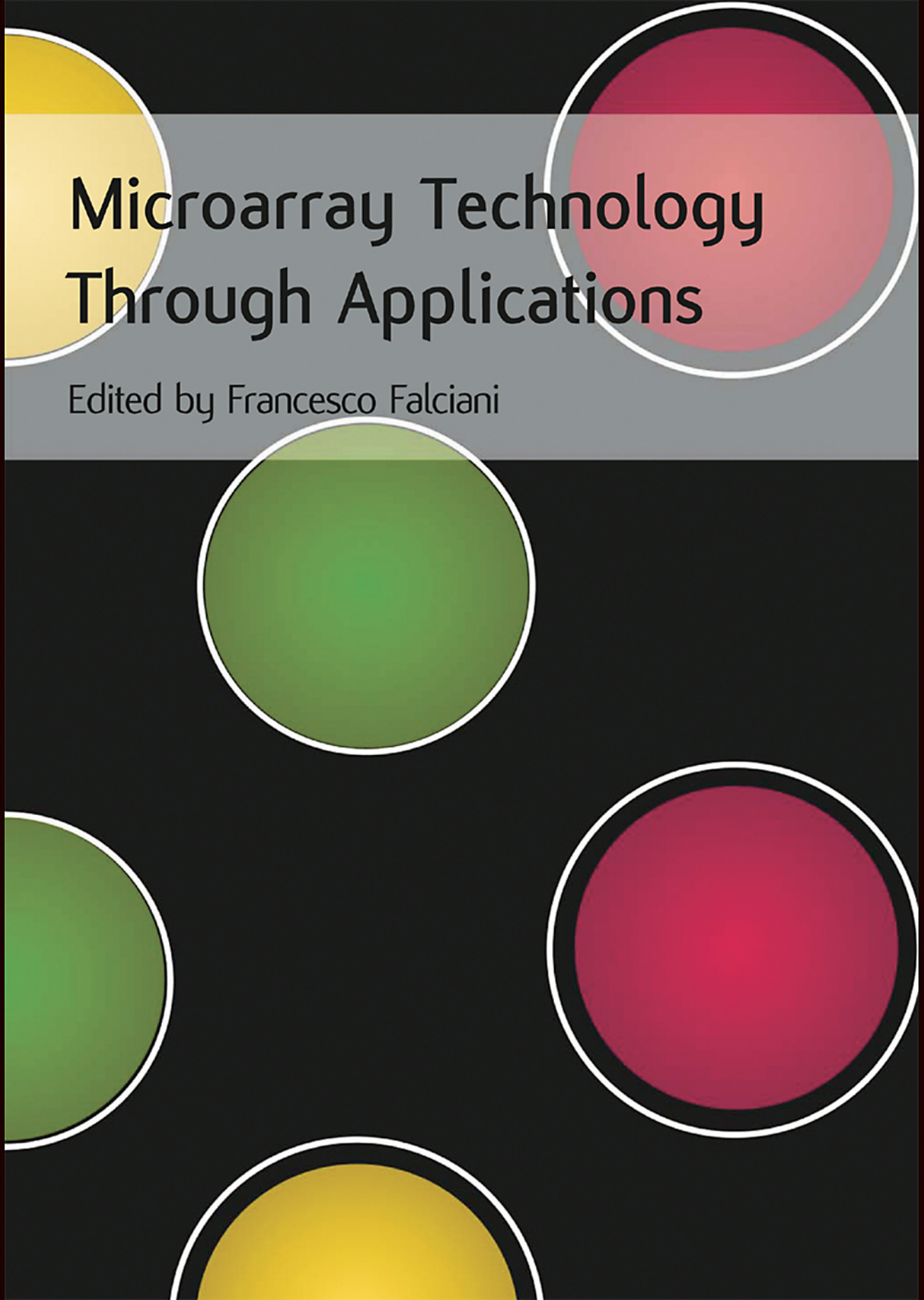
# Microarray Technology Through Applications

Edited by Francesco Falciani



# Microarray Technology Through Applications

# Microarray Technology Through Applications

**Francesco Falciani (Editor)**

School of Biosciences
University of Birmingham
Birmingham, UK

Taylor & Francis
Taylor & Francis Group

Published by:

**Taylor & Francis Group**

In US: 270 Madison Avenue
New York, N Y 10016
In UK: 2 Park Square, Milton Park
Abingdon, OX14 4RN

© 2007 by Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2007.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

ISBN 0–203–96727–5 Master e-book ISBN

ISBN: 9780415378536 or 0-4153-7853-2

All rights reserved. No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

A catalog record for this book is available from the British Library.

**Library of Congress Cataloging-in-Publication Data**

Microarray technology through applications / F. Falciani (ed.).
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-415-37853-6 (alk. paper)
1. DNA microarrays. I. Falciani, F. (Francesco)
[DNLM: 1. Gene Expression Profiling. 2. Oligonucleotide Array Sequence Analysis—methods. 3. Genomics—methods. 4. Research Design. QU 450 M6258 2007]
QP624.5.D726M5143 2007
572.8'636—dc22

2007000530

Editor: Elizabeth Owen
Editorial Assistant: Kirsty Lyons
Production Editor: Karin Henderson

T&F informa

Taylor & Francis Group, an informa business

Visit our web site at http://www.garlandscience.com

# Contents

**Table 0.1** Key to link cases to protocols and supplementary information

| Chapter | Protocols | Additional material |
|---|---|---|
| Chapter 1 | Protocols 1, 2 and 3 | Appendix 1 |
| Chapter 2 | Protocol 5 | |
| Chapter 3 | Protocol 6 | Appendix 2 |
| Chapter 4 | Protocol 8 | |
| Chapter 5 | Protocol 7 | |
| Chapter 6 | Protocols 2 and 3 | Appendix 4 |
| Chapter 7 | | |
| Chapter 8 | Protocols 2 and 3 | Appendix 3 |
| Chapter 9 | | |

# Contributors

**Fátima Al-Shahrour**, Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

**Sara Alvarez**, Molecular Group Cytogenetics, Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**David Blesa**, Molecular Group Cytogenetics, Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Christopher Buckley**, MRC Centre for Immune Regulation, Division of Immunity and Infection, The University of Birmingham, Birmingham, UK

**Thomas Carzaniga**, Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy

**Juan C. Cigudosa**, Molecular Group Cytogenetics, Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Ana Conesa**, Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias, Valencia, Spain

