# Chapter 88
# Towards Optimal Microarray Universal Reference Sample Designs: An *In-Silico* Optimization Approach

**George Potamias**
*ICS-Forth, Greece*

**Sofia Kaforou**
*IMBB-Forth, Greece*

**Dimitris Kafetzopoulos**
*ICS-Forth, Greece*

## ABSTRACT

*In this paper, the authors present an 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 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. Toward this target, the authors 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.*

## INTRODUCTION

Scientific experimental science is founded on the 'right' design of experiments where some common-sense principles should apply. 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 (Tsiliki, 2011). In this context, assessing the reli-

DOI: 10.4018/978-1-4666-3604-0.ch088

ability 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 (Khan, 2006).

With the 2ch microarray technology there are two alternatives for the comparison between different sample conditions, c1 vs. C2: (i) the '*loop*' design where, direct comparisons between a series of arrays are performed with a sample of type c1 in one channel and one of type c2 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 c1/c2 in one channel and the urr sample in the other. Then differences in gene-expression between c1 and c2 are acquired by comparing the ratios c1/c vs. C2/c (Manduchi, 2007). The hybridization of each (experimental) sample with an urr *mixture* serves as a common denominator between different microarray hybridizations (Eisen, 1999). 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. In addition, the urr design facilitates the normalization and the subsequent comparison between heterogeneous microarray data sets (Novoradvskaya, 2004).

Typically molecular scientists and experimentalists prepare their own array-specific reference samples, e.g., genomic dna (Williams, 2004; Gadgil, 2005), a mixture of clones spotted on the arrays (Sterrenburg, 2002) as well as complementary to every microarray spot short oligomers (Dudley, 2002). Many groups utilize a mixture of cell-lines for their urr, e.g., Stratagene's human urr (Novoradvskaya, 2004) or, clone-vectors (Khan, 2006). 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 (Yang, 2002). 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 c1/c vs. C2/c ratio comparisons, a fact that would force discarding those spots from the analyses (Manduchi, 2007).

It is natural to assume that as more cell-lines are included in the urr sample mixture 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 bet-