Towards Optimal Microarray Universal Reference Sample Designs: An In-Silico Optimization Approach

George Potamias, Sofia Kaforou, Dimitris Kafetzopoulos

Towards Optimal Microarray Universal Reference Sample Designs: An *In-Silico* Optimization Approach

George Potamias[1], Sofia Kaforou[2], Dimitris Kafetzopoulos[2]

[1] Institute of Computer Science, [2] Institute of Molecular Biology & Biotechnology,
Foundation for Research & Technology – Hellas (FORTH), N. Plastira 100,
GR - 70013 Heraklion, Crete, Greece
potamias@ics.forth.gr, kafetzo@imbb.forth.gr

Abstract. Assessment of the reliability of microarray experiments as well as their cross-laboratory/platform reproducibility rise as the major need. A critical challenge concerns the design of optimal Universal Reference RNA (URR) samples in order to maximize detectable spots in two-color/channel microarray experiments, decrease the variability of microarray data, and finally ease the comparison between heterogeneous microarray datasets. Towards this target we devised and present an *in-silico* (binary) optimization process the solutions of which present optimal URR sample designs. Setting a cut-off threshold value over which a gene is considered as detectably expressed enables the process. Experimental results are quite encouraging and the related discussion highlights the suitability and flexibility of the approach.

Keywords: Bioinformatics, microarrays, universal reference sample.

1 Introduction

Scientific experimental science is founded on the 'right' design of experiments where some common-sense principles should apply [1]. The observation is more than critical in the set-up and the complex design of microarray experiments. With more than fifteen years of research and experimentation in the field the real challenge as well as the actual tendency is moving towards the comparison of different microarray experiments originated from different platforms, from different laboratories, and from different designs. In this context, assessing the reliability of microarray experiments as well as their cross-laboratory/platform reproducibility rise as the major need.

Measuring reliable fluorescence intensities with microarrays is not a straightforward task. The basic problem rises from the variability in microarray spot geometry and the quantity of DNA deposited at each spot. So, absolute fluorescence intensity cannot be used as a reliable measure of RNA level. However, if two RNA samples are differentially labeled and co-hybridized to spots on the same microarray, the ratio of their signal intensities accurately reports the relative quantity of RNA targets in both samples – the so called, *two-color (two-channel)* hybridization set-up (2CH). Obtaining reliable and reproducible 2CH gene expression data is critically important for understanding the biological significance of perturbations made on a

cellular system. Moreover, and most critical, *maximizing detectable spots* on the reference image channel also decreases the variability of microarray data allowing for reliable detection of smaller differential gene expression changes. [2].

With the 2CH microarray technology there are two alternatives for the comparison between different sample conditions, C_1 vs. C_2: (i) the '*loop*' design where, direct comparisons between a series of arrays are performed with a sample of type C_1 in one channel and one of type C_2 in the other channel for all other arrays, and the comparison of samples in circular or multiple pair-wise fashion - it is a useful approach when a small number of samples is to be analyzed; and (ii) the *Universal* RNA *Reference* (URR) sample, C – devised by a pool of diverse (in terms of their ability to express different families of genes) cell-lines. With the URR approach a series of hybridizations are carried out with an experimental sample of type C_1 / C_2 in one channel and the URR sample in the other. Then differences in gene-expression between C_1 and C_2 are acquired by comparing the ratios C_1/C vs. C_2/C [3]. The hybridization of each (experimental) sample with a URR *mixture* serves as a common denominator between different microarray hybridizations [4]. With the employment of a reproducible URR the quantization of gene-expression levels is more reliable and offers a valuable tool for monitoring and controlling intra- and inter-experimental variation. The URR design has several practical advantages over the loop design: (i) it extends easily to other experiments if the common URR is preserved, (ii) is robust to multiple chip failures; and (iii) reduces incidence of laboratory mistakes as each (experimental) sample is handled the same way [1]. In addition, the URR design facilitates the normalization and the subsequent comparison between heterogeneous microarray data sets [5]. Typically molecular scientists and experimentalists prepare their own array-specific reference samples e.g., genomic DNA [6, 7], a mixture of clones spotted on the arrays [8], as well as complementary to every microarray spot short oligomers [9]. Many groups utilize a mixture of cell-lines for their URR (e.g., Stratagene's human URR [5]) or, clone-vectors [2]. In most of the cases the utilized URR samples are not reproducible between labs and may provide detectable signal for a low percentage of the microarray spots [10]. Subsequently, they do not provide good expression *coverage*, i.e., a high proportion of genes being adequately hybridized and expressed in order to avoid spots with zero denominators in C_1/C vs. C_2/C ratio comparisons, a fact that would force discarding those spots from the analyses [3].