**Chrystala Constantinidou**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Gianni Dehò**, Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy

**Joaquín Dopazo**, Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

**Francesco Falciani**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Javier Forment**, IBMCP, Technical University of Valencia, Valencia, Spain

**José Gadea**, IBMCP, Technical University of Valencia, Valencia, Spain

**Anna González-Neira**, Genotyping Unit. Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Nick Haan**, BlueGnome, Cambridge, UK

**Jon L. Hobman**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Antony Jones**, Functional Genomics Laboratory, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

**Cristina Largo**, Molecular Group Cytogenetics, Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Steve D. Minchin**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Shabnam Moobed**, Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA

**Mercedes Robledo**, Hereditary Endocrine Cancer Group. Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Sandra Rodríguez-Perales**, Molecular Group Cytogenetics, Spanish National Cancer Centre (CNIO), C/ Melchor Fernández Almagro 3, E-28029, Madrid, Spain

**Mike Salmon**, MRC Centre for Immune Regulation, Division of Immunity and Infection, The University of Birmingham, Birmingham, UK

**Donatella Sarti**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Nigel J Saunders**, Bacterial Pathogenesis and Functional Genomics Group, Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

**Graham Snudden**, BlueGnome, Cambridge, UK

**Lori A. S. Snyder**, The Sir William Dunn School of Pathology, University of Oxford, UK

**Victor Trevino**, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

**Joseph T. Wade**, Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA

**Jeroen van Dijk**, RIKILT, Wageningen, the Netherlands

**David Wild**, Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA and Warwick Systems Biology Centre, Coventry House, University of Warwick, Coventry CV4 7AL, UK

**Donling Zheng**, King's College London, Department of Life Sciences, Room 3, 132 Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NN, UK

# Abbreviations

| | |
|---|---|
| **ALO** | allelic-specific oligonucleotide |
| **AML** | acute myeloid leukemia |
| **BAC** | bacterial artificial chromosome |
| **BCCP** | biotin carboxyl carrier protein |
| **BSA** | bovine serum albumin |
| **CBM** | carbohydrate-binding molecule |
| **CCD** | charge-coupled device |
| **CC-RCC** | clear-cell RCC |
| **cDNA** | complementary DNA |
| **CGH** | comparative genomic hybridization |
| **ChIP** | chromatin immunoprecipitation |
| **CMOS** | Complementary metal-oxide-semiconductor |
| **CRP** | cAMP receptor protein |
| **CTAB** | hexadecyltrimethylammonium bromide |
| **CV/CD** | common variant/common disease (hypothesis) |
| **DIP-chip** | DNA immunoprecipitation chip |
| **DMD** | digital micromirror device |
| **DNA** | deoxyribose nucleic acid |
| **DOP** | degenerated oligonucleotide priming |
| **ELISA** | enzyme-linked immunosorbent assay |
| **ESE** | exomic splicing enhancer |
| **EST** | expressed sequence tag |
| **FDR** | false discovery rate |
| **FISH** | fluorescence *in situ* hybridization |
| **GalNAc** | *N*-acetylgalactosamine |
| **GC** | gas chromatography |
| **GMAT** | genome-wide mapping technique |
| **GST** | glutathione-*S*-transferase |
| **kb** | kilobasepairs |
| **LC** | liquid chromatography |
| **LC–MS** | liquid chromatography–mass spectrometry |
| **LD** | linkage disequilibrium |
| **LPS** | lipopolysaccharide |
| **MAGE** | microarray gene expression |
| **MAGE-OM** | MAGE object model |
| **MAGE-Stk** | MAGE software toolkits |
| **MALDI-TOF** | matrix-assisted laser desorption ionization/time-of-flight |
| **MAS** | maskless array synthesizer |
| **Mb** | megabasepairs |
| **MBP** | maltose-binding protein |
| **MGED** | Microarray gene expression data (society) |
| **MIAME** | minimum information about a microarray experiment |
| **MM** | mismatch oliogmer |
| **MTC** | medullary thyroid cancer |
| **Neu5Ac** | *N*-acetylneuraminic acid |
| **NMR** | nuclear magnetic resonance |
| **NHS** | *N*-hydroxysuccinimide |
| **NTA** | $Ni^{2+}$-Nitrailotriacetate |
| **OGT** | Oxford Gene Technology |
| **ORF** | open reading frame |
| **PBM** | protein-binding microarray |
| **PC** | principal component |
| **PCA** | principal component analysis |
| **PCR** | polymerase chain reaction |
| **PI** | predictive interval |
| **PM** | perfect match |
| **PMT** | photomultiplier tube |
| **PPI** | protein/protein interaction |
| **PVDF** | polyvinyldifluoride |
| **Q-RT-PCR** | quantitative reverse-transcriptase PCR |
| **RCC** | renal cell carcinoma |
| **RIP-chip** | RNA immunoprecipitation chip |
| **ROMA** | run-off transcription microarray analysis |
| **RT** | reverse transcriptase |
| **SAGE** | serial analysis of gene expression |
| **SAR** | system acquired resistance |
| **SDS** | sodium dodecyl sulfate |
| **SELDI** | surface-enhanced laser desorption/ionization |

| | |
|---|---|
| **SELEX** | systematic evolution of ligands by exponential enrichment |
| **SNP** | single nucleotide polymorphism |
| **SOM** | self-organizing map |
| **SOTA** | self-organizing tree algorithm |
| **SSC** | salt–sodium citrate |
| **SSDNA** | salmon sperm DNA |
| **STAGE** | sequence tag analysis of genomic enrichment |
| **TAP** | tandem affinity purification |
| **TDT** | transmission/disequilibrium test |
| **TIFF** | tagged image file format (also known as TIF) |
| **TSA** | thymidine signal amplification |
| **TTS** | triplet-forming oligonucleotide target sequence |
| **UV** | ultraviolet |
| **VHL** | von Hippel–Lindau |
| **YAC** | yeast artificial chromosome |

# Preface

In the last few years biology has experienced unprecedented technological and conceptual developments. This has coincided with the widespread application of functional genomic technologies that allow monitoring the expression and interaction of thousands of genes, proteins and metabolites in single experiments. Because of the relatively low cost, microarray-based platforms have made the greatest impact.

*Microarray Technology Through Applications* is an initial introduction to the wide spectrum of microarray technologies for the non-expert and will also be a useful tool to the expert user who is looking for better understanding of specific issues behind experimental design and data interpretation.

