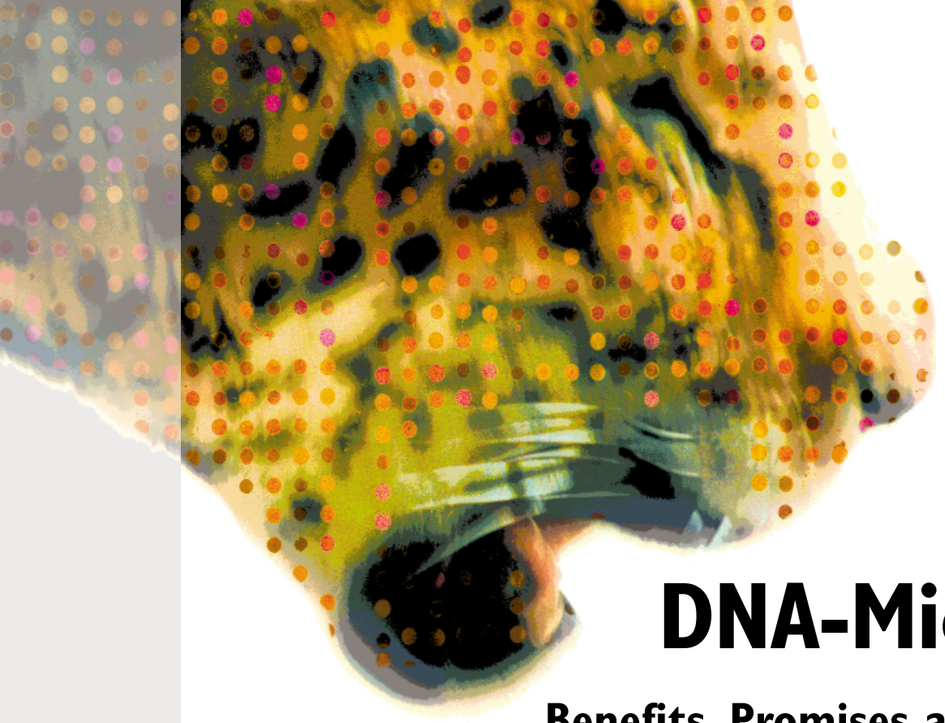


Jörg D. Hoheisel

Marcel J. Scheideler

# DNA-Microarray Analyses

## Benefits, Promises and Problems

**DNA-microarray technology is dramatically changing studies in molecular biology, permitting analyses on a global scale. There is a lot of enthusiasm about its power, since it brings us few (of still very many) steps closer to the understanding of the very complex and interrelated biological processes that define cellular life. However, despite – or rather because of – all the well-founded enthusiasm, one should also be aware of the technique's limitations in order to take proper advantage of the great chances offered by this technology.**

One important accomplishment in the bio-technical and bio-medical area is the development and diversification of microarray technologies. Basically, microarray assays are nothing else but parallelised blotting techniques. In contrast to earlier methods, generally the status of blotted and free-floating partners is turned on its head in the analysis in order to achieve parallelism. Currently, wide-spread applications include studies on transcriptional variations and genotyping experiments for diagnostic purposes, for example. Various processes for production and detection exist. The principle, however, remains the same. Pieces of nucleic acids are arranged in an ordered grid that allows an immediate allocation of a binding event to the relevant sequence. Upon incubation with a usually rather complex mixture of sample molecules, mostly again nucleic acids, a sorting process takes place by formation of a duplex between complementary molecules. The result is visualised as an intensity pattern produced by attached labels such as fluorescence dyes. A thorough quality assessment of data sets of this size and complexity is not a trivial task, especially if done by others. Therefore, false or only partially correct information could quickly accumulate. In such a scenario, the technical advance would not necessarily enhance science but – in the extreme case – could actually even slow it down. Careful consideration of microarray data, alertness to the problems attached, defined standards and the free exchange of the full data sets by means of a central depository are therefore necessary to make use of the techniques' full potential.

