



Sixth Framework Programme for Quality of Life and
Management of Living Resources

Project no. LSHG-CT-2006-037686

EMERALD

Empowering the Microarray-Based
European Research Area to Take a Lead in
Development and Exploitation

EU Deliverable: D6.1

Due Date: Jan. 2008
Delivery Date: 30.04 2008

Version 1

Partner responsible: UU

Implication for New Technologies. Survey of the Microarray Field

Olle Ericsson and Ulf Landegren

ABSTRACT

Microarray analysis is a powerful approach to interrogating large sets of biomolecules, and it has provided the basis for analysis of a wide range of molecular states at levels of DNA, RNA and protein. It is important to understand the limitations of current microarray platforms in order to develop enhanced technologies, and to define the role of microarrays compared to e.g. new parallel DNA sequencing approaches. In this brief survey we describe some new trends in microarray-based analyses, and we compare these to other approaches towards high-throughput molecular biology.

INTRODUCTION

The use of microarrays traces its roots to several trends during the late 20th century, following Richard Feynman's famous dictum "There's plenty of space at the bottom" (<http://www.zyvex.com/nanotech/feynman.html>). Particularly important inspirations are the reverse dot blot techniques¹, Roger Eakin's miniaturized sandwich elisas², and the efforts to establish parallel methods for sequencing by hybridization³. Ever since the introduction of nucleic acid microarrays as we recognize them today^{4,5}, feature density, probe types and application areas have expanded continuously. Initially microarray quality control issues produced data that frequently was inconsistent among users, a problem that is being countered by community efforts, including the EU sponsored EMERALD project. While initial applications focused on analyses of transcript levels, more recent technologies also serve to investigating factors like splice variation and the presence of single nucleotide polymorphisms (SNPs)^{6,7}. Novel technologies, improvements of current protocols and manufacturing methods will have major effects on how the microarray analysis format can be incorporated in future applications. We will review recent trends in the microarray field and discuss how microarrays can be improved as well as inherent limitations of the technology. The microarray technology is specifically discussed in the context of the new nucleic acid sequencing platforms that are currently rapidly gaining popularity.

RECENT TRENDS

Microarray-based techniques for gene expression profiling have evolved continuously and currently synthetic oligonucleotide arrays are generally favored over microarrays generated by printing PCR products. Oligonucleotide arrays allow more exact *in silico* design of probe sets and several oligonucleotide probes can be used to target individual genes. Probes can be designed to tile each gene to be investigated. Alternatively, probes targeting separate exons⁸ or exon junctions⁹ can be used to interrogate splice variation. The density of microarrays has been successively improved and currently microarrays comprising between 100.000 and one million features are routinely produced. Probe

density therefore is no longer a limiting factor for most expression profiling applications. Early inter-platform comparisons investigating differentially regulated genes often identified poorly overlapping gene sets^{10,11}. This has been improved by community efforts such as the use of common RNA standards¹² and standards for reporting minimum experiment information from microarray-based analyses¹³. The increased commercialization of and competition among manufacturers of microarray technologies, along with centralization of the application of the technology to core centers have improved production standards and handling. As a consequence, recent comparisons of microarray platforms paint a more promising picture of the available techniques¹⁴. The most widely used providers of devices for microarray analyses currently is Affymetrix, followed by Agilent, Illumina, GE, Applied Biosystems.

The development of robust nucleic acid microarray techniques for gene expression profiling has provided an inspiration for many other microarray application areas. Currently, microarrays are used for calling a million single nucleic acid polymorphisms (SNPs) in parallel, thus enabling the analysis of the genetic basis of complex traits like pigmentation and height^{15,16}. The microarray platform has been applied for analyses of splicing, transcription factor binding sites¹⁷, chromatin immunoprecipitation with microarray readout¹⁸ (ChIP on chip), and recently microarrays have become popular for capturing exons for subsequent high-throughput sequencing of selected parts of genomes¹⁹.

ROOM FOR IMPROVEMENT

There remains considerable room for improvement of microarray analyses, however. Current microarray platforms typically enable detection of transcripts that are present at 1-10 copies per cell as determined using quantitative PCR^{20,21}. Although this detection limit may appear sufficient at a glance, this may not be the case for many applications. In a study by Zhang et al. where the SAGE technology was used to sequence bits of transcripts, an estimated 86% of all transcripts were expressed at levels below five transcripts per cell²². The problem is compounded by the fact that investigated tissue samples do not comprise homogenous cell populations, resulting in lower expression of transcripts from minor subpopulations. Even in homogenous cell populations, temporal regulation like cell cycle specific transcription and transient expression bursts separated by periods with little or no expression can produce average transcript abundances below one copy per cell. Weak microarray signals typically display higher variation and they are also more vulnerable to cross-hybridization of abundant targets, reducing assay performance.