Expression profiling was at first the only microarray-based technology to be of general use in the scientific community but in the last few years, we have seen a proliferation of microarray-based technologies and their application to a diverse set of tasks, ranging from studying the interaction of proteins with nucleic acids to measuring the concentration of a large number of proteins in biological samples. As an increasingly large number of research groups are considering using these approaches we felt that there was the need for a book that would provide a comprehensive overview of the theoretical and practical basis of microarray technology. This book will provide such an overview by means of a comprehensive introductory chapter (Chapter 1) followed by a series of case studies that represent the main application of this technology in biology. These case studies have been written by leaders in the field and describe prototypic projects simply and rigorously, with indication of how to generalize the approach to similar studies. The book is also designed to be a useful reference in the laboratory by providing a series of protocols for manufacturing and using microarrays in a number of applications (for a key to link case studies to protocols and supplementary information see Table 0.1 on page xi). Chapter 2 describes the use of immunoprecipitation techniques used in conjunction with DNA microarrays to determine the genome-wide binding profile of DNA-associated proteins. This technique (chip-on-chip) is now becoming a crucial tool for understanding gene regulation at a genome level and has applications from bacteria to human cells. Chapter 3 describes the use of microarray technology to compare the structure of different genomes. The case study in this chapter reports the application of this technique to identify chromosomal deletions and rearrangements in the genome structure of healthy versus diseased human cells. The application of this approach to the comparison of bacterial genomes and more generally in the analysis of comparative genomics datasets is discussed in Appendix 2. In Chapter 4, the use of microarray technology for identifying single nucleotide polymorphisms (SNPs) is discussed. Chapter 5 explains a recently developed technique for the identification of direct targets of transcription factors. This technique is based on the analysis of RNA produced from an *in vitro* transcription reaction that is catalyzed by the addition of a purified transcription factor. Chapter 6 contains two case studies with relevance to bacterial pathogenesis and provides the necessary information for the design, execution and analysis of experiments with two-color and single channel arrays.

Chapter 7 describes a set of bioinformatics tools designed to facilitate the biological interpretation of the result of a microarray experiment. Such tools are of fundamental importance

of functionally annotate large functional genomics datasets. Chapter 8 exemplifies the application of microarray technology in agricultural research. This case study represents a potential practical application of expression profiling that uses a combination of some of the techniques described in previous two chapters. Chapter 9 deals with an application of microarray technology in proteomics providing indications on the potential issues behind this application.

Francesco Falciani

# Useful links

| Resource | Site | Remarks |
|---|---|---|
| **Construction of cDNA library** | | |
| Lambda ZAP | www.stratagene.com | Commercial Product |
| CloneMiner | www.invitrogen.com | Commercial Product |
| SMART | www.clontech.com | Commercial Product |
| **Clone LIMS and tracking** | | |
| AlmaZen | almazen.bioalma.com | Commercial Product |
| CloneTracker | www.biodiscovery.com/index/clonetracker | Commercial Product |
| B.A.S.E. | base.thep.lu.se | OpenSource Product |
| **EST trimming, assembly, and processing pipelines** | | |
| *phred, phrap, cross-match* | www.phrap.org | Trimming algorithms |
| *lucy* | www.tigr.org/software/ | Trimming algorithms |
| CAP3 | genome.cs.mtu.edu/cap/cap3.html | Trimming algorithms |
| CLU | compbio.pbrc.edu/pti | Trimming algorithms |
| d2-cluster | www.ccb.sickkids.ca/dnaClustering.html | Clustering algorithm |
| TGICL | www.tigr.org/tdb/tgi/software/ | Clustering algorithm |
| ESTIMA | titan.biotec.uiuc.edu/ESTIMA/ | OpenSource Product |
| PHOREST | www.biol.lu.se/phorest | OpenSource Product |
| ESTWeb | bioinfo.iq.usp.br/estweb | OpenSource Product |
| ESTAnnotator | genome.dkfz-heidelberg.de | OpenSource Product |
| PipeOnline | bioinfo.okstate.edu/pipeonline/ | OpenSource Product |
| ESTree | www.itb.cnr.it/estree/process.php | OpenSource Product |
| **Microarray printing** | | |
| Genomic Solutions | www.genomicsolutions.com | Commercial Product |
| Corning | www.corning.com | Commercial Product |
| ArrayIt | www.arrayit.com | Commercial Product |
| **Annotation** | | |
| Gene Ontology | www.geneontology.org | Gene Ontology Consortium |
| InterPro | www.ebi.ac.uk/interpro | Protein Family Annotation |
| Blast2GO | www.blast2go.de | Freeware tool for GO annotation |
| **General database systems** | | |
| MySQL | www.mysql.org | OpenSource Product |
| PostgreSQL | www.postgresql.org/ | OpenSource Product |

| | | |
|---|---|---|
| Oracle | www.oracle.com/database/index.html | Commercial Product |
| **Plant genomics databases** | | |
| PLANET | mips.gsf.de/projects/plants/PlaNetPortal/databases.html | Network of European Plant Databases |
| TAIR | arabidopsis.org/ | The Arabidopsis Information Resource |
| MAIZEGDB | www.maizegdb.org/ | Maize genome database, homepage |
| GRAMENE | www.gramene.org/ | Resource for Comparative Grass Genomics |
| SoyBase | soybase.agron.iastate.edu/ | Soybean database project |
| **Microarray sites and data repository** | | |
| MGED | www.mged.org/ | International Microarray Gene Expression Data Society |
| MIAME | www.mged.org/Workgroups/MIAME/miame.html | Minimum information about a microarray experiment, published by the MGED |
| SMD | genome-www5.stanford.edu/index.shtml | Homepage of the Stanford microarray database of Stanford University |
| ArrayExpress | www.ebi.ac.uk/arrayexpress/ | EBI repository |
| RED | red.dna.affrc.go.jp/RED/ | Rice Expression Database |
| NASCArrays | affymetrix.arabidopsis.info/ | International Affymetrix transcriptomics service |
| **Microarray protocols** | | |
| General information | www.microarrays.org | Site with information on microarrays, University of California at San Francisco |
| KRL | www.rbhrfcrc.qimr.edu.au/kidney/Pages/Microarray protocol.html | Microarray protocols used by the kidney research laboratory (Australia) |
| Microarray protocols | research.nhgri.nih.gov/microarray/hybridization.shtml | Microarray protocols used by the National Human Genome Research Institute |
| **Data analysis** | | |
| Bioconductor | www.bioconductor.org | Freeware Project for Genomics Data Analysis |
| GEPAS | gepas.bioinfo.cnio.es | Web resource |
| CyberT | visitor.ics.uci.edu/genex/cybert | Web resource |
| Expressionist | www.genedata.com | Commercial Product |
| Rosetta | www.rosettabio.com/products/resolver/default.htm | Commercial Product |
| **Institutions** | | |
| CropNet | flora.life.nottingham.ac.uk/agr/ | UK Bioinformatics Resource for Crop Plants |

| | | |
|---|---|---|
| ILSI | www.ilsi.org/ | International Life Sciences Institute |
| NCBI | www.ncbi.nlm.nih.gov/ | National Center for Biotechnology Information |
| EBI | www.ebi.org | European Bioinformatics Institute |
| TIGR | www.tigr.org/ | The Institute for Genomic Research, homepage |