### Microarray design

Microarrays come in various formats, mainly different in the kind of sensor molecule. Each design has its own characteristics in sensitivity, accuracy and reproducibility but also with respect to the kind of assays which are sensibly possible. However, an evaluation of microarray systems and their results requires not only a close look at design and production – still frequently the emphasis of quality-assuring measures – but at the complete workflow involved (Fig. 1). Short oligonucleotides have a high specificity. For quantification, however, redundant information is necessary, since individual oligomers differ strongly in their hybridisation behaviour. Also, their selection can be anything but trivial. These effects are less pronounced with oligonucleotides longer than about 50 nucleotides, but at a cost of specificity, rendering them useless for typing experiments. PCR-products or other DNA-fragments, finally, are least affected by stability and kinetic differences but most prone to cross-hybridisation events.

Another difference in design is set by the actual application. Global analyses are mostly aiming at the accumulation of new (basic) knowledge, such as information on genes involved in the establishment of a disease. Alternatively, functionally defined sets of probes may be assembled on an array in order to cut cost and simplify analysis. However, preselection risks that important and relevant information is missed. For routine screening of large sample numbers, usually focussed arrays with a well-defined probe set of little complexity are being used.

A third categorisation factor is, whether the design is based on known sequence or not. To date, 95 completed genome sequences are available in public databases, of which 16 are archaeal, 69 bacterial and 10 eukaryotic genomes, and another 529 projects are underway (http://ergo.integratedgenomics.com/GOLD). Knowledge of genomic sequence allows the conscious selection of capture probes, if the annotation is sufficient. Alternatively, one can design arrays without sequence information. For transcriptional profiling, for example, this comprises arrays made of cDNAs or genomic shotgun clones. Pre-