Amplification and specificity

One frequent misconception among microarray users is that amplification of either the target nucleic acid or the detection signal are sure means to solve problems with limits of detection and dynamic ranges. In practice, however, enhanced amplification that is not accompanied by the appropriate specificity cannot enhance microarray performance. If a microarray probe fails to discriminate a target transcript from a homologous transcript present at much higher concentration, then amplification of absolute transcript concentrations or signals generated in the microarray feature cannot prevent that the more

abundant transcripts is detected in place of the intended one. Amplification of non-specific signals will also increase background signals, and it may therefore reduce overall microarray performance.

Complexity of the transcriptome vs. the genome

The scoring of SNPs in genomic DNA requires additional means besides conventional hybridization due to the sample complexity. These means may include reduction of DNA complexity prior to microarray hybridization⁶ or introduction of enzymatic steps that serve to enhance discrimination^{7,23}. Recent reports indicate that a large proportion of the non-repetitive genome is transcribed²⁴, rendering the complexity of the transcriptome similar to that of the genome. This may explain the relative difficulty of interrogating rare transcripts as the specificity required to detect these would be similar to that required in SNP analysis. Considering that the levels of transcript expression can vary over a million fold²⁰, the molar fraction of a rare transcript in the transcriptome may be even lower than that of a single-copy genomic locus subjected to SNP analysis. It becomes increasingly difficult to maintain specificity and to avoid cross-hybridization the more rare the target transcript is in the analyzed sample.

Improving microarrays

The problem of cross hybridization of targets in solution to erroneous microarray probes can be improved in several ways. The combination of probe-target hybridization with on-chip enzymatic discrimination⁷ is one strategy and the use of multiple probes for targeting each transcript is another⁶. Approaches have also been established where nucleic acid probes recognize target molecules in solution and thereby become converted to amplifiable probe molecules that include standard nucleic acid zip codes for subsequent sorting on tag microarrays^{23, 25}. Although tags designed *in silico* can serve to improve hybridization specificity, it is apparent that not all mechanisms underlying cross hybridization are clearly understood, and *in silico* designed tag sets still display some levels of cross-hybridization²⁶. The chip-based hybridization of tagged probe amplification products has been successfully combined with enzyme-based recognition steps that enhance discrimination in order to eliminate cross-hybridization signals.

In one recent approach our group used ligation of so-called padlock probes, directed at specific cDNAs to measure gene expression with microarray read-out via a dual-tag approach. After amplification of circularized probes in solution by rolling circle amplification (RCA), amplified and monomerized single-stranded reporter molecules were in turn circularized upon recognition of dual-tag containing oligonucleotides in microarrays. Reporter molecules whose tag sequences at each end perfectly matched the immobilized probes were thus converted to circular DNA strands, and could be detected after signal amplification by a second, on-chip RCA (Figure 1). This procedure depends on highly specific probe ligation reactions, templated first by cDNA targets, and then on arrays, and it was shown to enhance performance by both greatly reducing cross hybridization and enhancing signal output²⁷.

Microarray limitations

It appears plausible, although still controversial, that a certain background or leakage expression occurs from the genome at a level that is insignificant for cellular function. Random transcription of genomic sequences or leakage transcription from *bona fide* promoters could generate transcripts and/or proteins at concentrations or in contexts where the products lack functional effects without resulting in evolutionarily deleterious effects. The consequence of such transcription events may be that many transcripts are found that are expressed at low and non-functional levels in all cells. This average transcript noise level would then set the limit for detection of meaningful gene expression in cellular subpopulations. If the expression of biologically significant transcripts in a minor cell population does not exceed this background expression of the whole population, then these transcripts will be extremely difficult to analyze even with perfect specificity of detection. Novel technologies enabling analysis of targets in single cells will be important to investigate expression patterns of minor sub-populations or transient temporal expression bursts. However so far these technologies typically only allow analysis of single or few targets per analysis compared to microarrays.

NEXT GENERATION SEQUENCING & FUTURE TRENDS

Recently several new platforms for DNA sequence analysis have been introduced allowing analysis of up to ten billion bp per instrument run. Microarrays are closely tied to the development of these next generation sequencing approaches. Firstly, parallel sequencing techniques may become directly competitive with gene expression analysis using microarrays. Secondly, the microarray platform may complement next generation sequencing by assisting in sample preparation for sequencing by exons capture. Finally, ordered microarrays may also provide a basis for next generation sequencing by ordering templates in defined locations.

