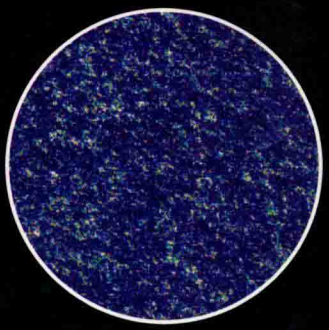
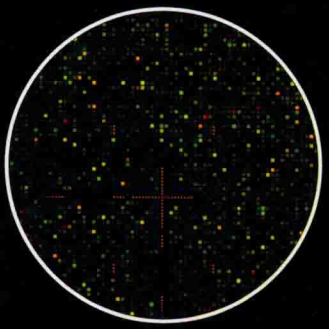
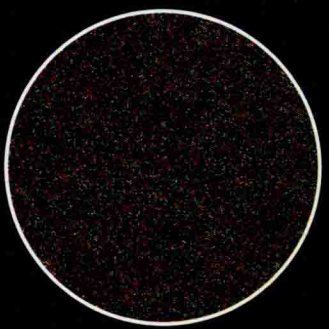
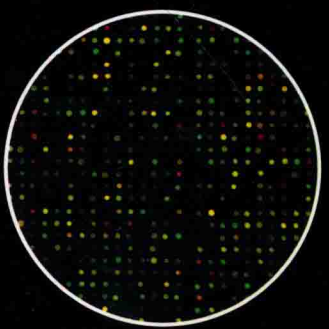
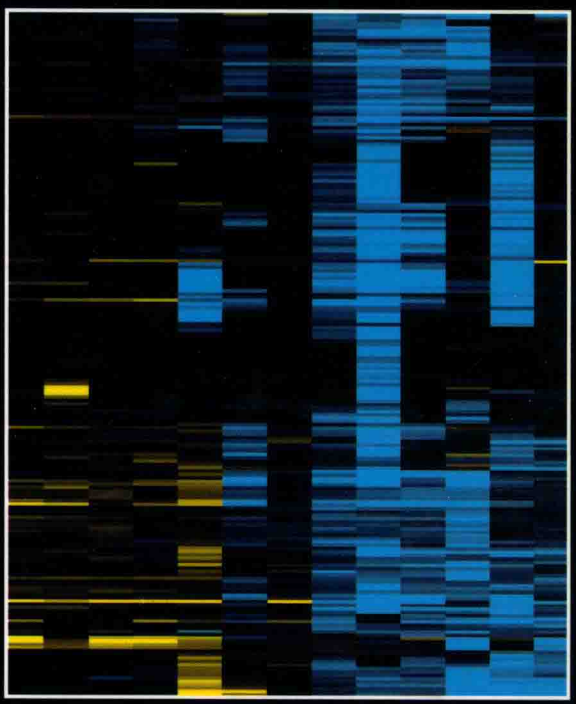








Microarray Technology in Practice

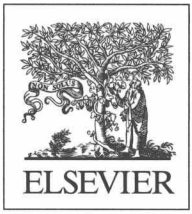


Steven Russell • Lisa A. Meadows
Roslin R. Russell

AP

Microarray Technology in Practice

Steven Russell, Lisa A. Meadows and Roslin R. Russell
Department of Genetics and Cambridge Systems Biology Centre
University of Cambridge
Cambridge, UK



AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an Imprint of Elsevier

Academic Press is an imprint of Elsevier
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
32, Jamestown Road, London NW1 7BY, UK

First edition 2009

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress
ISBN–13: 978-0-12-372516-5

Printed and bound in USA
09 10 11 12 13 10 9 8 7 6 5 4 3 2 1

For information on all Academic Press publications
visit our website at elsevierdirect.com

Microarray Technology in Practice

Foreword and acknowledgments

We have written this book because we believe that a single volume over viewing the current state-of-the-art in microarray technology is currently lacking. While there are many excellent books and reviews covering microarrays, most of the books were written a few years ago and do not cover the more recent developments in very high-density arrays for gene expression and genomic DNA analysis. As we hope to demonstrate, there has been tremendous progress in the widely available platform technologies so that whole-genome analysis is now available for virtually any organism with a sequenced genome at a cost affordable by many research laboratories. Hand-in-hand with technology developments, the analytical tools for data processing and analysis have also progressed and we hope that this book provides a comprehensive overview of the current approaches that is accessible to the beginner. To accompany the book we have generated a website that contains supplementary information, additional details on some of the analysis methods and a series of useful links. The site can be accessed at http://www.flychip.org.uk/mtip/