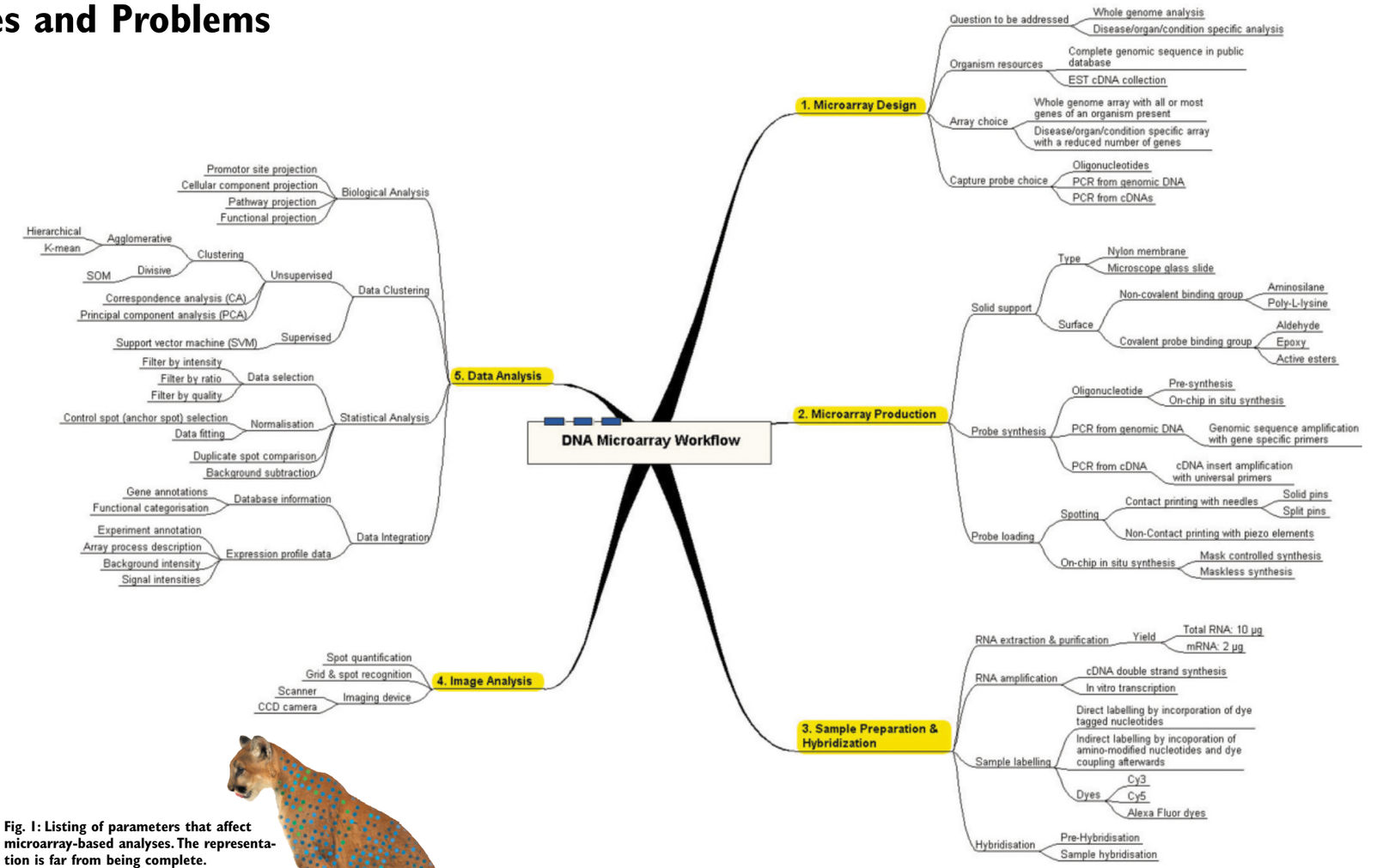





**Fig. 1: Listing of parameters that affect microarray-based analyses. The representation is far from being complete.**

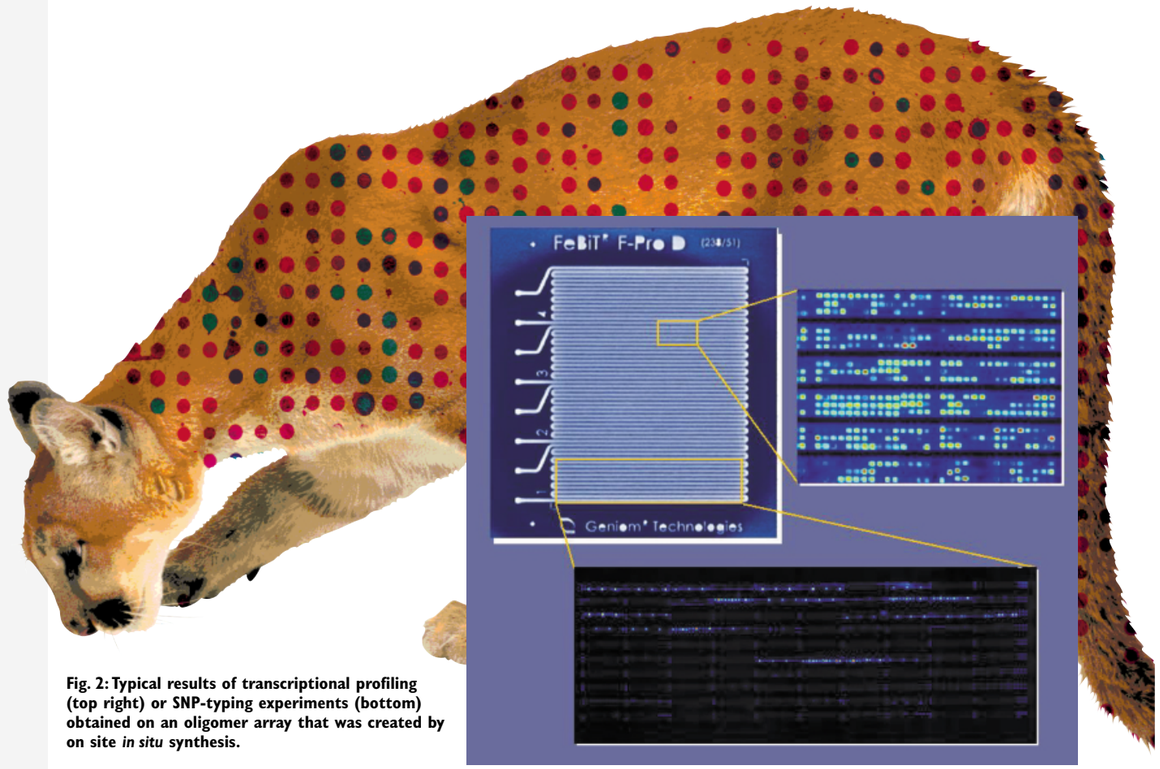



**Fig. 2: Typical results of transcriptional profiling (top right) or SNP-typing experiments (bottom) obtained on an oligomer array that was created by on site *in situ* synthesis.**

selection by partial (EST-) sequence produces a viable, intermediate format. As a matter of fact, a minimal tiling path of genomic fragments might well be the optimal format for such studies, especially on microbial organisms, since by definition no open reading frame is missing and intergenic regions could be analysed on the very same microarray. For genotyping experiments, the equivalent is represented by comprehensive oligomer arrays consisting, for instance, of all 16,384 possible heptamers.

