MANUAL
of
AQUATIC VIRAL ECOLOGY

MAVE Chapter 5, 2009, 34–56

Methods in aquatic virus ecology: Construction of microarrays and their application to virus analysis

Michael J. Allen[1], Bela Tiwari[2], Matthias E. Futschik[3], and Debbie Lindell[4]*

[1]Plymouth Marine Laboratory, Prospect Place, Plymouth, UK

[2]NERC Environmental Bioinformatics Centre, NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, UK

[3]IBB—Institute for Biotechnology and Bioengineering, Centre for Molecular and Structural Biomedicine, University of Algarve, Faro, Portugal

[4]Faculty of Biology, Technion—Israel Institute of Technology, Haifa, Israel

Abstract

DNA microarray is the term used to describe a microscopic collection of DNA probes arrayed onto a solid surface. Microarrays take advantage of the highly selective nature of nucleic acid interactions and are commonly used for expression profiling, for comparative genomic hybridization, to aid genomic annotation, and for detection of mutations within genomes. In this virus-focused chapter, we deal primarily with the use of microarrays for expression analysis (the most popular usage) of host and virus systems during infection. We examine aspects related to array platform choice (spotted and oligonucleotide arrays), probe and array design considerations, experimental procedures and data analysis, normalization, processing, and curation. We also provide in-depth examples for the study of viral transcriptome analysis for both spotted long oligonucleotide (coccolithoviruses) and Affymetrix GeneChip (cyanophage) arrays.

Introduction

A DNA microarray is a microscopic collection of DNA probes (or features) arrayed onto a solid surface. Microarrays take advantage of the highly selective nature of nucleic acid interactions (either DNA–DNA or DNA–RNA) and are most commonly used for expression profiling and comparative genomic hybridization. Although the technology is cutting edge, it is based on the basic principles developed in the Southern (DNA) and northern (RNA) blots (Southern 1975, Alwine et al. 1977). These techniques are still widely used but are usually limited to the study of a handful of target sequences/genes. Microarrays, on the other hand, provide the opportunity to screen tens to hundreds of thousands of targets simultaneously. The high cost involved in obtaining genomic information and constructing microarrays originally limited their use to model organisms such *E. coli*, mouse, and *Drosophila*. Technological developments, however, have created an explosion in the genomic information available and have significantly reduced the costs associated with developing arrays for other organisms. This is particularly relevant to the field of virology, where the genomes studied are smaller (and hence more accessible) and the associated arrays can be developed on a limited budget. To put this into context, there are approximately 57 million bases of sequence information from complete virus genomes in the current GenBank database (as of 13 December 2007). This equates to the amount of sequence generated from a single sequence run on the newly developed pyrosequencing systems.

Microarrays have been a revelation in the field of molecular biology. The simultaneous analysis of thousands of genes

*Corresponding author: E-mail: *mija@pml.ac.uk

Acknowledgments

M. J. Allen is supported by grants from Natural Environment research Council (NERC) through the Environmental Genomics program (NE/A509332/1 and NE/D001455/1) and Oceans 2025. B. Tiwari is supported by the NERC Environmental Bioinformatics Centre. M. E. Futschik is supported by the Ciência 2007 initiative of the Fundação para a Ciência e a Tecnologia, Portugal. D. Lindell is supported by the Morasha Program of the Israel Science Foundation (grant 1504/06) and a Marie Curie International ReIntegration Grant (MARCYV) within the sixth European Community Framework Programme and is a Shillman Fellow. There are no declared conflicts of interest.

Publication costs for the Manual of Aquatic Virus Ecology were provided by the Gordon and Betty Moore Foundation. This document is based on work partially supported by the U.S. National Science Foundation to the Scientific Committee for Oceanographic Research under Grant OCE-0608600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation (NSFG).

ISBN 0000000000000, DOI 00.0000/00000000000

has provided a wealth of information that we are only now coming to terms with. Microarrays have allowed the fine-tuned mapping of transcriptional pathways and cascades from a variety of organisms responding to a plethora of environmental stimuli such as physical, chemical, and biological stressors. Yet microarrays should be thought of as a technique to be used in combination with others. Typically, a large group of genes is analyzed to identify a small group of genes involved in a particular process, response, or even phenotype. Of course, assaying huge numbers of genes simultaneously will never be as accurate as assaying each gene individually, but microarrays offer a relatively cheap and quick way to assay genes on a genomewide scale, allowing future focus on the genes identified; i.e., microarrays can be thought of as a molecular compass to guide research in the right direction. Thus, when performing genomewide transcriptional profiling, it is essential that techniques such as quantitative real-time PCR and northern blotting are used to confirm the findings from a subset of genes from microarray experiments. In the case of comparative genomic hybridization, an array can pinpoint where variation occurs in a genome but cannot determine what the basis of the variation is. In this case, it is important to verify the results through direct sequencing.