# Introduction to microarray technology

# 1

*Jon L. Hobman, Antony Jones, and Chrystala Constantinidou*

'Man is a tool-using animal...Without tools he is nothing, with tools he is all' – Thomas Carlyle (1795–1881)

'The mechanic who wishes to do his work well, must first sharpen his tools.' – *Analects of Confucius* 15: 9

## 1.1 Introduction to the technology and its applications

### 1.1.1 Microarrays as research tools

Large-scale DNA sequencing projects and the completion of increasing numbers of genome sequences is having a major impact on biological research. The 'post-genomic era' is characterized by exploitation of genomic DNA sequence data as a research resource, and the use of high throughput experimental methods to study organism-wide events and interactions. These technical advances combined with increasing amounts of available genomic data have started to influence the direction that biological research is taking. Until quite recently, there has been a concentration on the reductive ('bottom up') view of understanding how an organism grows, adapts to changing conditions, or interacts with other organisms. This has been achieved by research groups studying single genes or regulons, or by determining the structure and function of small numbers of proteins, or by studying interactions between small numbers of cellular components. The use of the data generated in these experiments has led us to an understanding of how many cellular components work, and how some of these components interact with each other. However, in much the same way that understanding what a component in a radio does, or what happens when that component part of that radio is damaged or removed, does not lead to an understanding of how the radio works, we are faced with similar problems in describing how organisms work by looking at their components in isolation (Lazebnik, 2002). Now, there is a momentum towards the whole organism view of biology (Twyman, 2004a), using the holistic ('top down') approach of trying to understand how an organism works in its entirety and how the networks of physical and functional interactions occur between gene promoters, proteins, and noncoding RNAs (Brasch *et al.*, 2004). Attempts to understand the whole organism have led to the emergence of systems biology as a new cross-disciplinary research area, which encompasses experimental research, systems and control theory,

bioinformatics, and theoretical and computational model building and prediction.

The development of appropriate technologies (experimental research tools) that exploit genomic data is playing a major role in the development of whole organism studies, and will begin to allow us to dissect networks of gene regulation (transcriptomics), understand protein production patterns and interactions with other proteins (proteomics), study the interactions of small molecules with proteins (chemical genomics), and start to catalog the small molecules and metabolites found in cells during normal and abnormal function (metabolomics).

One of the most important research tools used in transcriptomics and proteomics studies is the 'array', which is a powerful, high-throughput, massively parallel-assay format used for studying interactions between biological molecules. Arrays are a good example of how an advance in technology has allowed researchers to test hypotheses and interrogate organisms on a scale which prior to the development of this technology would have been impossible.