## Production matters

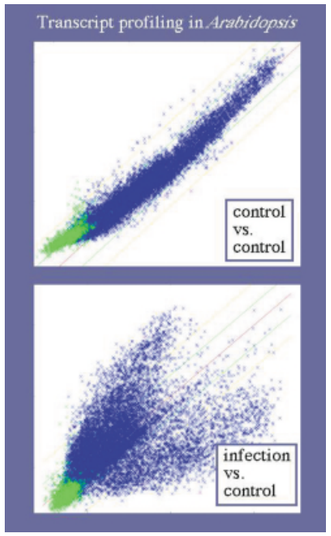
The above sub-heading can be read two ways, both to be covered briefly. Production consists of two aspects: solid support as well as capture probe synthesis and immobilisation. Nylon membranes were initially used. Mainly for the aspects of miniaturisation and automation, glass became the favourite surface. With other detection modes – mass spectrometry, measurement of impedance or the electric potential – other surfaces will become more appropriate. While monolayers are superior in most aspects, the amount of attached probe molecule can be limiting, thus influencing kinetics and equilibrium, which in turn affect sensitivity and accuracy.

Another important aspect in production is the purity of the probe molecules. Protocols are available, for example, for avoiding purification of PCR-products prior to spotting, without a significant loss in performance. This not only avoids the expense of purification but also the tremendous loss of material during this process. Oligonucleotide pre-fabrication allows for purification, either prior to or during attachment, the latter achieved by taking advantage of active groups present only in full-length molecules, a method that is also applicable to *in situ* synthesised oligomers. For some assays, impurity can be neglected anyway. Shorter oligonucleotide derivatives, for example, usually do not affect typing quality other than in terms of sensitivity. The dynamic range of a measurement, on the other hand, is strongly influenced.

Throughput and flexibility are other critical factors. Application of pre-fabricated probes limits flexibility merely because of the cost involved in producing new molecules. *In situ* synthesis can offer extra flexibility, if the control mechanism does. Photo-chemistry, for example, permits high throughput production. In case of photolithography, however, flexibility is limited by the expensive masks. Mask-free *in situ* synthesis permits a fully flexible design, coming on-line in central, large-scale facilities and even for on-site microarray production, the ultimate in flexibility (Fig. 2).

## Hybridisation samples

In many microarray assays, genomic DNA is considered a static medium.

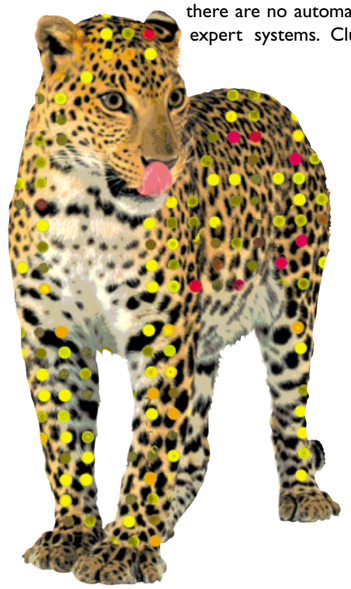



However, there are differences of dynamic nature, introduced by re-arrangements, for example, or reversible modifications such as methylation. The latter, as many other sequence features, may also have an effect only in combination with structural variants. Topological constrains in DNA, for instance, can be essential regulative factors. At current, they cannot be analysed, however. Therefore, some data obtained from microarray analyses might only be descriptive or even not open to interpretation at all for the lack of the appropriate assay.

For analyses on dynamically varying – parameters e.g. the amount of RNA present in a tissue – sample handling has a large effect on the eventual analysis. Only an appropriate data annotation, which in turn requires a commonly accepted ontology, will really make data comparable, irrespective of the problem caused by the various procedures and platforms that are in use. The MGED standards (www.mged.org) formulated for transcriptional analyses, for example, only represent an initial step toward that goal. What is eventually needed for quantitative analyses are algorithms that permit the determination of the absolute amount of each individual molecule type, an achievement not beyond imagination.

