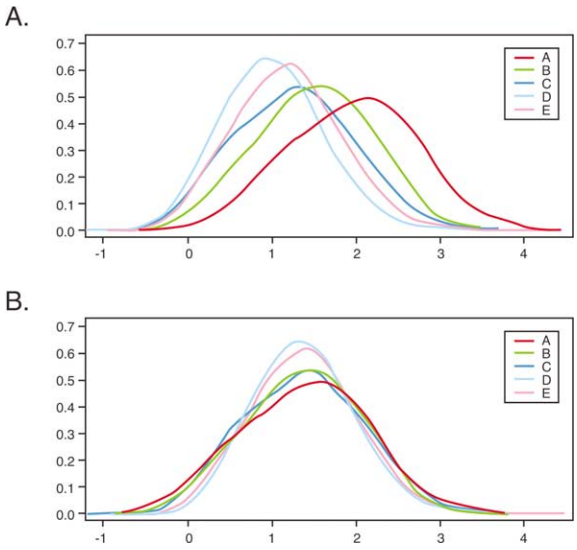

Fig. 1. Intensity distribution in the different arrays set in heart sample. Density plots of one heart sample before and after normalization are presented. Panel **A** presents the density plot for $\log_{10}$ raw signals while panel **B** presents the normalized signals.

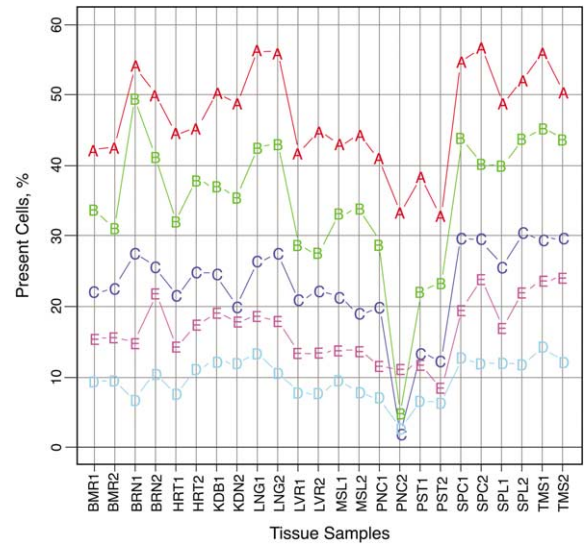

Fig. 2. Percent of present calls in all samples according to the array type. 'Present' call percentages were taken from the MAS 5.0 report files and are presented for the various array types in all of the samples. The A–E letters represent the array type. The following are the tissue samples shortcuts: **BMR** for bone marrow, **BRN** for brain, **HRT** for heart, **KDN** for kidney, **LNG** for lung, **LVR** for liver, **MSL** for muscle, **PNC** for pancreas, **PST** for prostate, **SPC** for spinal cord, **SPL** for spleen, and **TMS** for thymus. The numbers attached to the sample name represent the replicate number.

Fig. 1 shows a comparison of the intensities distribution of one heart sample for the five different arrays, before and after normalization. In this density plot, it is observed that the highest intensities are in arrays A and B, while arrays C, D and E produce considerably lower intensities. These results are similar to the earlier observations by Bakay et al. [8]. Aspects of the decreased array intensities are also presented in Fig. 2 which displays the percentage of 'present' calls along the samples for each array type. For example, it is well observed that the percentage of 'present' calls in array D is consistently below 10%. This could be the result of intrinsic hybridization differences. It could also be due to the average quality of the probe-sets, since most of the well-studied genes are found on array A while arrays B to E are constructed with increasing propensities of less well-characterized ESTs (Expressed Sequence Tags).

Tissue specific genes are supposed to have a significant role in tissue functionality. Hence, we examined the counts of such genes in the different tissue samples (Fig. 3). One representative sample for each tissue was chosen, and tissue specific genes were determined on the basis of the 'absent–present' calls of the Affymetrix MAS 5.0 software. A similar criterion was used in the past for housekeepings genes selection [14]. In this paper, we define tissue specificity as having a 'present' call in only one tissue. It is observed that brain and thymus had the highest number of tissue specific genes, almost twice the number observed for other tissues. The propensity of tissue specific genes is not much lower in arrays B–E when compared to array A, suggesting that important information also resides within the former. However, one should also take into consideration that the amount of 'present' calls is reduced on arrays C–E (Fig. 2). Consequently, the increased effect of tissue specificity may be an artifact of the array quality and design.