Methods like SAGE (Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW (1995). Serial Analysis Of Gene Expression. *Science* 270, 484-487) or MPSS (Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., *et al.* (2000) *Nat. Biotechnol.* **18**, 630-634) have been available for some time for measuring gene expression by sequencing short segments of individual expressed genes in order to identify these. The new next generation sequencing instrument have been used in a similar fashion to provide digital data on gene expression (www.illumina.com) and the new sequencing techniques have also been used to sequence full length cDNA synthesized from polyadenylated mRNA²⁴. The advantages of these sequencing-based approaches are the high specificity due to sequence identification and precise digital data acquisition of read counts. Gene expression profiling by sequencing also allows novel transcripts to be identified, while microarrays are limited to interrogating predefined sets of genes via complementary probes. Drawbacks of sequence analysis for gene expression profiling include difficulties to detect rare transcripts in the presence of abundant genes, necessitating vast oversampling of highly expressed genes, and the fact that sequencing still remains expensive and time consuming compared to microarray analyses.

Returning to the topic of microarrays, it is clear that they can be used to achieve extraordinary sensitivity and specificity, particularly upon combination with enzymatic approaches to enhance selectivity and signal amplification. Table 1 outlines approaches that have been applied to improve microarray performance. As discussed, microarray-based ligation has been demonstrated to dramatically improve selectivity and to essentially eliminate cross hybridization. However, so far enzyme-assisted microarray analyses have primarily been exploited on a larger scale for SNP scoring, not for expression profiling that typically depend on regular probe-target hybridization reactions. It is likely that future microarray platforms will be used for sequencing of the captured targets, not merely for capturing of interesting regions for subsequent off-chip sequencing¹⁹. Extension of the single nucleotide polymerization approach currently used for SNP scoring⁷ to sequencing of several consecutive nucleotides is probably not far away. In this approach target selection by microarray capture and sequencing occurs in the same procedure with may enable resequencing of targeted regions without a first selection step.

In conclusion, microarrays will most likely remain a competitive solution for nucleic acid analysis in the foreseeable future, and they may increasingly move from research applications to clinical routine. However, the array-based methods will have to be compared to and sometimes combined with new technologies, such as methods for high-throughput sequencing.

1. Zhang, Y., Coyne, M.Y., Will, S.G., Levenson, C.H. & Kawasaki, E.S. Single-base mutational analysis of cancer and genetic diseases using membrane bound modified oligonucleotides. *Nucl. Acids Res.* **19**, 3929-3933 (1991).
2. Ekins, R.P. Multi-analyte immunoassay. *Journal of Pharmaceutical and Biomedical Analysis* **7**, 155 (1989).
3. Bains, W. & Smith, G.C. A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology* **135**, 303 (1988).
4. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).
5. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-1680 (1996).
6. Kennedy, G.C. et al. Large-scale genotyping of complex DNA. *Nat Biotechnol* (2003).
7. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549-554 (2005).
8. Shoemaker, D.D. et al. Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922-927 (2001).
9. Johnson, J.M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141-2144 (2003).
10. Vogel, G. Stem cells. 'Stemness' genes still elusive. *Science* **302**, 371 (2003).

11. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-412 (2002).
12. Baker, S.C. et al. The External RNA Controls Consortium: a progress report. *Nat Methods* **2**, 731-734 (2005).
13. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-371 (2001).
14. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151-1161 (2006).
15. Sanna, S. et al. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* **40**, 198-203 (2008).
16. Sulem, P. et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39**, 1443-1452 (2007).
17. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* **98**, 7158-7163 (2001).
18. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309 (2000).
19. Albert, T.J. et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903-905 (2007).
20. Holland, M.J. Transcript Abundance in Yeast Varies over Six Orders of Magnitude. *J. Biol. Chem.* **277**, 14363-14366 (2002).
21. Czechowski, T., Bari, R.P., Stitt, M., Scheible, W.-R. & Udvardi, M.K. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *The Plant Journal* **38**, 366-379 (2004).
22. Zhang, L. et al. Gene Expression Profiles in Normal and Cancer Cells. *Science* **276**, 1268-1272 (1997).
23. Hardenbol P, B.J. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* (2003).
24. Nagalakshmi, U. et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* (2008).
25. Yeakley, J.M. et al. Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol* **20**, 353-358 (2002).
26. Hardenbol, P. et al. Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269-275 (2005).
27. Ericsson, O. et al. A dual-tag microarray platform for high-performance nucleic acid and protein analyses. *Nucleic Acids Res* **36**, e45 (2008).