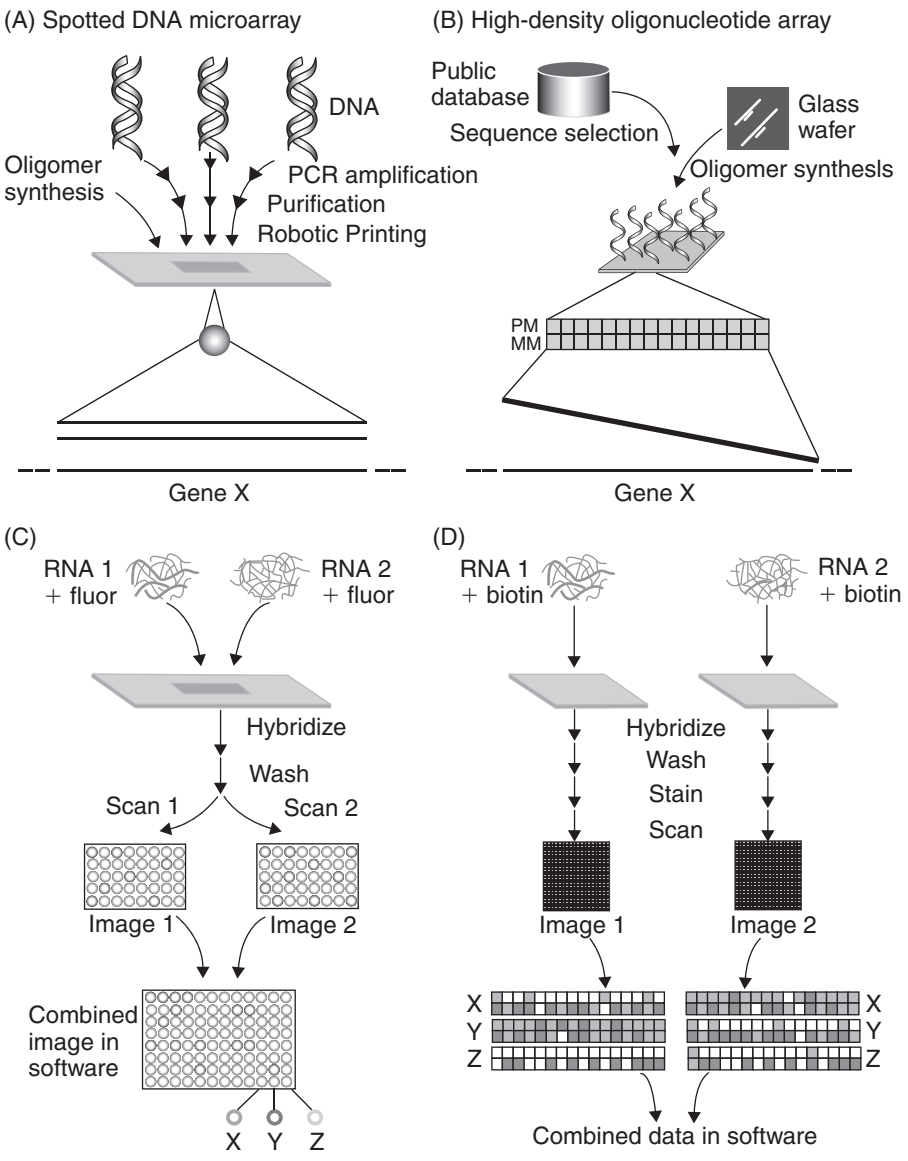
### 1.1.2 Arrays and microarrays

The use of ordered arrangements or 'arrays' of spatially addressed molecules in parallel assays is becoming an increasingly popular technology for studying interactions between biological molecules. These assays are commonly referred to as arrays, and by extension assays that have been miniaturized to a small format are called microarrays. The biological materials used in these arrays can be nucleic acids, proteins or carbohydrates, whilst arrays of chemicals and other small molecules have also been made. These materials are most commonly deposited on the surface of a planar solid substrate in an ordered arrangement so that each positional coordinate where material has been deposited contains material that represents a single gene or protein or other molecule. The most widely used solid substrates for arrays are nylon or nitrocellulose membranes, glass slides or silicon/quartz materials.

For DNA arrays, as the numbers of different nucleic acid molecules that are printed on arrays has become larger, so that the array represents the whole genome of the organism, and each spot on the array represents a single gene, the imperative has been to miniaturize the array format. There are two drivers for miniaturization: the first is simply so that all of the features (e.g. spots of DNA, each representing one gene) could be fitted onto a conveniently sized solid substrate, and the second is to use smaller and smaller amounts of biological material on the arrays, because high throughput methods of sample preparation tend to be restricted in the amount of biological material that can be purified using them. Concomitant with miniaturization of the array there has also been a trend towards decreasing the amount of biological material (e.g. RNA) used in the array experiments, as the less material that needs to be used, the easier it is to extract and purify it, and the more economical it is in terms of reagent costs.

There are several widely established formats that are used for DNA microarrays. These array types fall into two categories: those that are constructed within laboratories, and those that are produced under industrial manufacturing conditions by commercial companies. The first type,

developed by the Pat Brown lab at Stanford University (Schena *et al.*, 1995, 1996; DeRisi *et al.*, 1997; Heller *et al.*, 1997) is the so-called 'home brew' or 'roll your own' glass slide microarrays, which are produced in-house, often in a core facility. The most popular technology for printing in-house arrays appears from anecdotal evidence to be contact printing, which is used by a large number of university research laboratories, and will be covered in some detail in this section (*Figure 1A*). The second format is the manufactured array, of which the best known is the Affymetrix GeneChip™ format, which is discussed in more detail in Section 1.4 (*Figure 1B*). In addition to these two well-known array formats there are other formats offered by commercial companies, such as Agilent, Nimblegen, Oxford Gene Technology, Xeotron, Combimatrix, Febit, and Nanogen. Each of these formats is more or less related in concept to the spotted array or Affymetrix formats. Agilent technologies (http://www.chem.agilent.com/Scripts/PCol.asp?lPage=494) have developed a method for depositing long oligonucleotides (60-mers) on to glass slides, using ink jet printing (Hughes *et al.*, 2001). This printing method is also used by Oxford Gene Technology (OGT) to create arrays (http://www.ogt.co.uk/). OGT is a company created by Professor E.M. Southern, which owns fundamental European and US patents on microarray technology. Nanogen (http://www.nanogen.com/) have developed a method of electronically addressing oligonucleotides to positions on a chip, and enhancing hybridization by using electronic pulsing (Edman *et al.*, 1997; Sosnowski *et al.*, 1997; Heller *et al.*, 1999, 2000). Nimblegen (http://www.nimblegen.com/technology/) use a proprietary maskless array synthesizer (MAS) technology, to synthesize high density arrays, using photodeposition chemistry for oligonucleotide synthesis on a solid support. This system uses a digital micromirror device (DMD) in which an array of small aluminum mirrors are used to pattern over 750 000 pixels of light. The DMD creates 'virtual masks' at specific positions on a microarray chip that protects these regions from UV light that is shone over the array surface. In positions on which the UV light shines, it deprotects the oligonucleotide strand already synthesized, allowing the addition of a new nucleotide to the lengthening oligonucleotide (Nuwaysir *et al.*, 2002; Albert *et al.*, 2003). Nimblegen uses short oligonucleotide (25-mer) technology in their arrays, and are producing high density tiling arrays for resequencing and ChIP-chip experiments (see Section 1.1.4). Xeotron (who have recently been acquired by Invitrogen http://www.invitrogen.com/content.cfm?pageid=10620) use a proprietary platform technology for synthesis of DNA microarrays. The arrays are made by *in situ* parallel combinatorial synthesis of oligonucleotides in three-dimensional nano-chambers. The process of oligonucleotide synthesis uses photogenerated acids to deprotect oligonucleotide capping, and uses digital projection photolithography to direct deprotection and parallel chemical synthesis (Gao *et al.*, 2001; Venkatasubbarao, 2004). This method has also been used to produce peptide arrays and can be used for other syntheses. Combimatrix (http://www.combimatrix.com/) uses a different technology to generate the acids used to detritylate capped oligonucleotides during *in situ* phosphoramidite synthesis. Rather than using light-directed acid generation, Combimatrix uses a specially modified 'CMOS' semiconductor to direct synthesis of DNA in response to a digital command. Each feature on the




array is a microelectrode, which can selectively electrochemically generate acid, during oligonucleotide synthesis using phosphoramidite chemistry. Febit (http://www.febit.de/index.htm) market an all-in-one machine that synthesizes oligonucleotide arrays using maskless light activated synthesis of microarrays controlled by a digital projector, hybridizes the fabricated arrays, and analyzes the data. New microarray technologies such as nonplanar DNA microarrays made by companies such as Illumina (http://www.illumina.com), PharmaSeq (http://www.pharmaseq.com), and SmartBead Technologies (http://www.smartbead.com) are constantly evolving, and a recent review of the state of the art and future prospects