Despite the insight that microarrays can help provide, the marine and aquatic sciences have been relatively slow to embrace this technology. In the past, marine-focused microarray work has been limited to a handful of isolated laboratories spread around the globe. We hope this review will promote the use of microarrays to answer important aquatic science questions, in particular in the field of aquatic virology. The potential of microarrays for use in studying virus biology is enormous (Allen and Wilson 2008). Relatively small virus genomes can have profound and dramatic effects on host global gene expression, making microarrays a powerful tool in a molecular biologist's armory.

In this chapter, we deal primarily with the use of microarrays for expression analysis of host and virus during infection. We give in-depth description of the most commonly used platforms and methodology and discuss issues that must be considered when designing and using arrays for aquatic virus studies. We do not attempt to give a comprehensive overview of all available technologies and options, but invite the reader to investigate other options using the issues discussed herein as a guide (Karsten and Geshwind 2002, Liu 2007, Page et al. 2007, Sipe and Saha 2007, Bier et al. 2008, Coppee 2008, Gresham et al. 2008, Simon 2008, Dufva 2009a, 2009b). Microarrays are ideal for determining the transcriptional program of viruses during infection (as well as the host's transcriptional response to this infection) because of their ability to receive information on the transcription of all genes from both genomes from the same sample. This requires that probes for both host and viral genomes are designed on the one array. In these experiments, where temporal dynamics of transcription during infection are being investigated, it is important to ensure synchronous infection. The time scale for such experiments is generally the length of the latent period. When investigating host responses, it is also important to ensure that the vast majority of cells are infected so that transcriptional changes in the infected cells are not masked by expression levels in uninfected cells. The contamination status of cultures is also worth considering; it is best to work with axenic cultures if possible. With environmentally relevant host–virus systems, actual host infection levels at a particular multiplicity of infection (MOI) do not always match those based on theoretical considerations from Poisson distribution. Therefore, a number of preliminary experiments are essential before embarking on a microarray experiment beyond determining infection parameters (such as the length of the lytic cycle and latent period). These experiments include determining the conditions for maximizing the percent of infected cells and the actual degree of infection, as well as conditions for synchronous infection by assessing various cell and virus concentrations in addition to the appropriate MOI. It is also very useful to determine the timing of different stages of the infection cycle to which transcriptional data can be compared. For example, viral genome replication, production of structural proteins, host genome degradation, and changes in host cell morphology can all help place the transcriptional information in the wider context of the infection process.

Materials and procedures

Microarray platforms—At every stage of microarray design and construction, there are various options that can be taken depending on the available budget, the number of probes on the array, and the local infrastructure and facilities available (Fig. 1). It is crucial that before microarray design begins, the scientific questions of interest are considered and defined, so that a microarray fit for the purpose is constructed. For example: Can the virus system be accurately studied on a microarray independent of the host system? Are there any genes common to both virus and host genomes? How much virus message will be present in relation to host message at various stages of infection? Will any amplification of message be needed? The answers to these sorts of questions can have a profound impact on the nature of the microarray developed. Other questions that may affect your microarray design include the following: How many microarray experiments will be run? Will the microarray be used by just one research group, or will it be made available to other interested parties? What local microarray infrastructure is available?

Here we focus primarily on two array platforms: spotted arrays and Affymetrix GeneChip arrays. We also occasionally comment on Agilent arrays when relevant. Other array types exist, but these three are common platforms that provide a good overview of the different approaches currently employed. Each array platform offers its own particular advantages and disadvantages, which we discuss. We start off describing high-density custom oligonucleotide systems and