The preparation of this book has relied on advice and suggestions from several of our colleagues but we are particularly grateful to Bettina Fischer, Richard Auburn and Natasha Karp for their insightful comments on the manuscripts. Our other colleagues in the Cambridge FlyChip facility and more widely, provided an excellent environment for developing the technology and we thank Boris Adryan, Nuno Barbosa-Morais, David Kreil, Gos Micklem, Santiago Sevillano Matilla, Peter Sykacek, Natalie Thorne, Sarah Vowler and Rob White for intellectual and technical input. A special mention has to go to François Guillier whose unflappable systems administration has kept us honest. Steve couldn't have done any of this without the unfailing support of Michael Ashburner and John Roote: drinks are on me boys! Of course, we could not have considered approaching the compilation of this work without the support of our families, who have had to put up with our absence on many weekends and evenings so our heartfelt thanks to Fiona, Shuggie, Jonathan, James, Andrew, Struan and Isla. Lisa would particularly like to thank her mum, Ann, for all her help with childcare.

It is difficult to get to grips with the complexities of microarray analysis without getting your sleeves rolled up and doing experiments yourself. In the early days of the technology this was a relatively expensive business, requiring considerable infrastructure and Steve would like to particularly acknowledge the prescience of

the UK Biotechnology and Biological Sciences Research Council, who invested considerable funding to establish core genomics infrastructure in the UK, with a special mention to Alf Game for his energy and enthusiasm.

Steve Russell is a Reader in Genome Biology in the Department of Genetics, University of Cambridge and founder of the Cambridge Systems Biology Centre. After a PhD at the University of Glasgow with Kim Kaiser (a true technological visionary) where he was trying to do fly functional genomics before functional genomics was really invented, he came to Michael Ashburner's lab in Cambridge and has never left! He established the FlyChip Drosophila microarray facility in 1999 and has been involved in fly functional genomics and microarray analysis ever since. He helped found the International Drosophila Array Consortium to make core array resources widely available to the worldwide research community.

Lisa Meadows is currently a Research Associate in the FlyChip microarray group, primarily focusing on the technical and 'wet lab' aspects of microarray technology. She established and led implementation of the FlyChip laboratory protocols including array printing, RNA extraction, amplification, labeling, hybridization, scanning and image analysis. Prior to this she obtained a strong grounding in molecular biology techniques and Drosophila research with a PhD from the University of Cambridge and postdoctoral research at the University of Mainz, Germany.

Roslin Russell is currently a Senior Computational Biologist at the Cambridge Research Institute, Cancer Research, UK, specializing in the experimental design and analysis of expression, ChIP-chip, CGH and SNP array data from various commercial platforms. She completed her PhD at the University of Cambridge, where she developed her own spotted microarrays for the study of the human major histocompatibilty complex (MHC) region in health and disease and gained experience in comparative microarray platform analysis. Following this, she participated in the establishment of a microarray facility for the Leukaemia Research Fund at the Genome Campus in Hinxton, Cambridge and then joined FlyChip where she led the development of a microarray bioinformatics and analysis pipeline using R and various packages in Bioconductor.

Contents

Chapter 1

Introduction

Although there may have been a considerable amount of hype surrounding the emergence of DNA microarray technology at the end of the 20th century, the dawn of the 21st century has seen the realization of much of the initial promise envisaged for this technique. We now have a set of stable platform technologies that allow virtually any nucleic acid assay at a genome-wide scale for any organism with a sequenced genome, with the past 5 or so years seeing both the stabilization of the available platforms and the convergence in the data derived from different array technologies. In this book we gently introduce the reader to all aspects of modern microarray technology from designing the probe sequences on the array through to submitting the data to a public database. On route we hope to provide a comprehensive overview of the current state-of-the-art in the field and help guide the novice through the myriad of choices available for conducting experiments and analyzing the subsequent data. Although the primary focus is on the use of microarray technology for gene expression profiling, we describe platform technologies for other nucleic acid analysis, principally genome exploration and high-throughput genetics, as well as introducing more recent technologies for arraying proteins. We hope there is material in this volume that is useful for both those new to the field and those with more experience.