**Figure 1.1**

Expression analysis experiments using spotted glass DNA microarrays, and Affymetrix DNA microarrays. (A) Spotted glass microarrays are produced by the robotic spotting of PCR products, cDNAs, clone libraries or long oligonucleotides onto coated glass slides. Each feature (spot) on the array corresponds to a contiguous gene fragment of 40–70 nucleotides for oligonucleotide arrays, to several hundred nucleotides for PCR products. (B) Affymetrix high-density oligonucleotide arrays are manufactured using light directed *in situ* oligonucleotide synthesis. Each gene from the organism is generally represented by ten or more 25-mer oligonucleotides, which are designed to be a perfect match (PM) or a mismatch (MM) to the gene sequence. (C) For spotted arrays, gene expression profiling experiments commonly involve the conversion of RNA or mRNA to cDNA and labeling of the cDNA with a fluorescent dye for two samples. These are cohybridized to the probes on the array, which is then scanned to detect both fluorophores. The spots X, Y, and Z at the bottom of the image represent (X) increased levels of mRNA for gene X in sample 1, (Y) increased levels of mRNA for gene Y in sample 2, and (Z) similar levels of mRNA of gene Z in both samples. (D) During Affymetrix GeneChip transcription experiments, cRNA is biotinylated, and hybridized to the GeneChip. The GeneChip is then stained with avidin conjugated to a fluorophore, and scanned with a laser scanner. Results show: (X) Increased levels of expression of genes in sample 1, (Y) Increased levels of gene expression for sample 2, and (Z) similar gene expression levels for both samples. Reprinted from Harrington *et al.* (2000) *Curr Opin Microbiol* **3**: 285–291, with permission from Elsevier. (A color version of this figure is available at the book's website, www.garlandscience.com/9780415378536)

indicates that the evolution of the technology is proceeding rapidly (Venkatasubbarao, 2004).

Aside from differences in the method of array manufacture (deposition of prepared material versus *in situ* synthesis) the major differences between DNA array types is the nucleic acid material that is deposited onto the solid surface. Early-spotted DNA arrays deposited PCR products amplified from genes or open reading frames (ORFs), or spotted plasmid preparations from gene libraries, cDNAs, or expressed sequence tag clones (ESTs). As more complete genome sequences have been deciphered, complete genome sequence data is being used for the design of the materials deposited on the arrays. Primer design software can be used to design PCR primers to amplify regions from each gene from a sequenced genome for arraying (see http://colibase.bham.ac.uk/ as an example of a website that integrates genome analysis and primer design software tools), and PCR arrays have the advantage that they will represent both the sense and antisense strands of DNA. Single stranded oligonucleotide arrays by their nature can only be sense or antisense arrays, so their design requires careful thought because transcriptomics experiments use labeled complementary DNA (cDNA) made from mRNA to hybridize onto the array, so sense oligonucleotide arrays will hybridize to these cDNAs, but antisense ones will not. Genome sequence data and bioinformatics techniques have been used in the rational design and chemical synthesis of long oligonucleotides (40–100-mer) to represent genes on arrays, so that each has a matched melting temperature and length, which is claimed to improve the reliability of hybridization signals. Both of these two 'longmer' methods rely on a single, long nucleotide fragment to represent each gene or ORF on the array. The alternative strategy employed for DNA arrays (such as Affymetrix and Nimblegen arrays) has been the use of multiple short oligonucleotides (generally 25-mer) on an array to represent a gene. The use of high density oligonucleotide arrays such as these to

'tile' across a genome so that intergenic regions as well as ORFs and genes are represented by multiple oligonucleotides, has led to greater flexibility in the experiments that can be performed on these arrays, which is detailed below.

### 1.1.3 Principles of DNA array technology

All arrays are used as a tool to determine interactions between molecules immobilized to the solid surface, and molecules that are in a complex mixture in a solution, which is in contact with the array (*Figure 1.1*). Those interactions that occur are then detected and quantified. All DNA arrays harness the ability of nucleic acids with complementary sequences to hybridize to each other under suitable conditions. In DNA arrays, one nucleic acid is immobilized on a solid surface, and the nucleic acid in the hybridization solution is labeled with a radioactive isotope, chemical or dye molecule that can be detected quantitatively. Throughout this chapter we will be adopting the convention that the nucleic acid material tethered to the solid surface of the microarray is termed the 'probe', and the nucleic acid that is labeled by a reporter molecule such as a dye or isotope will be termed the 'target' (Phimister, 1999). The theory behind this nomenclature is that the 'probe' nucleic acids immobilized to the array surface are used to interrogate the complex mixture of labeled nucleic acids (e.g. cDNAs in transcriptomics experiments) for nucleic acids in the hybridization solution that are complementary to, and will associate with, the immobilized (tethered) probe. Unfortunately, this widely adopted nomenclature for description of probe and target in microarrays is the opposite of that commonly adopted (in our laboratory at least) for Southern/northern blotting, where a radioactively labeled DNA or cDNA probe would be used to hybridize to an immobilized nucleic acid on a nylon membrane. Clearly, this difference in nomenclature can lead to some confusion.

Key to DNA microarray technology is nucleic acid hybridization, which occurs where single stranded (denatured) nucleic acids are incubated together under conditions that promote the formation of base paired duplex molecules by C:G or A:T base pairing. The double stranded nucleic acid hybrids are therefore composed of sequences that are complementary to each other. Conditions that favor hybridization between nucleic acids can be promoted by manipulating time, temperature and ionic strength of the hybridization buffer (stringency) and will be affected by the concentration and complexity of the sample (Young and Anderson, 1985; Stoughton, 2005). Hybridizations are commonly conducted at low stringency (high salt concentration) often in the presence of formamide, because formamide allows the hybridization to be carried out at relatively low temperatures (42°C). The stability of the DNA hybrids is dependent on whether there are any mismatches in the nucleic acid duplex, with stability decreasing as the number of mismatches increases. The stability of the nucleic acid duplex determines how strongly attached the target DNA is to the immobilized probe nucleic acid, and how easy or difficult it is to wash away labeled target DNA from nonspecific binding to the attached probes. Washes to destabilize and consequently remove mismatched hybrids are conducted under increasing stringency (for practical purposes this is generally decreasing salt concentration and/or increasing temperature), so that nonspecific hybridization is minimized.