## Data analysis

Microarray experiments provide unprecedented quantities of data. Their interpretation should therefore be based on three main processes: statistical quality analysis, data integration and a presentation that makes it accessible to human thinking. Because of the shear wealth of information, only statistical procedures allow an assessment and filtering of microarray data. Data integration is essential, since tying connections that make subsequent interpretation feasible. This requires a modular data warehouse concept combining the storage of data like raw signal intensities, gene annotations and their functional categories as well as experimental annotations. For later queries, the annotations should be determined by a pre-defined and catalogued vocabulary. Sophisticated computational tools for data visualisation and reduction of data complexity are important to make the information accessible to a human mind, at least as long as there are no automatic expert systems. Clustering, such as correspondence analysis, and algorithms that project the data onto biologically relevant information like pathways, molecular functions, cellular components or promoter sites represent tools to such end.



The capability of combining freely individual experiments gets increasingly important the more data becomes available. Only then, new queries are made possible, which had not been considered when producing the results in the first instance. Currently, most information remains unused, even if the results are published, since most data are irrelevant to the specific cellular phenomenon that is discussed in detail. Only a recycling of data will fully release the gigantic potential of microarray analyses.

## Personalised medicine

Frequently, it is claimed that microarrays will soon allow a more personal disease diagnosis followed by a tailor-made medical treatment. Expression analyses on cancer tissues, for example, demonstrated that patients that seemingly suffered from the same form of cancer actually fall into distinct molecular sub-classes. Chip-based comparative genome hybridisation (CGH) produced patterns of amplification or loss of genomic regions, which were different between groups of patients and indicative of disease development, therefore being useful to prognostic purposes. While such approaches are very valuable and exciting, one should be aware that more comprehensive analyses, based on sample sizes, collection procedures and annotation practices of epidemiologists are needed in many cases before results of wider implication can be expected. The same is true for typing analyses. In addition, a set of single nucleotide polymorphisms (SNPs) needs to be defined for each assay. Millions of SNPs are known in the human genome. Which of these are highly relevant, however, is still being determined. And also the production and – once applied simultaneously to the arrays – detection of large sample numbers is not trivial, but nevertheless needed for taking advantage of the parallelism of microarrays. Advanced procedures such as an on-chip polymerase reaction are bound to make this profiling routinely possible. In an adaptation to SNP-typing, even cellular regulative processes can be studied. Un-methylated cytosine, for example is transformed into uracil and then – upon PCR-amplification – thymidine when treated with bisulfite, while methylated cytosine remains unaffected. Thereby, a polymorphism is artificially introduced and made detectable by comparing treated and untreated samples.

However, will the above-mentioned studies immediately lead to personal disease treatments or even drugs? The answer is "no". For once, no company currently has the capacity to create all these drugs. Second, even if individualised drugs could be produced, most likely they would not be produced for the astronomical cost involved. Solving this dilemma, as possibly by means of combinatorial techniques, is certainly a promising field for innovative developments. Without immediate personal remedy, is having a personal molecular diagnosis therefore a waste of time? Again, the answer is a clear and emphatic "no". While tailor-made drugs for individual patients might be far off still or even never materialise, knowledge about molecular sub-groups will be useful right away. One only needs to think about the huge number of drugs that never made it to the patients due to side-effects. Frequently enough, however, these side-effects were restricted to few probants or patients, while the vast majority experienced good results. The ability to identify the group of patients, who should not be treated with a particular drug, could therefore increase the number of useful drugs immensely. Both the health of the patients and the commercial success of drug companies will gain from this, the latter being able eventually to make revenues from drugs they had invested in heavily without return.

## Prospects

Despite of some limitations, the power and width of DNA-microarray technology made it a predominant factor in genomics, transcriptomics, pharmacogenomics and systems-biology, simultaneously getting ever more important in pre-clinical research and meanwhile even clinical studies. Additionally, the principle is also put to good use in other areas of investigation, including studies on proteins, metabolic compounds or small chemical entities.

**References are available from the authors.**

**Dr Marcel J. Scheideler**
**Dr Jörg D. Hoheisel**
**Functional Genome Analysis**
**Deutsches Krebsforschungszentrum**
**Im Neuenheimer Feld 506**
**69120 Heidelberg**
**Germany**
**Phone +49 6221 42 4680**
**Fax +49 6221 42 4687**
**j.hoheisel@dkfz.de**
**www.dkfz.de/funct_genome**