Not Working on Mice or Humans?

A Pipeline to Rapidly Generate Species-specific Microarrays from Sequence Databases

Microarrays technology is fast developing and its application is expanding from *Homo sapiens* to a wide number of species where enough information on sequences and annotations are available. The number of species for which a dedicated platform exists is not high. The Expressed Sequence Tags (ESTs) databases represent a collection of anonymous sequences that can be used to build species specific microarrays for species whose genome sequences are largely unknown.

Susana Bueno, Researcher, CASPUR (Inter-University Consortium for the Application of Super-Computing for Universities and Research), Rome, Italy

Lorraine Pariset, Researcher, Department of Animal Science, Tuscia University, Viterbo, Italy

Silvia Bongiorni, Post Doctoral Scientist, Department of Animal Science, Tuscia University, Viterbo, Italy

Alessio Valentini, Director of Department of Animal Science, Tuscia University, Viterbo, Italy

We have developed a pipeline that allows the production of oligos for *in situ* synthesizing microarrays starting from unannotated, redundant EST sequences. The system was tested by constructing the first annotated microarray with a covering of most of the sheep genome. This method can be easily extended to other species of which genetic sequences are present in public databases. As a perspective, the approach can be applied also to species of which no sequences are available to date, thanks to high-throughput "next generation" sequencing methods.

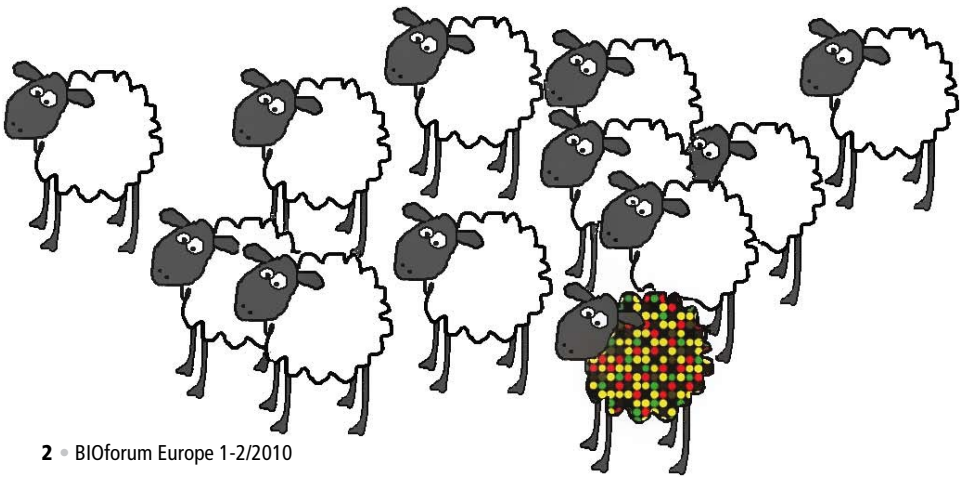
What Is a Microarray?

Microarray technology, since its introduction in 1995 [1], has been employed for many different applications, such as gene expression profiling, microbial detection, SNP genotyping, comparative genome hybridization, ChiP on chip analysis and miRNA detection.

Microarray manufacturing is based on spotting cDNA or pre-synthesized oligonucleotides, inkjet depositing technologies [2], light-directed synthesis processes [3], and local electrochemistry [4]. Schematically, a gene expression microarray works as follows: mRNA, extracted from a sample and labeled with a fluorescent dye as it is or converted into cDNA, is hybridized to a platform harboring targets corresponding to genes of interest arranged in an easily-coded template (the array). When the microarray is excited by a laser with a specific wavelength, the hybridized probe emits fluorescence. The raw intensities of the fluorescence give an estimation of the level of gene expression. A typical microarray experiment uses cDNA or oligonucleotides arrayed on supports that are usually glass microscope slides or silicon chips. The platform employed influences the number of gene probes that can be hosted on the array, which is higher in the case of *in situ* synthesized oligos. In spotted microarrays, the gene probes are formerly synthesized and then „spotted" onto glass. This technique is commonly used to produce customized printed microarrays at a relatively low-cost per unit.

Today, the most densely populated arrays are produced by *in situ* synthesis through light-directed process [5, 6] or, at a lower scale, through electrochemistry [4]. The latter technology, based on a silicon microchip that includes also a circuitry, can be employed for the oligonucleotide synthesis and also for electrochemical detection of target molecules bound to the microarray, beside the conventional fluorescent scanner method [4].

A standard microarray experiment compares mRNA abundance between two different samples on the same support, designed to work with either a single or dual detection system. In inkjet-printed and spotted microarrays two different samples are simultaneously hybridized using a two-color hybridization method.



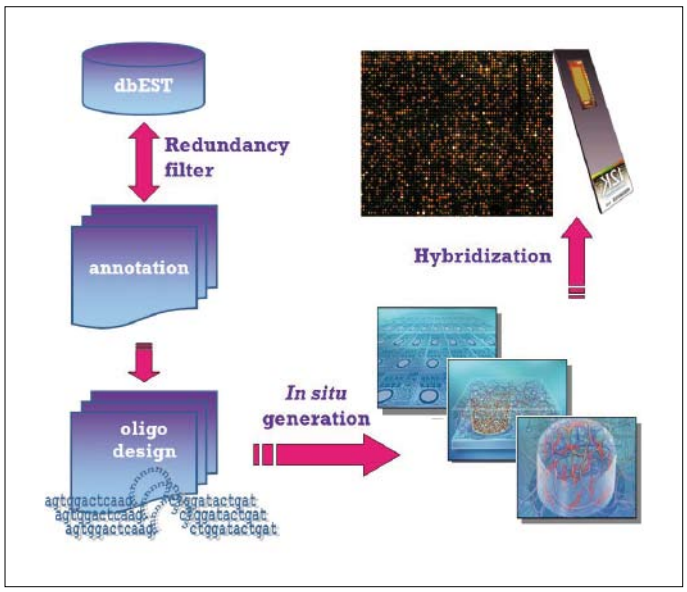


Fig. 1: Microarray pipeline flowchart: from unannotated, redundant sequences to oligonucleotides design, in situ generation on chip and probe hybridization.