It is natural to assume, and this is actual the case, that as more cell-lines are utilized to prepare the URR sample mixture then, the biggest the chance to get dilution and saturation effects. So, it is of fundamental importance to determine an *optimum number* of cell-lines so that such effects are avoided as much as possible. This is the target of the present work. In particular, our aim was to devise a general methodology that guides the determination of an optimal ("the less the better") number of cell-lines the mixture of which is capable to hybridize and express the highest possible number of array genes. The paper is organized as follows: section 2 outlines the state-of-the-art research on the field; in section 3 we introduce and formally present our optimization approach, and in section 4 we present and discuss the results of the performed experiments; finally, in section 6 we conclude and point to future research plans.

2 Background to the URR sample design

The MAQC (MicroArray Quality Control) project[2] systematically evaluated the reliability of microarray technology by profiling two human URR samples using different microarray and QPCR (Quantitative PCR) platforms [11]: Stratagene's (acquired by Agilent Technologies[3]) human Universal Reference RNA™ - derived from a mixture of cell lines [5], and Ambion's human brain URR - pooled from multiple donors and several brain regions[4].

According to Sterrenburg et al., a common reference for DNA microarrays would consist of a mix of the products being spotted on the array. Their polymerase chain reaction (PCR) reference was generated by pooling a fraction of all amplified microarray probes prior to printing. A very low number of spots could not be analyzed because reference, as opposed to target, did not give a significant signal [8]. Yang et al., suggested the use of a limited number of cell lines, each expressing a large number of diverse genes. They constructed two reference pools from those cell lines with the greatest representation of unique genes, which are also easy to grow and yield high quantities of RNA. They found that adding more cell lines to the pool would not necessarily improve the overall gene representation because some genes were diluted below the detection limit. The first reference sample exhibited similar coverage with Statagene's human URR sample (75%), and the second reference sample had coverage equal to 80%. Thus, according to Yang et al., a simple pool of RNA from diverse cell lines can provide a superior reference [10]. Human, mouse and rat reference RNA samples were considered by Novoradovskaya et al. [5] and were prepared from pools of RNA derived from individual cell lines representing different cell-lines (ten human, eleven mouse and fourteen rat). They evaluated microarray coverage based on a pre-specified threshold equal to the background intensity or twice the background intensity of each channel, using different microarray platforms. Probes with intensities above threshold were characterized as present. They reported microarray coverage greater than 80% for all arrays tested when threshold was equal to background, and greater than 60% when threshold was equal to twice the background. Consequently, they agreed with Yang et al., that pools of RNA derived from a limited but diverse set of cell lines result in an optimal reference sample.

Furthermore, in a recent study we examined whether a small number of cell-lines (in a total of 22), each of which expresses different genes, could outperform more complex reference mixtures [12]. Following different techniques (exhaustive combination of cell-lines, heuristic search and stochastic simulated annealing) we achieved an optimal cell-line mixture of 11 cell-lines with theoretical gene coverage of 57.84%. The mixture was tested with a wet-lab experiment (i.e., it was hybridized with the mixture of all 22 cell-lines), and showed coverage of 57.12% - quite near to the theoretical one. Testing (with a similar experimental set-up) the standard Stratagene URR mixture an inferior coverage of 54.02% could be reached.