The MAS 5.0 package is one of the most commonly used software tools for analyzing high-density micrroarray results. However its use of 'present–absent' calls is a controversial issue in the literature [15–17]. To address this, we have begun to develop alternative methods utilizing a normalized entropy-related Tissue Specificity Index (TSI), computed directly from the tissue vectors. This algorithm is currently being evalu-

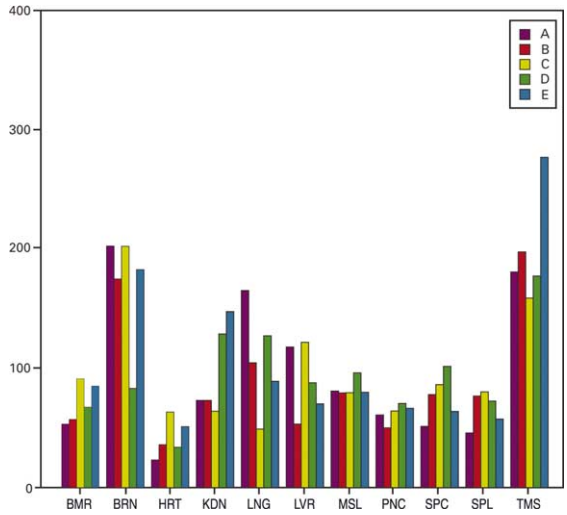

Fig. 3. Tissue specific genes in all of the tissues sorted by the array set. Tissue specific genes were found based on the 'absent–present' call of the Affymetrix MAS 5.0 software. Tissue specificity was defined as having a 'present' call in only one tissue. One sample for each tissue was chosen for the search. The tissue specific genes distribution on the arrays set, HG-U95A-E is shown for all tissues.

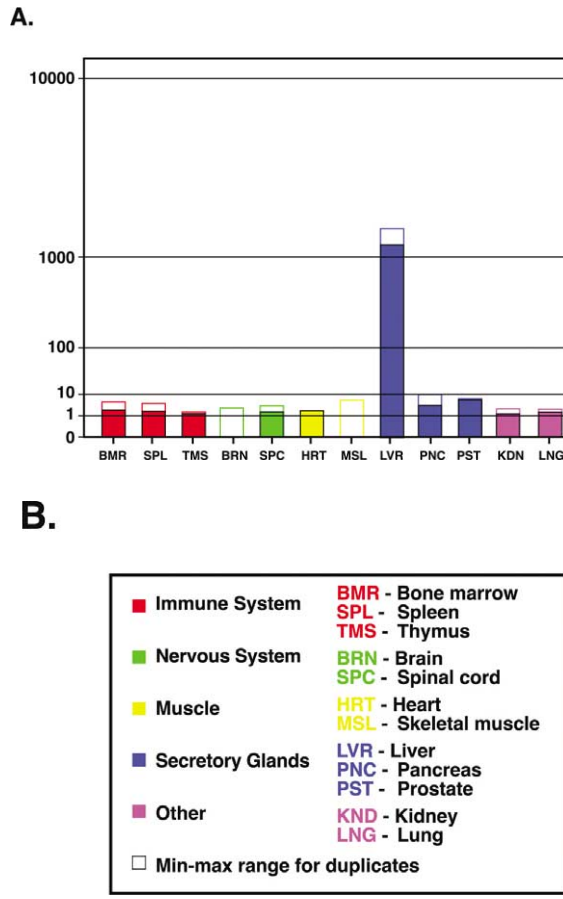

Fig. 4. Tissue vector for the gene APCS [21]. Tissue vector for the gene APCS was calculated from the normalized signals and is presented in a graphical way (panel **A**). Tissues were grouped according to their origin and the groups colored accordingly (e.g., nerve tissues in green). The range between the lower and higher measurements was represented by a white box above the colored minimal measurement bar. The graph is presented on the y-axis with a special root scale [7] $Y = X^{(1/B)}$ where $B = \log_2 10$. Panel **B** presents the tissue vector colored map.

ated as compared to other methods (O. Shmueli et al., manuscript in preparation).

Criticism has also been directed towards the subtraction of the mismatch probes (MM) from the perfect match probes (PM) creating PM-MM values as an underlying computation for assessing background levels and signal computations. Several methods have been developed to evaluate the background level of intensity above which the value truly reflects the gene expression value [15,16,18–20]. Some of these models fit log(PM-Background), instead of PM-MM, though in low-intensity regions the current models are still very questionable [16]. Consequently, the validity of the zone of low-intensity probes, which is close to the background level, remains undetermined.