1.1 TECHNOLOGY

In his book '*The Sun, The Genome, and The Internet: Tools of Scientific Revolution*', Physicist Freeman Dyson argues that it is principally the development of new technologies that drive scientific progress rather than revolutionary new concepts (Dyson, 2001). Nowhere is this truer than in modern biology, where advances in molecular biology have opened up incredible vistas on the living world by providing a set of tools that have allowed us to begin deciphering the genetic code underpinning the molecular basis of life. Technologies such as

Microarray Technology in Practice

DNA cloning, nucleic acid hybridization, DNA sequencing and polymerase chain reaction, facilitate the isolation and analysis of virtually any nucleic acid molecule and have culminated in the determination of the complete human genome sequence along with the sequences of over 4000 other species. The emergence of DNA microarray technology promises much more in this vein of discovery with the ability to interrogate entire genomes or gene complements in a comprehensive high-throughput way.

While we can now determine the DNA sequence of any organism with relative ease, interpreting that sequence is an entirely different matter. From the sequence we may be able to assemble a reasonable description of the repertoire of protein coding genes encoded in a genome, however, understanding how those genes are deployed to generate the incredible diversity of cell types and organisms that characterize the living world remains a daunting challenge. Trying to decipher this complexity is one of the goals of functional genomics, a branch of molecular biology that may be defined as the study of gene functions and inter-relationships. If we take as an example the relatively simple eukaryote, the bakers yeast *Saccharomyces cerevisiae*, over 6000 genes, including the 100 or so that code for transcriptional regulators, need to be coordinated to allow the cell to survive and replicate in a variety of different environments. It is fair to say that despite its relative simplicity, we are only just beginning to scratch the surface in terms of understanding the gene regulatory systems that govern the behavior of this organism. When trying to decipher the complexity of living systems it must be recognized that defining gene functions and their interactions is only part of the story. As Denis Noble cogently argues in '*The Music of Life: Biology Beyond the Genome*' (Noble, 2006), it is the property of a biological system as a whole, the multilayered interactions between genes, their products and the cellular environment that generates biological function. Thus the rapidly developing science of integrative systems biology aims to understand biological processes in terms of such systems and in doing so provides robust mathematical descriptions of the networks and interactions underpinning the systems. Of course in order to completely describe a system one must have a reasonably complete inventory of the components of the system and it is therefore desirable to be able to accurately define the set of genes expressed in a particular system and how the expression of these genes changes over time or in response to a particular stimulus. Prior to the advent of microarray technology, comprehensive analysis of gene expression was not really possible since classical methods can only really accommodate the analysis of a few tens of genes rather than the tens of thousands encoded by a metazoan genome. In their current state, microarrays facilitate just such an analysis. We would argue that the combination of bottom up functional genomics and top down systems approaches will be necessary if we are to truly understand how the genetic information locked in the genome sequence is translated into the incredible complexity that is a cell and how collections of cells cooperate to elaborate multicellular organisms.

As Eric Davidson beautifully illustrates in *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution* (Davidson, 2006): 'All morphological features of adult bilaterian body plans are created by means of pattern formation processes. The *cis*-regulatory systems that mediate regional specification are thereby the keys to understanding how the genome encodes development of the body plan. These systems also provide the most fundamental and powerful approach to understanding the evolution of bilaterian forms.' If we accept this line of reasoning, we need to understand how the information in the genome is dynamically deployed during development. In the first place, we must discern where and when genes are turned on and off: defining the spatio-temporal profile of the gene complement. Perhaps more importantly, we need to understand how the *cis*-regulatory systems regulating gene expression are encoded in the genome sequence and how regulatory molecules interpret this information to accurately control gene expression. The comparatively recent development of microarray-based methods for genome-wide mapping of in vivo transcription factor binding sites will undoubtedly contribute to such efforts and significantly advance our exploration of this complexity.

Along with our fascination with basic biology, the biosciences seek to uncover fundamental aspects of human biology and nowhere is this more evident than in our desire to understand the molecular basis underlying human disease. While some diseases are inherited in a straightforward Mendelian fashion, the majority of human conditions are much more complex traits influenced by many different genes and external environmental factors. Getting to grips with such complexity is very much akin to the challenges we face in understanding developmental processes: what genes are involved and how does the genome of particular cell types respond to changes in internal and external stimuli. Unfortunately, unlike the model organisms used for studying developmental problems, humans are not good experimental systems and we must therefore rely on genomic differences between individuals to provide us with clues about the genes involved in particular diseases. In this area we see an explosion in data accumulation over the past two or three years with the advent of microarray-based high-density genotyping platforms facilitating the identification of genetic differences in large-scale case-control studies. While this technology is in its infancy, the studies already performed promise considerable insights into complex multigenic traits. Although there is certainly a possibility that the new generation of ultra high throughput sequencing may overtake microarray based assays by allowing relatively cheap whole human genome resequencing, the possibility for using microarray-based assays as diagnostic or prognostic tools in the clinic is very attractive in terms of speed and cost. Of course not all-human diseases have a genetic basis and infections by pathogenic organisms remain a major contributor to human mortality, especially in the developing world. In this case, as well as understanding the biology of the infectious agent and the affected human tissues we need to get to grips with the interactions between