The relatively low gene coverage (around 57%) but also, the need to offer a more natural, universal and robust design approach forced us to consider a straightforward

[2] http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm
[3] http://www.genomics.agilent.com/
[4] http://www.ambion.com/catalog/CatNum.php?6050

optimization approach. In particular, we utilized the mathematical programming framework and devised a *linear/binary programming* optimization methodology. Our methodology is flexible enough and could be easily tuned to fit the specifics of different microarray platforms and heterogeneous experimental settings.

3 Towards Optimal Reference Sample Designs

Assume a set of genes, G, and a set of cell-lines, C. In addition, we assume that the URR sample is to be devised using equal shares of fixed quantity from each cell-line. A series of C two-channel microarray experiments are performed, each time hybridizing an individual cell-line with the reference-mix. After standard, cross-experiment normalization of the acquired intensities, we get the gene-expression profile for each cell-line, and the respective gene-expression matrix is formed - a matrix with G rows and C columns. Our fundamental concern is twofold: which genes are expressed over all cell-lines (i.e., across rows), and which genes are expressed by each cell-line (i.e., across columns). We address the problem by setting a *cut-off* value, with gene-expression values greater than the cut-off to be considered as detectably expressed (with the remaining considered as not detectable). In most microarray experiments expression cut-off values could be acquired either during the microarray image analysis phase where, the (manual) inspection of the underlying background intensities guides to the determination of a specific cut-off value in order to consider the gene signal as detectable or, by the inspection of the normalized gene-expression matrix and the calculation of different cut-off values for different fold-change gene expression rates. As the determination of the cut-off value is beyond the scope of the current work, we assume an absolute cut-off based on the *median* over all gene-expression values.

We consider a gene as *detectably expressed* if an *adequate number of selected cell-lines express it*, i.e., the corresponding gene's expression values for these cell-lines are over the cut-off. In the present paper we report results with two different cut-off values: one that equals the median, and one being set to the 70% of the median of all gene-expression values. Elaborating on the previous discussion, we post two *criteria* in order to consider the number of cell-lines (to be utilized in the URR) as optimal:

(i) each array gene should exhibit an *adequate* number of its corresponding expression-values over the cut-off threshold - in this case we consider the gene as expressed, and

(ii) the number of expressed genes is over a (user-defined) pre-specified percentage of all array genes. All genes considered as expressed by (i) are considered as covered, and their percentage over all the array genes presents the coverage figure of the final URR mixture.

3.1 The Optimization Methodology

In order to meet and fulfill the aforementioned criteria we devised an optimization process enabled by a linear/binary programming problem formulation.

Notation. Cell-line: given a set of m cell-lines, assume, C_j, a set of *binary* variables, $C_j \in \{0,1\}$, $0 \leq j \leq m$ to denote the *presence* ($C_j = 1$) or, *absence* ($C_j = 0$) status of a cell-line. We do not cope with the case of multiple (i.e., integer or, real) values for C_j.

This is a more difficult problem that calls for a more elaborate and complex modeling. In such a case we need more microarray experiments in order to assess exact dilution and saturation effects when the cell-lines are taking part in a reference-mix in non-equal shares; *Gene-Expression value*: given a set of n genes, G_i, $0 \le i \le n$, we denote with G_{ij} the expression value of gene i for cell-line j; *Cut-Off*: the pre-specified cut-off threshold is denoted with T. Based on this notation we form the following Binary Programming Optimization Problem (BPOP):