Table 1

Approach	Advantages	Disadvantages
Conventional hybridization	Straight forward, inexpensive	Cross-hybridization & off-spot signals
On-chip primer extension	Enzyme enhanced specificity	Background due to self priming
Multiple MM/PM probes	Bioinformatically enhanced specificity	Several probes, based on background subtraction not specific signal generation
On-chip ligation	Enzyme enhanced specificity & strong signal amplification using RCA	Requires solution-phase detection probes
Sequencing	Highly specific, high precision, de novo detection	Expensive, requires oversampling of abundant transcripts

Table 1. Brief outline of different approaches to enhance performance of nucleic acid analyses along with advantages & disadvantages (MM/PM – mismatch and perfect match probe pairs)

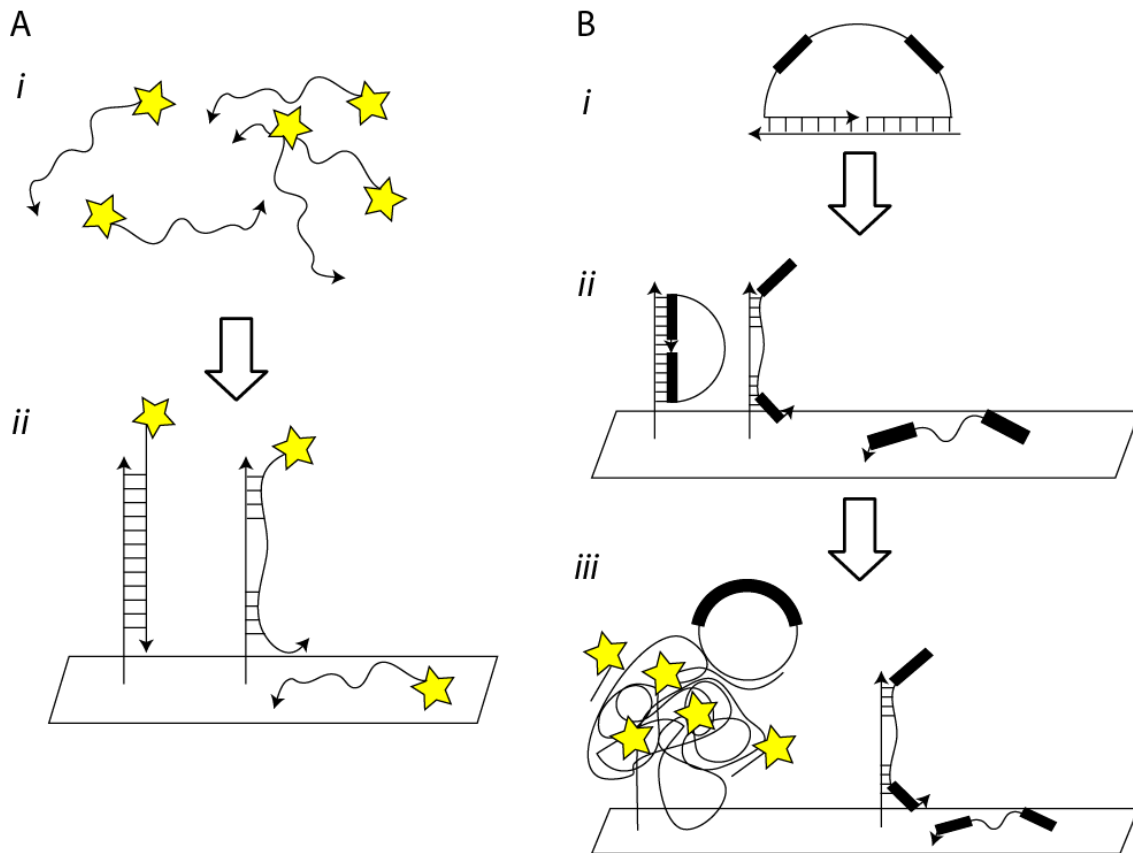


Figure 1. Comparison of microarray analysis by conventional hybridization (A) and enzyme assisted readout (B). A) Target nucleic acids are labeled in solution and hybridized to the microarray (*i*). Cross-hybridized nucleic acids generate signals in features other than those intended and off-spot probe binding may elevate background (*ii*). B) Target nucleic acids are converted into nucleic acid reporter molecules carrying tag pairs (*i*). Reporter molecules are amplified and ligated to a tag microarray (*ii*). Signals are generated by amplifying reporter molecules circularized by ligation with rolling circle amplification. Cross-hybridizing reporter molecules and off-spot reporter molecules do not generate signals (*iii*).