Solution nucleic acid hybridization was used widely in the 1960s and 1970s to determine DNA homology, and investigate nucleic acid structure (Anderson, 1999). The colony blot (Grunstein and Hogness, 1975), Southern blot (Southern, 1975), and dot blot were methods that were primarily used to identify identical or closely related DNA sequences by hybridization, and immobilized DNA onto a solid (though permeable) surface such as a nitrocellulose or later a nylon membrane, and then hybridized radioactively labeled denatured nucleic acids to the immobilized and denatured nucleic acids. The DNA macroarray is essentially a direct development from the dot blot and Southern blot. The Panorama® macroarrays marketed by Sigma-Genosys consist of PCR products immobilized on nylon membranes; where each PCR product represents all, or part, of a single gene. In transcriptomics experiments these immobilized PCR products are hybridized with radioactively labeled cDNA products, washed, and the resulting membrane scanned using a phosphor-imager system. For macroarrays, comparisons between two samples have to be conducted using two separate hybridizations, because the 'test' and 'control' samples are labeled with the same reporter molecule. Aside from miniaturization, the key difference between a standard glass slide DNA microarray experiment and experiments conducted using systems such as the Panorama® macroarray, is that the glass slide arrays use competitive hybridization between cDNAs derived from two sources, and the immobilized probe DNA. In these experiments, each cDNA from a different source (e.g. wild-type and mutant) is labeled with a different fluorescent dye. Competitive hybridization has a number of advantages: it allows for a direct comparison between two samples, e.g. a wild-type (control) and a mutant (test) on the same array, and very importantly from the point of view of data reliability and reproducibility, competitive hybridization overcomes irregularities in probe spot properties or local hybridization conditions on the array, which may adversely affect hybridization signal intensities. In competitive hybridization, the signal from each spot on the array is therefore a ratio of signals from both the control and test samples, where the control and test have undergone matched conditions. This means that even if the probe deposition, or hybridization conditions are irregular, accurate signal intensity ratios can be obtained (Stoughton, 2005).

Although the protein (Cutler, 2003), carbohydrate (Wang, 2003), and small molecule (Spring, 2005) microarray formats are different from DNA microarrays, and interactions between biological molecules will be different, the principle of how these other arrays work is essentially identical to the concept of DNA arrays in terms of the idea of using the immobilized biological molecule (antibodies, proteins, small molecules or carbohydrate) 'probes' to interrogate a complex mixture of labeled 'target' biological molecules for those that will bind to, or associate with them (see *Figure 1.2*).

### 1.1.4 Microarrays as tools for biological research applications

The major modes of use of DNA microarrays are:

- expression profiling,
- pathogen detection and characterization,

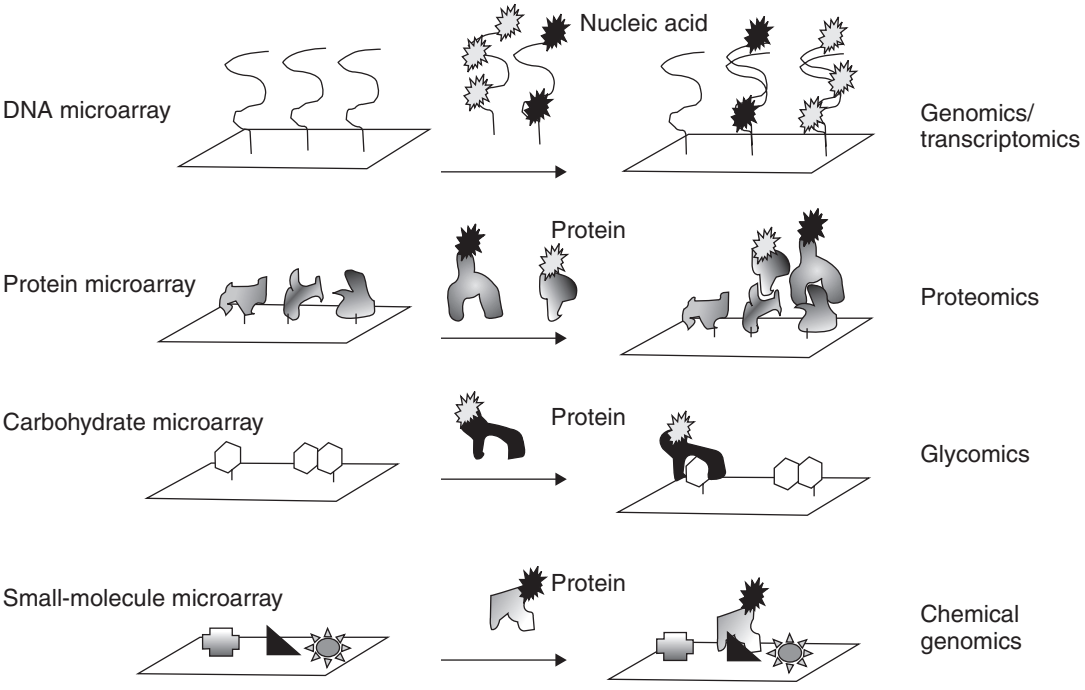



**Figure 1.2**

Microarray-based technologies used for the study of biological interactions. Modified from Shin *et al.* (2005) *Chemistry, a European Journal* **11**: 2894–2901, with permission from Wiley-VCH.

- comparative genome hybridization (CGH),
- genotyping,
- whole genome resequencing,
- determining protein DNA interactions (ChIP-chip) (Stears *et al.*, 2003; Buck and Lieb, 2004; Stoughton, 2005) (see *Figure 1.3* for an overview of some of the applications of DNA microarrays).

Several other DNA array-based applications are developing, which include:

- regulatory RNA studies (Wassarman *et al.*, 2001; Zhang *et al.*, 2003),
- alternative splicing and RNA binding protein studies,
- methylome analysis (reviewed in Mockler and Ecker, 2005).

DNA microarrays are probably best known as a research tool to study whole organism/tissue genome-wide transcriptional profiling, where the transcription profiles of all of the genes on a genome, from a subset of genes from the genome, or from a particular tissue, can be simultaneously assayed. The nucleic acid that is being assayed in transcriptional profiling by the arrays is messenger RNA, which is generally converted to cDNA