$$\text{minimize OF:} \quad \sum_{j=1}^{m} C_j \tag{1}$$

$$\text{subject to:} \quad \sum_{j=1}^{m} C_j * G_{ij} \geq kT \tag{2}$$

$$i\ , 1 \le i \le n$$

$$C \in \{0,1\}, G_{ij} \in \mathrm{R}$$

$$0 \le k \le m$$

Solutions to the BPOP problem will assign values to the binary variables C_j. If $C_j=1$ then, the respective cell-line is to be utilized in the URR sample, otherwise it will not. The objective function (1) seeks for the minimum number of cell-lines to be utilized. Each constraint in (2) states the following requirement: the weighted sum of gene-expression values (over all present cell-lines) should be greater or equal to kT, i.e., k times the cut-off value T. In other words, we try minimize the number of cell-lines, and at the same time to keep the gene-expression values over k times the (pre-specified) cut-off threshold. In this way, the constraints guide the minimization process to select those cell-lines for which at least one or more of its gene-expression values is over the cut-off. In the ideal, the corresponding gene-expression values will be greater than the cut-off in all present cell-lines – of course this may happen if all genes exhibit at least one of their expression values (across all present cell-lines) over the cut-off. What we really seek is an adequate number of cell-lines for which each gene exhibits an expression value over the cut-off. Tolerating (by varying the k control number) the required number of cell-lines covering a gene the BPOP process results into different solutions. For example, we may consider a gene as expressed if *at-least-one* of its expression values is over the cut-off, i.e., at-least-one cell-line covers it. This requirement is guaranteed by the formulation of BPOP - the constraints presented by (2) guides the optimization process to select those cell-lines for which, not only just one value is over the cut-off but also all other values are high enough in order to pass the kT threshold. So, for different and increasing values of k the actual BPOP solutions meet the aforementioned criteria for the design of an optimal reference sample.

4 Experiments

A set of 22 tissue-specific cell-lines were targeted as candidates for the design of the optimal URR (their code, name and tissue-origin are shown in Table 1). With these cell-lines a set of 22 microarray experiments were conducted on a custom-made microarray platform comprising 34771 gene transcripts (for details of cell-line hybridizations, acquisition of signal-intensities and normalization across experiments the reader may refer to [12]). As a number of missing gene-expression values were

present, and in order to avoid as much as possible noise effects, we conducted a data cleaning procedure: (a) genes having more than half, across all cell-lines, missing values were eliminated, and (b) for the remaining genes we replaced missing values with the average of the respective gene's expression values. The data cleaning process concluded into a final set of 34673 genes.

Table 1. The target cell-lines

	Cell line	Tissue origin
1.	HL60	Bone marrow
2.	Hs578T	Mammary Gland
3.	McF7	Mammary Gland
4.	OVCAR3	Ovary
5.	Panc1	Pancreas
6.	SKMEL3	Skin
7.	SKMM2	Bone Marrow
8.	T47D	Mammary Gland
9.	TERA1	Testis
10.	U87MG	Brain
11.	Raji	B-lymphoblasts
12.	JAR	Genital
13.	Saos2	Bone
14.	SW872	Liposarcoma
15.	THP1	Peripheral Blood
16.	HCT116	Colon
17.	HUVEC	Umbelical Vein
18.	HepG2	Liver
19.	HeLa	Cervix
20.	LNCap	Prostate
21.	Molt4	T-lymphoblasts
22.	WERI1	Retina

We set the cut-off threshold to be equal to the median (T=368.64) of all gene-expression values, and to a proportion of 70% of the median (T = 258.05). Because of space limitations we do not report the full spectrum of results but just some observations should be reported: (a) a percentage of 28.1% and 5.0% of genes are not covered by any cell-line, (b) the percentage of genes being covered by just-one (k=1) time the respective cut-off threshold is 7.4% and 4.5%, respectively – in this case the respective genes are considered as *tissue-specific*, and are not taking part in the optimization process, and (c) the percentage of genes being covered by all, (k=22) times the respective cut-off threshold is 31.5% and 46.7%, respectively - these genes are considered as *tissue-independent,* and again are not taking part in the optimization process (whatever cell-line is to be included in the final URR, these genes will always exhibit gene-expression levels over the respective cut-off threshold; and (iv) the percentage of genes being covered by at-least-one ($k\geq1$) times the respective cut-off threshold is 71.9% and 95.0%, respectively - this observation validates the determination of the median as a rational cut-off value.