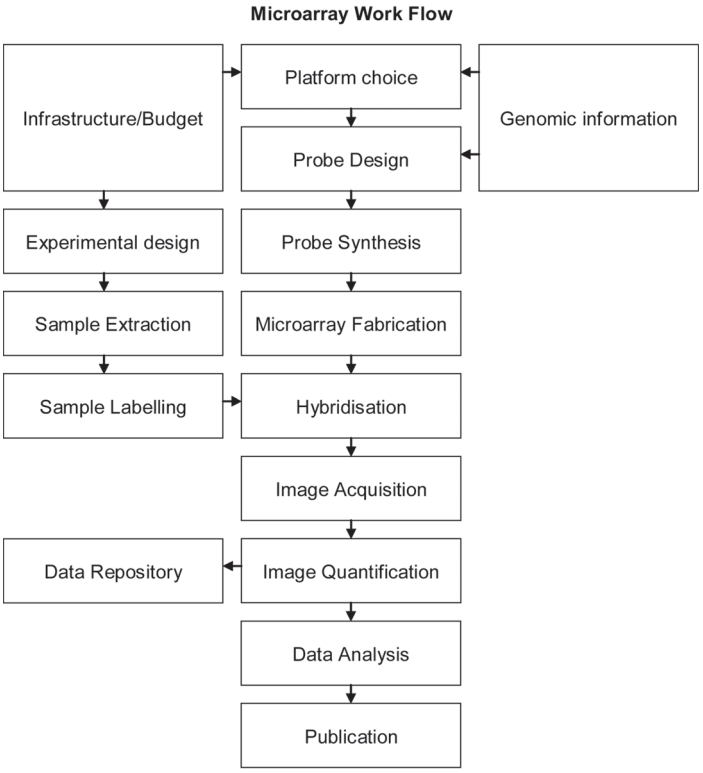

Fig. 1. Microarray work flow.

then move on to the custom spotted microarrays. One issue we will not touch on in this review is specific costs, which are highly subjective and prone to changes in the future; here, we discuss only the current costs of each system relative to other options presently available. With respect to cost calculations, the stripping and reuse of previously used microarrays is technically feasible for many types of microarray system, but we strongly advise against this practice as it can introduce uncontrollable variability.

High-density oligonucleotide microarrays—High-density oligonucleotide microarray systems (such as the Affymetrix and Agilent systems) usually offer the best reliability, reproducibility, and coverage. Many high-density microarrays are constructed using a process known as photolithography (Affymetrix and Nimblegen), whereby light is used to stimulate in situ DNA synthesis in defined positions; others use inkjet technology (such as Agilent) (Table 1). In both these cases, single-stranded oligonucleotides are sequentially synthesized base by base, directly on the solid surface of the array. Whereas Affymetrix technology uses a number of short 25-mer probes per gene (generally 8–11), Agilent arrays consist of a single 60-mer probe for each gene. The manufacturing process is an extremely robust procedure with no noticeable differences between arrays. A powerful application of high-density microarrays is to produce what is referred to as a tiling array. Tiling arrays are designed independently of annotation data, cover entire stretches of genomic sequence (usually total genomes) in an unbiased manner (e.g., 25-mer probes designed with a space of approximately 50 bases in between, along the length of the whole genome irrespective of annotation), and allow the identification of novel transcribed sequences (often unannotated) as well as regulatory elements. By determining which probes generate positive signals and their intensity relative to neighboring probes, regions of the genome that are transcribed (i.e., the genes) or are regulatory can be easily identified. Tiling arrays, developed by companies including Affymetrix, Nimblegen, and Agilent, are incredibly powerful but are commonly restricted to model organisms for which there is a large commercial market.

Affymetrix expression arrays can produce highly reproducible data. These arrays are designed and manufactured by Affymetrix based on empirical but proprietary information and therefore are the easiest for the researcher to "design," although also the most costly. After RNA is extracted, it can be labeled by the researcher or, for a cost, at an Affymetrix array facility. Hybridization of the arrays is carried out at a specialized array facility, generally by facility personnel. This makes this procedure relatively simple, especially for researchers not so familiar with RNA work, but requires that the researcher find a reliable facility. The greatest disadvantage of these arrays is the high cost incurred for the custom design necessary for nonmodel systems. Cost per array is also quite high (with a minimum order of 90 arrays).

Agilent arrays—both probes and array layout—can be designed for free (using their Web program at earray.chem.agilent.com/earray) or can be designed by Agilent for a reasonable fee. A major advantage of the Agilent platform is the flexibility for probe and array design it provides, being feasible to order any number of arrays at a time (even a single array) and redesign probes for subsequent arrays. RNA labeling can be carried out by the researcher or at an array facility; however, hybridization and scanning are generally carried out at an array facility, again making the process quite simple for the researcher. The biggest disadvantage of the Agilent arrays is the cost per array, which is considerably higher than for the other platforms. It is possible, however, to hybridize multiple samples in different compartments on one slide if the number of probes is small enough (for example, if investigating the expression dynamics of the viral genome alone), making the price more reasonable per sample.

Thus, for high-density oligonucleotide arrays, Affymetrix is the platform of choice for systems when a large number of experiments will be carried out, although the cost for array design is quite high. Agilent is the platform of choice when high design flexibility is desired and few samples will be investigated, although the cost per array is quite high.

Custom spotted microarrays—The costs associated with developing high-density microarrays can make them financially prohibitive. The development of a custom spotted array is an economical alternative, although there is a price to pay in reduced array reproducibility. A spotted microarray is also the platform of choice for large-volume experiments when

Table 1. Commercial suppliers of high-density microarrays.