the host and the pathogen. This is a fascinating area of biology that is beginning to yield to functional genomics approaches with the prospect that weaknesses in the parasite can be harnessed for the development of effective therapeutic agents. This is a pressing need in the developing world: for example, it is estimated by the World Health Organization that there is a new person infected with tuberculosis every second with approximately one third of the world's population currently infected with the TB bacillus. Similarly, more than one million people die each year of malaria with the majority of disease occurring in sub-Saharan Africa: two children every minute are killed by this preventable infection. We desperately need to develop cheap and effective therapies that can combat such diseases and understanding the complex lifecycle of the responsible pathogens is certainly one step along the way.

It is, in our view, obvious that microarray technologies have much to offer in the exploration of biological complexity. While 'the secrets of life' are unlikely to be completely uncovered in our lifetime, the application of modern genome exploration methods are sure to shed some light on the complexity of life. We should bear in mind that it is only a little over 10 years since the publication of the first widely accessible microarray technology and in this time we have witnessed an incredible acceleration in our ability to explore gene expression and other areas of genome biology. A demonstration of this can be seen from some yearly PubMed searches using simple terms such as *microarray* or *genomics* (Figure 1.1), with well over 5000 papers a year now published containing these terms. It is clear that the technology is making a major impact on biology and we suspect that soon the use of DNA microarrays for expression profiling will be a universally deployed integral part of any molecular biology-based investigation.

1.2 A BRIEF HISTORY

There are many reviews and books that cover the brief history of microarray technology and it is not our intention here to extensively go over this ground. However, a few words about the basics are worth revisiting to set the scene before we overview the contents of the book. Nucleic acid microarrays are founded on the exquisite specificity inherent in the structure of the DNA duplex molecule (Watson and Crick, 1953) since complimentary single strands are able to recognize each other and hybridize to form a very stable duplex. It was Ed Southern (Southern, 1975) who first realized that this specificity could be used to detect specific sequences in a complex mixture by labeling a known DNA fragment (the probe) and using this to interrogate a fractionated sample of, for example, genomic DNA. The Southern blot technique was soon adapted so that specific RNA molecules in a fractionated cellular extract could be similarly detected using Northern blots (Alwine et al., 1977) and the routine analysis of mRNA transcripts was established. It was realized that the concept of using a

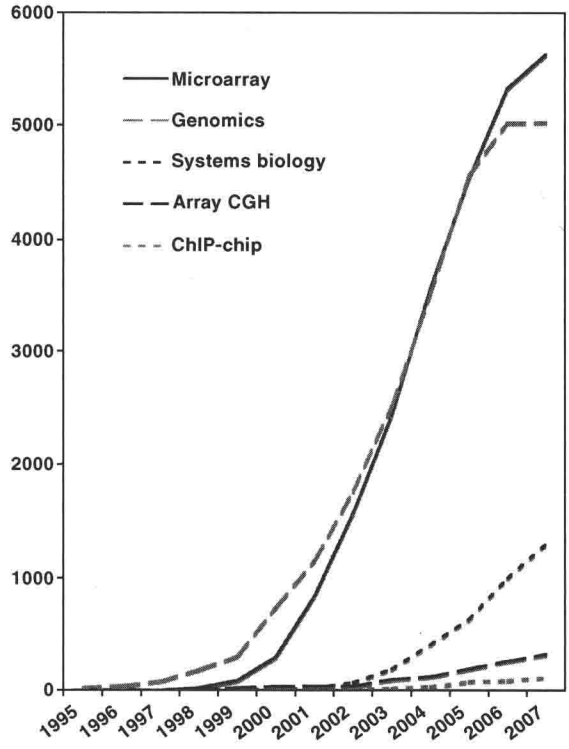


FIGURE 1.1 Impact of microarray technology in PubMed citations. Number of publications (*y*-axis) per year retrieved from PubMed using search terms to capture the indicated areas.

labeled probe fragment to identify complimentary sequences was adaptable to parallel processing of DNA clones with Grunstein and Hogness (1975) providing the first demonstration of using hybridization to isolate specific clones from plasmid libraries. Together, these methods and their subsequent developments provide the foundation for virtually all aspects of current molecular genetics and the conceptual basis for DNA microarray technology.