On the Affymetrix GeneChip, only one sample per chip can be hybridized using a single-color detection system; Combimatrix chips employ either two-color or single-color scheme and the same array can be stripped and re-hybridized up to four times.

Available Platforms

In NCBI Gene Expression Omnibus (GEO) more than 100 species are present, with most (~70%) platforms represented by spotted DNA/cDNA or oligonucleotides. This is a weak point in microarray generation in that, before spotting, libraries have to be prepared and sequenced, while oligonucleotides should be synthesized on a large scale. Therefore, a considerable lag is expected between the starting of information collection and the microarray applications.

Compared to cDNA arrays, *in situ* synthesized oligonucleotides offer increased specificity and sensitivity and minimize chip-to-chip variations, even if one drawback is the price that can be up to 10 fold higher than in house spotted arrays [7]. The ~30% of GEO platforms represented by *in situ* generated oligonucleotides are mainly produced by big companies interested in species for which there is high attention worldwide and, as a result, they are 65% human, 37% model species, 8% pathogens, and 3% agricultural species.

Cross species hybridization has been used for comparative analysis of transcriptome between divergent genomes. Different methods have been proposed to select unbiased probes for interspecific transcriptome analysis, as those based on genomic DNA hybridization [8].

Anonymous sequences of species whose genome sequences are largely unknown are deposited in gene banks, particularly as a result of Expressed Sequence Tags (EST) sequencing projects. EST databases are redundant and errors are likely, nevertheless they represent accessible data that can be exploited with some bioinformatic work. Part of it is already carried out by NCBI, which eliminates the redundancy by converting raw sequences in the Reference Sequences (Refseq). However, most of the work still remains to be done, i.e. the annotation of the anonymous sequences.

A Pipeline to Rapidly Generate Custom Microarrays from Sequence Databases

We have developed a pipeline of software instruments that allow starting from unannotated, redundant sequences as those found in public databases or generated by parallel sequencing, to yield oligonucleotides suitable for *in situ* generation on chip. Microarrays designed from not fully annotated genome become quickly obsolete, due to the increasing availability of transcript sequences in public databases. Our developed pipeline software answers this problem, since the target oligonucleotides can be easily and rapidly regenerated, including new information just before the microarray production.

The method was tested by generating a chip from sheep ESTs deposited at NCBI. Oligonucleotides of 40 nucleotides length were designed using the GoArrays software [9], which designs two short sequences interleaved by a random DNA spacer to achieve a better annealing of the cDNA, and *in situ* generated using the Combimatrix (Seattle, WA, USA) equipment. The 40-mer length was chosen because shorter oligonucloetide probes might be more deeply influenced by a single base pair mismatch, while linkers have been used to extend probes and provide greater specificity. The chip, carrying 21,743 non-redundant features in quadruplicate, 73.4% of which are fully annotated and correspond to 10,190 genes, represents a good coverage of the sheep genome [10]. The NCBI sheep sequences have been annotated in a sequential procedure by blasting anonymous ESTs on the sheep specific database, and subsequently on databases of homologous species in phylogenetic order, if sequences were not covered by the closer database.

The microarray efficiency was assessed by performing pilot experiments using RNA of two sheep breeds [11] and achieving very good technical outcomes, such as in slide replicates coefficient of variation <0.25 for differentially expressed genes with P<0.01 (fig.1).

We conclude that the method is very efficient and can be easily extended to other species of which genetic sequences are present in public databases [10]. With this procedure, even a microarray in single copy can be generated with a moderate cost. Therefore, we believe that this method allows the study of species so far neglected with advanced devices like microarrays. As a perspective, the approach can be applied also to species of which no sequences are available to date, by using high-throughput deep sequencing methods.

References

[1] Schena M. *et al.*: Science 270, 467-70 (1995)
[2] Hughes T.R. *et al.*: Nat Biotechnol. 19, 342-347 (2001)
[3] Tan P.K. *et al.*: Nucleic Acids Res. 3119, 5676-5684 (2003)
[4] Ghindilis A.L. *et al.*: Biosensors and Bioelectronics 22, 1853-1860 (2007)
[5] Pease A.C. *et al.*: Proc Natl Acad Sci U S A. 91, 5022-5026 (1994)
[6] Nuwaysir E.F. *et al.*: Genome Res. 12, 1749-1755 (2002)
[7] Lee N.H. and Saeed A.I.: Methods Mol Biol. 353, 265-300 (2007)
[8] Chain F.J.J. *et al.*: PLoS ONE 3, e3279 (2008)
[9] Rimour S. *et al.*: Bioinformatics 21, 1094-1103 (2005)
[10] Pariset L. *et al.*: New Biotechnol. 25 (5), 272-279 (2009)
[11] Bongiorni S. *et al.*: Ital. J. Anim. Sci. 8 (2), 33-35 (2009)

Authors:
Lorraine Pariset, Susana Bueno, Silvia Bongiorni, Alessio Valentini

CONTACT:
Dr. Lorraine Pariset
Dipartment of Animal Science
Università degli Studi della Tuscia
Viterbo, Italy
pariset@unitus.it
www.unitus.it