4.1 Results and Discussion

We run the BPOP process for each of the specified cut-off thresholds using the MOSEK optimization suite[5]. MOSEK optimization software solves large-scale mathematical optimization problems, and provides specialized solvers for

[5] http://www.mosek.com

linear/binary programming and mixed integer programming problems. In the sequel we present and discuss on the respective results.

T = 368.64 cut-off. Varying the *k* control number from 2 to 21 (times the cut-off threshold value - recall that genes with just one, k=1, and all, k=22 of their expression values over the cut-off are discarded), we came up with the results presented in Table 2. The '*k*' row refers to the number of times the sum over all weighted gene-expression values exceeds the cut-off – it controls the lower limit for the formed BPOP constraints. Note that for k>13 no solutions could be found, i.e., there are no genes covered by more than 13 cell-lines.

Table 2. Gene coverage results of the BPOP process for cut-off *T*=368.64

at-least	**one**				**two**				**three**					**three**
k	1	2	3	4	5	6	7	8	9	10	11	12	13	
C_j**=1**	2	4	**6**	**7**	9	10	12	**13**	15	16	18	19	**21**	22
0	45.4%	40.4%	38.1%	36.7%	34.7%	33.9%	33.0%	32.3%	30.9%	30.1%	29.4%	29.2%	28.3%	28.1%
1	54.6%	59.6%	**61.9%**	**63.3%**	65.3%	66.1%	67.0%	67.7%	69.1%	69.9%	70.6%	70.8%	71.7%	71.9%
2	45.9%	51.9%	55.0%	56.2%	58.1%	58.1%	59.7%	**60.5%**	61.7%	62.2%	63.1%	63.4%	64.2%	64.5%
3		47.1%	51.3%	52.4%	54.5%	54.5%	56.1%	57.0%	58.0%	58.5%	59.4%	59.7%	**60.4%**	**60.7%**
4		40.4%	48.0%	49.6%	51.9%	51.9%	53.7%	54.7%	55.7%	56.2%	57.0%	57.3%	58.1%	58.3%
5			44.4%	46.7%	49.6%	49.6%	51.7%	52.9%	53.9%	54.5%	55.3%	55.6%	56.3%	56.5%
6			38.7%	43.2%	47.4%	47.4%	49.9%	51.3%	52.5%	53.0%	53.8%	54.1%	54.9%	55.2%
7				37.7%	44.8%	44.8%	48.1%	49.8%	51.2%	51.8%	52.6%	53.0%	53.8%	54.1%
8					41.7%	41.7%	46.2%	48.3%	49.9%	50.5%	51.5%	51.9%	52.7%	53.0%
9					36.2%	36.2%	43.8%	46.6%	48.5%	49.2%	50.4%	50.9%	51.7%	52.1%
10							40.6%	44.8%	47.2%	48.0%	49.3%	49.8%	50.8%	51.2%
11							35.2%	42.6%	45.7%	46.7%	48.2%	48.8%	49.8%	50.3%
12							34.7%	39.8%	44.0%	45.2%	46.9%	47.5%	48.8%	49.4%
13								34.5%	41.8%	43.5%	45.7%	46.4%	47.8%	48.4%
14									38.9%	41.1%	44.4%	45.2%	46.7%	47.5%
15									33.8%	38.1%	42.7%	43.9%	45.7%	46.4%
16										33.1%	40.4%	42.2%	44.5%	45.4%
17											37.5%	39.9%	43.2%	44.3%
18											32.6%	37.1%	41.3%	42.9%
19												32.2%	39.2%	41.2%
20													36.4%	39.0%
21													31.6%	36.3%
22														31.5%