In its current forms, microarray technology derives from two complimentary approaches developed in the 1990s. The first 'cDNA microarrays' were produced in Patrick Brown's laboratory in Stanford (Schena et al., 1995), utilizing gridding robots to 'print' DNA from purified cDNA clones on glass microscope slides. The slides were interrogated with fluorescently labeled RNA samples and the specific hybridization between a cDNA clone on the slide and the labelled RNA in the sample used to infer the expression level of the gene corresponding to each cDNA clone. Thus was born the spotted array, a technology that can be implemented in individual research labs and has provided an accessible route for high throughput gene expression profiling in many areas of biology. Since the

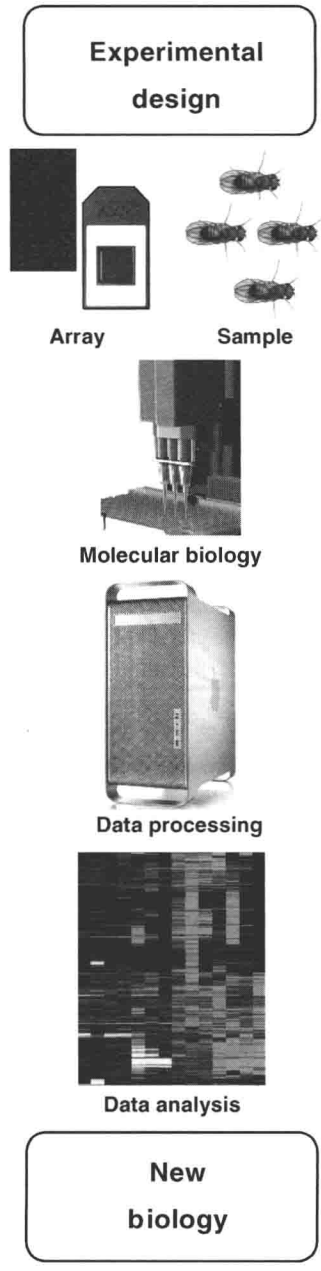


FIGURE 1.2 A microarray study. An overview of the stages involved in a typical microarray gene expression study. Experimental design is of paramount importance and is derived from a consideration of the biological question, the available samples and the type of microarray platform used. Molecular biology techniques are used to isolate and label samples, which are hybridized to the array

DNA fragments printed on the array correspond to the specific labelled molecule used to interrogate Southern or Northern blots, the clones on the array are referred to as probes and the labelled complex RNA mixture as the target. In parallel work at Affymetrix, in situ synthesis of defined oligonucleotides probes at very high density on glass substrates (Fodor et al., 1991) was shown to provide a reliable route for measuring gene expression (Lockhart et al., 1996) and set the scene for the development of the current generation of ultra high density microarrays now employed for gene expression, genome tiling and genotyping.

In essence, a microarray study consists of a series of defined stages as outlined in Figure 1.2. An experimental question is formulated and a design phase combines the desired biological objectives with the type of array platform selected and the available biological samples to generate a robust experimental design. Some relatively straightforward molecular biology is used to extract and label the RNA samples, hybridize to the array and acquire the primary data. These data are processed and a variety of statistical and analytical tools used to identify changes in gene expression thus leading to new biological insights. In terms of the mechanics of a microarray study, there are two fundamental approaches that initially arose from the type of microarray platform used (Figure 1.3). In the case of the cDNA or spotted array, the analysis generally involves a comparative hybridization, where two RNA samples are separately labelled with different (generally fluorescent) reporters: both labelled samples are then combined and hybridized together on the same array. In this way these so-called dual channel arrays allow for a comparative analysis between samples (wild type and mutant for example). While the data for each sample are collected separately during the acquisition stage, they are generally combined during the analysis stage to yield a ratio of gene expression between one sample and another. Such a comparative approach allows for very sensitive detection of differences in gene expression between two samples. The alternative, or single channel approach, was developed in concert with the Affymetrix GeneChip platform though it is not unique to this technology. In this case, RNA from each biological sample is individually hybridized to an array and comparisons are made at the subsequent data analysis stage. Obviously single channel experiments generally require twice as many arrays as dual channel experiments though this is dependent upon the experimental design. A particular benefit of single channel experiments is that they can provide far more flexibility when comparing a variety of different samples hybridized to the same platform: it should be obvious that each channel of a dual channel experiment can be independently treated as a single channel, though caution must be exercised when doing this.

---

and the data acquired via a dedicated scanner. A variety of software tools are used to process and normalize the array data prior to statistical and meta-analysis to identify and classify differentially expressed genes. The end result is hopefully new insights into the biology under investigation.