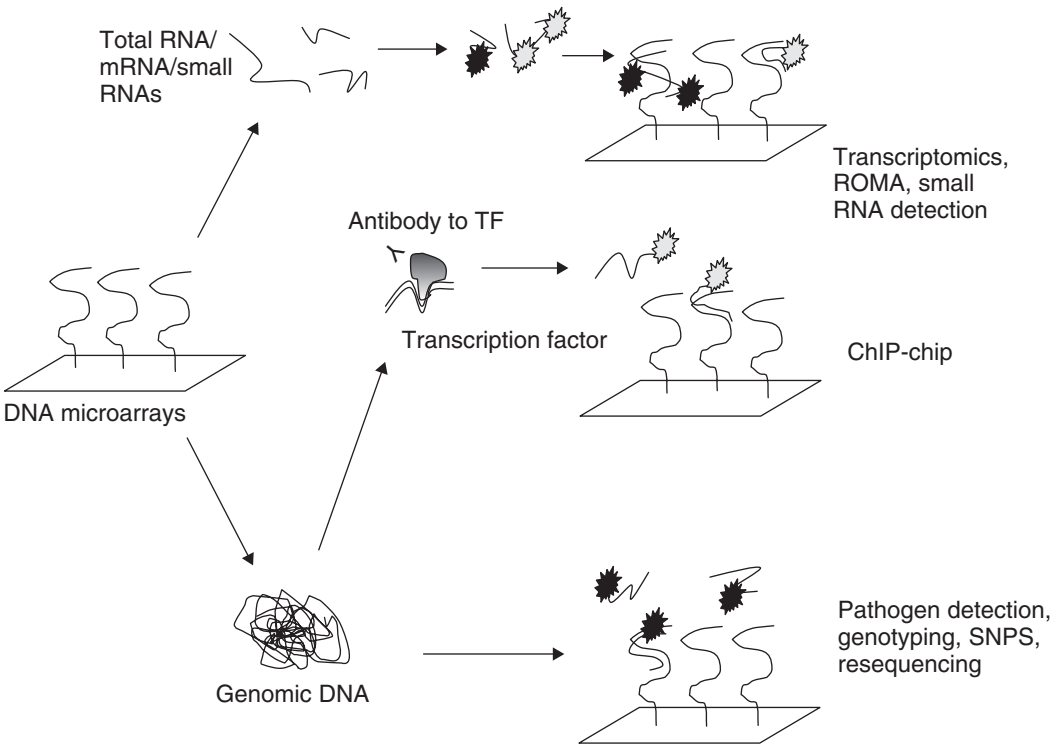



**Figure 1.3**

Examples of applications for DNA microarrays.

before hybridization onto the array. It is common in these types of experiments to compare the transcription profiles of two or more individuals or cultures of cells, in order to determine the transcriptional differences between them. In their simplest form (sometimes referred to as type I experiments (DeRisi *et al.*, 1997; Yang and Speed, 2002) DNA array transcriptomics experiments compare, for example, a wild-type with a mutant, a healthy with a diseased individual, or unstressed with stressed cells. More complicated transcriptomics experiments, which use an invariant control, such as genomic DNA, or pooled RNA for one channel in the experiment (sometimes referred to as type II experiments (DeRisi *et al.*, 1997; Yang and Speed, 2002)), can be used to compare multiple individuals, conditions or treatments. The number of published *in vivo* transcriptomics papers in which microarrays have been used is growing at an extremely rapid rate, for both prokaryotic and eukaryotic systems (Stoughton, 2005), with thousands of papers now published using this technology. One variation of *in vivo* transcriptional profiling is the *in vitro* transcriptional profiling method: run-off transcription microarray analysis (ROMA), (Cao *et al.*, 2002; Zheng *et al.*, 2004, and this book), which uses microarray technology to profile the abundance of run-off transcripts generated *in vitro* using DNA template, purified RNA polymerase, a regulatory protein and nucleotides.

There are a number of array-based methods that use DNA arrays to interrogate the DNA content of cells, tissues or other samples. These methods rely on the high density of probes contained on a DNA array to multiplex hybridizations, in order to either assay for a large number of DNA sequences (detection of the presence or absence of DNA sequences), or to use the high density of probes to increase resolution in locating the presence of a hybridizing piece of DNA (mapping onto the chromosome). Comparative genome hybridization (CGH) is a method that has been extensively used to detect the absence or presence of particular genes or chromosomes, or variations in gene copy number in eukaryotes, as these gene deletions or duplications are often associated with diseases such as cancer, and with developmental abnormalities, such as Down's syndrome. CGH uses co-hybridization onto metaphase chromosomes of labeled total genomic DNA from a 'test' and 'reference' population of cells to localize and quantitatively measure DNA copy number differences between these populations, and associate the copy number aberrations with the disease phenotype (Kallioniemi *et al.*, 1992). One disadvantage of CGH is that the use of metaphase chromosomes results in a relatively low resolution of detection using CGH (Pinkel *et al.*, 1998). Array CGH is a further refinement of the technique that allows high resolution of detection of where deletions or gene duplications occur on the chromosome, by using high density or tiling DNA arrays (Pinkel *et al.*, 1998; Albertson *et al.*, 2000; Dunham *et al.*, 2002; Ishkanian *et al.*, 2004). Similarly, the use of microarrays in genotyping and detection of single nucleotide polymorphisms (SNPs) is becoming widely used, because arrays offer rapid, parallel allele discrimination (Fan *et al.*, 2000; Hirschhorn *et al.*, 2000; Syvänen, 2001; Lindroos *et al.*, 2002; Kennedy *et al.*, 2003; Matsuzaki *et al.*, 2004). A further development of DNA rather than RNA-based uses for arrays is in the detection and characterization of pathogenic microorganisms. Use of DNA arrays for detection is a technology that has become popular because of the ability of DNA arrays to be both flexible and multiplexing (Call *et al.*, 2003; Korczak *et al.*, 2005). DNA microarrays have also been used to compare pathogenic and non-pathogenic variants of related bacterial species using comparative genomic hybridization (Behr *et al.*, 1999; Salma *et al.*, 2000; Schoolnik, 2002; Call, 2005), and for resequencing of pathogen strains, detailed below.

A fundamental goal of understanding how organisms regulate gene expression is the study of how regulatory proteins interact with DNA, and influence transcription from promoters within their cognate regulon. The site specificity of regulator interaction with genomic DNA, and what influences their binding to genomic DNA *in vivo* is fundamental to studies on the control of gene expression. This can be problematic when mutations or a deletion of the gene encoding the regulator, or overexpression of regulatory proteins is lethal to the host cells (Lieb *et al.*, 2001). Many of these studies on transcription factor interactions with DNA have been conducted *in vitro*, using purified proteins and DNA in gel shift assays, DNAase I footprinting assays or SELEX (systematic evolution of ligands by exponential enrichment (Gold *et al.*, 1997)) and other methods. These methods have been limited by the need to overproduce and purify regulatory proteins for assays, and there are indications that *in vitro* binding of a protein to DNA is not always an accurate predictor of a regulator's binding sites *in vivo* (Lieb *et*