COVER

The respective non-zero binary variables in each solution are shown in row 'C_j=1', for example, when *k*=4 the solution includes 7 non-zero, i.e., the sum of genes' expression values is greater or equal to four times the cut-off is satisfied. In other words, all constraints are satisfied at an optimal selection of seven cell-lines. Setting a lower (user specified) coverage limit of 60%, and in the case that all 22 cell-lines are to be included in the final URR mixture, all genes should exhibit gene-expression levels over the cut-off in at-least-three cell-lines. For *k*=3 the corresponding adequate number of cell-lines is equal to at-least-one. On this level of adequacy, i.e., when we want the genes to exhibit in at-least-one cell-line an expression value over the cut-off, we achieve coverage of 61.9% with the inclusion of just 6 cell-lines. For *k*=4 the corresponding adequate number of cell-lines is still to at-least-one. On this level of adequacy we achieve coverage of 63.3% with the inclusion of just 7 cell-lines. For *k*=8 the corresponding adequate number of cell-lines is equal to at-least-two. On this level of adequacy, i.e., when we want the genes to exhibit in at-least-two cell-lines expression values over the cut-off, we achieve coverage of 60.5% in the presence of 13 cell-lines. When the corresponding adequate number of cell-lines is equal to at-least-three, we may achieve a coverage of 60.4% but with a costly reference sample design of 21 cell-lines being present (*k*=13). With a careful inspection of the results we devised a map of the cell-lines being present in the aforementioned solutions (Table 3).

Table 3. Present cell-lines (marked) in the solutions for the design of an optimal reference sample for cut-off T=368.64

k	3	4	7	8	12	13
# CELL-LINES	6	7	12	13	19	21
COVERAGE	**61.9%**	**63.3%**	**59.7%**	**60.5%**	**59.7%**	**60.4%**
CELL-LINE	**PRESENCE**					
HL60	✓	✓	✓	✓	✓	✓
Hs578T			✓	✓	✓	✓
McF7					✓	✓
OVCAR3					✓	✓
Panc1	✓	✓	✓	✓	✓	✓
SKMEL3			✓	✓		
SKMM2				✓	✓	✓
T47D				✓	✓	✓
TERA1		✓	✓	✓	✓	✓
U87MG			✓	✓	✓	✓
Raji						✓
JAR	✓	✓	✓	✓	✓	✓
Saos2			✓		✓	✓
SW872			✓	✓	✓	✓
THP1				✓	✓	✓
HCT116	✓	✓	✓		✓	✓
HUVEC						✓
HepG2	✓	✓	✓	✓	✓	✓
HeLa	✓	✓		✓	✓	✓
LNCap					✓	✓
Molt4					✓	✓
WERI1			✓		✓	✓

From table 3 we may observe that there are exactly four cell-lines, HL60, Panc1, JAR and HepG2, being present in all solutions. So, it is natural to consider them as the most promising and the only components of the optimal URR sample mixture. With just these cell-lines in the final URR reference mixture we were able to achieve coverage of 60%. Note, that two of these cell-lines (Panc1, HepG2) are also included in the experimentally tested URR mixture that achieves coverage of 57.12% (refer to [12]).

T=258.05 cut-off. The BPOP process was also applied on the same data with a lower cut-off threshold T = 258.05 (70% of the median) but with a higher coverage limit of 75%. Because of space limitations we do not report results analogue to Tables 2 and 3. With a careful inspection of the results we again came up with a map of the cell-lines being present in the solutions that meet the aforementioned criteria. There are thirteen of them with just six cell-lines, HL60, Hs578T, Raji, HepG2, LNCap, and WERI1 being present in at least eleven of the thirteen solutions. So (as in the case of the 368.64 threshold) it is natural to consider them as the most promising and the only components of the optimal URR sample mixture. With just these six cell-lines we were able to achieve a coverage of 87.8%. Again, five (Hs578T, Raji, HepG2, LNCap, and WERI1) of the six cell-lines are included in the experimentally tested URR mixture (refer to [12]). So, and provided that with the presence of less cell-lines saturation and dilution effects are avoided as much as possible, we may hypothesize that the presented optimization approach greatly improves previous URR sample designs.