The tissue vectors for each gene were drawn on a root scale representation [7] as exemplified in Fig. 4 for the gene APCS (Amyloid P-component, serum) [21]. APCS is a precursor for amyloid component P which is found in basement membrane and associated with amyloid deposits. A root scale enables one to visualize several orders of magnitude, similar to a logarithmic scale, but preserves some advantage of the linear scale, i.e. a differential increase with the different orders of magnitude. This affords an effective view-at-a-glance of the tissue vectors.

A major aim of this work is to enable prediction of the function of novel genes based on their expression profiles. It is expected that genes that display similar expression patterns are functionally related, since they are co-regulated under all of the developmental conditions [22,23]. On the basis of the expression profile similarities, ten tissues were ordered in a tissue correlation matrix (Fig. 5). It is clearly seen that

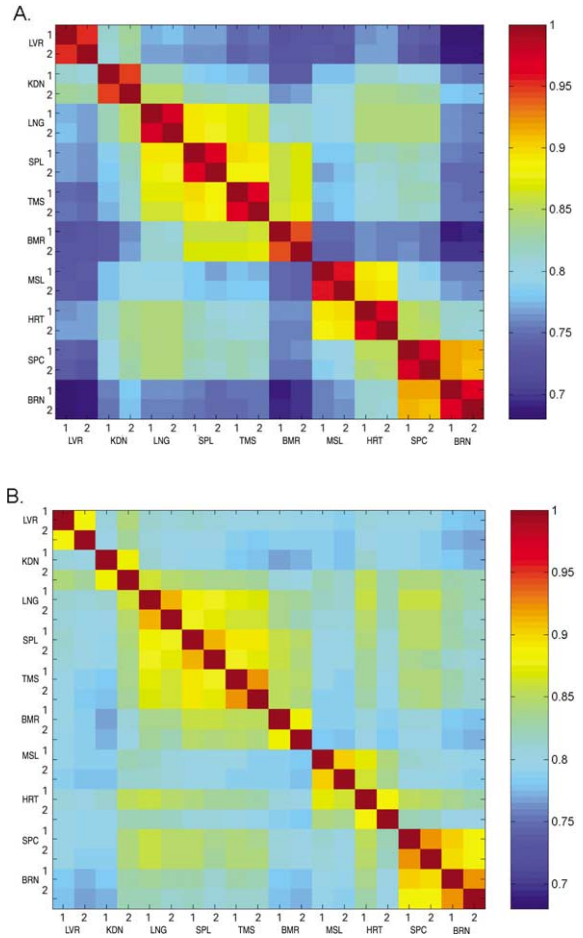

Fig. 5. Tissue correlation matrix. Tissue vectors of the normalized signals (log transformed and scaled) were used to calculate correlations between twenty tissue samples. The correlations were calculated using the Pearson correlation coefficient and presented in a tissue vector correlation matrix. Different correlation levels are presented according to the color scale on the right. Panel **A** presents the tissue vector correlation matrix for array A, while panel **B** presents arrays B–E.

array A (Fig. 5A) and arrays B–E (Fig. 5B) show the same pattern of correlations. The high correlations found between tissue replicates is an indication of the validity of the results [24,25]. In addition, the correlation matrix revealed three groups of closely related tissues. The first group with high inter-tissue correlation includes brain and spinal-cord, the second is heart and muscle and the last includes thymus, spleen, bone-marrow and lung. This is in agreement with the expected biological origin of the tissues, since they represent the nervous system, muscle, and immune system respectively.

All expression data, raw as well as normalized, have been stored in the GeneNote database (http://genecards.weizmann.ac.il/genenote/). GeneNote tissue vectors are provided in GeneCards (http://bioinfo.weizmann.ac.il/genecards) for all genes that could be explicitly assciated via the Affymetrix annotation (currently ∼20 000 genes).

The preliminary results presented above provide a clear indication of the importance of surveying gene expression in an as broad as possible a gene set. The combination of comprehensive experimental data gathering, computer analysis, and database display provide relevant and useful tools for the transcriptome community.

Acknowledgements

The work described in this paper was supported by the Abraham and Judith Goldwasser foundation. Doron Lancet is the incumbent of the Ralph and Lois Silver Chair in Human Genomics. Eytan Domany is the incumbent of the Henry J. Leir Professorial Chair.