5 Concluding Remarks

We have presented an *in-silico* methodology for the design of optimal URR samples suited for 2CH microarray experiments. The biological problem to overcome relates to the avoidance (as much as possible) of dilution and saturation effects taking place when a big number of cell-lines are present. In the final URR mixture sample. We approached the problem with the careful formation of a binary programming optimization methodology (BPOP). In particular, we tried to minimize a set of binary

variables that represent the inclusion or not of a cell-line in the final URR mixture, and maximize the overall expression of genes. Setting different cut-off values, to determine if a gene is detectably expressed or not, we achieved high enough gene coverage percentages. In particular, setting the cut-off to the 70% the median of gene expression values, we were able to achieve a quite high coverage of 87.8% with the presence of just six cell-lines. A figure that outperforms current state-of-the-art URR sample designs.

Our future R&D plans include: (a) testing and assessment of the introduced optimization methodology on other cell-line configurations, and especially on integrated microarray data from different platforms and studies, and (b) elaboration of more advanced optimization techniques, e.g., integer and/or real programming optimization in order to extend our approach to varying cell-line shares in the final URR mixture.

References

1. Design of Microarray Experiments. Genomics & Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, NCI, NIH, DHHS, http://discover.nci.nih.gov/microarrayAnalysis/Experimental.Design.jsp
2. Khan, R.L., Gonye, E.E., Gao, S., Schwaber, J.S. A universal reference sample derived from clone vector for improved detection of differential gene expression. BMC Genomics 2006, 7:109 doi:10.1186/1471-2164-7-109 (2006)
3. Manduchi, E., White, P. Issues Related to Experimental Design and Normalization. WhitePaper-20040629, Report, Computational Biology and Informatics Laboratory, University of Pennsylvania. http://www.cbil.upenn.edu/downloads/EPConDB/download/Protocols/WhitePaper-20040629.doc
4. Eisen, M.B., Brown, P.O. DNA arrays for analysis of gene expression. Methods Enzymol. 1999;303:179--205 (1979)
5. Novoradovskaya, N., Whitfield, M.L., et al. Universal Reference RNA as a standard for microarray experiments. BMC Genomics 2004, 5:20 doi: 10.1186/1471-2164-5-20 (2004)
6. Williams, B.A., Gwirtz, R.M., Wold, B.J. Genomic DNA as a cohybridization standard for mammalian microarray measurements. Nucleic Acids Res 2004, 32(10):e81 (2004)
7. Gadgil, M., Lian, W., Gadgil, C., Kapur, V., Hu, W.S. An analysis of the use of genomic DNA as a universal reference in two channel DNA microarrays. BMC Genomics 2005, 2005, 6:66 (2005)
8. Sterrenburg, E., Turk, R., Boer, J.M., van Ommen, G.B., den Dunnen, J.T. A common reference for cDNA microarray hybridizations. Nucleic Acids Res. 2002, 30(21):e116 (2002)
9. Dudley, A.M., Aach, J., Steffen, M.A., Church, G.M. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci 2002, 99(11), 7554--7559 (2002)
10. Yang, I.Y., Chen, E., et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biology 2002, 3 doi:10.1186/gb-2002-3-11-research0062
11. Shi, L, Reid, L.H., Jones, W.D., et al: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006, 24(9), 1151--1161 (2006)
12. Tsiliki, G., Kaforou, S., Kapsetaki, M., Potamias, G., Kafetzopoulos, D.A. A computational approach to microarray universal reference sample. In: 8th IEEE International Conference on BioInformatics and BioEngineering, pp. 1--7 doi: 10.1109/BIBE.2008.4696690 (2008)