References

[1] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl Acad. Sci. USA 95 (1998) 14863–14868.
[2] G. Getz, E. Levine, E. Domany, Coupled two-way clustering analysis of gene microarray data, Proc. Natl Acad. Sci. USA 97 (2000) 12079–12084.
[3] M. Kurella, L.L. Hsiao, T. Yoshida, J.D. Randall, G. Chow, S.S. Sarang, R.V. Jensen, S.R. Gullans, DNA microarray analysis of complex biologic processes, J. Am. Soc. Nephrol. 12 (2001) 1072–1078.
[4] A. Sturn, J. Quackenbush, Z. Trajanoski, Genesis: cluster analysis of microarray data, Bioinformatics 18 (2002) 207–208.
[5] G. Zweiger, Knowledge discovery in gene-expression-microarray data: mining the information output of the genome, Trends Biotechnol. 17 (1999) 429–436.
[6] T. Strachan, M. Abitbol, D. Davidson, J.S. Beckmann, A new dimension for the human genome project: towards comprehensive expression maps, Nat. Genet. 16 (1997) 126–132.
[7] M. Safran, V. Chalifa-Caspi, O. Shmueli, T. Olender, M. Lapidot, N. Rosen, M. Shmoish, Y. Peter, G. Glusman, E. Feldmesser, A. Adato, I. Peter, M. Khen, T. Atarot, Y. Groner, D. Lancet, Human Gene-Centric Databases at the Weizmann

Institute of Science: GeneCards, UDB, CroW 21 and HORDE, Nucleic Acids Res. 31 (2003) 142–146.
[8] M. Bakay, P. Zhao, J. Chen, E.P. Hoffman, A web-accessible complete transcriptome of normal human and DMD muscle, Neuromuscul. Disord. Suppl. 12 (1) (2002) S125–S141.
[9] T.J. Mariani, V. Budhraja, B.H. Mecham, C.C. Gu, M.A. Watson, Y. Sadovsky, A variable fold-change threshold determines significance for expression microarrays, FASEB J. (2002).
[10] C.A. Iacobuzio-Donahue, A. Maitra, G.L. Shen-Ong, T. van Heek, R. Ashfaq, R. Meyer, K. Walter, K. Berg, M.A. Hollingsworth, J.L. Cameron, C.J. Yeo, S.E. Kern, M. Goggins, R.H. Hruban, Discovery of novel tumor markers of pancreatic cancer using global gene expression technology, Am. J. Pathol. 160 (2002) 1239–1249.
[11] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes, Proc. Natl Acad. Sci. USA 99 (2002) 4465–4470.
[12] P.M. Haverty, Z. Weng, N.L. Best, K.R. Auerbach, L.L. Hsiao, R.V. Jensen, S.R. Gullans, HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues, Nucleic Acids Res. 30 (2002) 214–217.
[13] Affymetrix, Inc., GeneChip Expression Analysis, 2001.
[14] L.L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z. Weng, G.L. Mutter, M.P. Frosch, M.E. Macdonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, G. Stephanopoulos, S.R. Gullans, A compendium of gene expression in normal human tissues, Physiol. Genomics 7 (2001) 97–104.
[15] C. Li, W.H. Wong, Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, Proc. Natl Acad. Sci. USA 98 (2001) 31–36.
[16] R. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data, 2002.
[17] R. Sasik, E. Calvo, J. Corbeil, Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model, Bioinformatics 18 (2002) 1633–1640.
[18] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.
[19] E. Hubbell, W.M. Liu, R. Mei, Robust estimators for expression analysis, Bioinformatics 18 (2002) 1585–1592.
[20] W.M. Liu, R. Mei, X. Di, T.B. Ryder, E. Hubbell, S. Dee, T.A. Webster, C.A. Harrington, M.H. Ho, J. Baid, S.P. Smeekens, Analysis of high density expression microarrays with signed-rank call algorithms, Bioinformatics 18 (2002) 1593–1599.
[21] E.C. Mantzouranis, S.B. Dowton, A.S. Whitehead, M.D. Edge, G.A. Bruns, H.R. Colten, Human serum amyloid P component. cDNA isolation, complete sequence of pre-serum amyloid P component, and localization of the gene to chromosome 1, J. Biol. Chem. 260 (1985) 7752–7756.
[22] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S.H. Friend, Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109–126.
[23] J.L. DeRisi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.
[24] M. Bakay, Y.W. Chen, R. Borup, P. Zhao, K. Nagaraju, E.P. Hoffman, Sources of variability and effect of experimental approach on expression profiling data interpretation, BMC Bioinformatics 3 (2002) 4.
[25] K.R. Coombes, W.E. Highsmith, T.A. Krogmann, K.A. Baggerly, D.N. Stivers, L.V. Abruzzo, Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays, J. Comput. Biol. 9 (2002) 655–669.