Company	Web site	Description
Affymetrix	http://www.affymetrix.com	High-density chips can be designed with up to 1.3 million features, with 25mer probes. Features are 11 μm in size.
Agilent	www.chem.agilent.com	High-density arrays with up to 243,504 features per array printed on standard glass slides. Features are 65 μm in size. Standard probe length is a 60mer, but any length between 25 and 60 bases possible.
CombiMatrix	www.combimatrix.com	Array chips featuring 12,000 35-40mer probes, feature size 44 μm.
Nimblegen	www.nimblegen.com	High-density arrays with up to 2.1 million features, 50–75mer probes printed on standard glass slides. Features are 13 μm in size.

high design flexibility is required. The researcher has the added advantage of having complete control over array design and total flexibility in its use. For spotted microarrays, the most common method is to use a glass slide as the basis for printing. Glass slide arrays offer researchers great control: they can generate their own labeled samples, perform their own hybridizations, and scan using their own equipment. Depending on the infrastructure, once microarrays have been fabricated, all that is needed is a microarray scanner and some basic laboratory equipment. Probes can be immobilized onto glass slides by a variety of techniques. Initially, this was done by physical contact between robotically controlled pins and the slide surface. It is now more common to use a technique known as piezoelectric printing, which is akin to ink-jet printing and provides greater control over spotting quality and quantity. As the print head moves across the array, electrical stimulation causes the DNA to be delivered onto the surface via tiny jets in a noncontact process. The process of "slide printing" is time consuming, technically demanding, and requires expensive robotic machinery. Often this is beyond the budget of most research laboratories; however, microarray printing facilities are now commonplace and offer a cheap and reliable method of fabricating microarrays.

Perhaps the greatest advantage of glass slide microarrays over their high-density counterparts is the flexibility they offer in the nature of the material printed on the slide. Many different types of material, including PCR products, plasmid libraries (cDNA/Expressed Sequence Tag [EST], plasmid, shotgun [randomly fragmented DNA]), or presynthesized oligonucleotides have been printed on glass slide microarrays depending on what was available and most cost-effective for a given project. We strongly recommend the use of long, presynthesized oligonucleotides when sequence information is available. Presynthesized oligonucleotides provide high specificity and allow design of probes with similar hybridization characteristics. A length between 50 and 70 bases generally provides a good balance between sensitivity, specificity, cost, and consideration of the decreased coupling efficiency during synthesis with increasing probe length. As the price of generating longer probes has decreased over recent years, we recommend probes of approximately 70 base pairs (although in theory they can be synthesized to any size required). Long oligonucleotide probes have the added advantages of not needing certain quality checks required for other material types, for example, amplifying and verifying the quality of PCR fragments, as well as avoiding the problems associated with hybridizing to probes of different length and different annealing temperatures, which arise when using PCR and plasmid probes.

For those who do not have access to preexisting sequence data, probes can be generated from plasmid libraries. Here, researchers can hone in on a small number of unknown probes using a microarray, and sequence only relevant plasmids of direct interest to them toward the end of their experimental work. This used to be a popular use of microarrays; however, the price of sequencing and of generating synthetic oligonucleotides have dropped considerably in recent years. Thus, for most purposes, it is preferable to generate, sequence, and design long oligonucleotides, rather than generate physical materials such as plasmid libraries or PCR products for spotting onto arrays. High-throughput sequencing methods have already been used to provide sequence for use in designing transcriptomic arrays (Vera et al. 2008), and this is likely to become a popular avenue in the future.

For long oligonucleotide arrays, researchers must choose how much of the process they want to undertake locally versus the cost in time and money. Designing long oligonucleotide probes is within the abilities of most researchers, thanks to the ready availability of probe design software, both commercial and free. However, one should not underestimate the time required to design an array layout and to quality check this array before any experimentation can begin. For many researchers, the services offered by a company, along with a quality guarantee for what they provide and the time scale within which they will provide it, may make the initially larger financial outlay worth it in the end.

When engaging a company, the normal rules stand: Ensure that you lay out clearly what you need and expect, and that you read the terms and conditions of their service in detail. Many companies and facilities have much experience working with model organisms and the types of chip designs one might desire for studying such organisms. The needs of the viral community can be somewhat different, however; for

example, the requirement to spot multiple, unrelated genomes (e.g., virus and host) on a single chip, where the characteristics of these genomes can be quite different. Will the company print probes of different lengths? If not, how do they plan to deal with the likelihood of different melting temperature profiles among genes in each organism? If you are doing a diversity study, or screening for new genotypes, will they help you design appropriate probes? How many probes do you anticipate spotting? What is the minimum number of arrays you can have made? (For example, study the implications on choice if you are only going to run a few experiments/hybridizations with this design.) Could/should the company do the hybridizations for you?

If you are designing your own long oligonucleotide spotted array, then you must take into account the physical characteristics of the probes to include on the array, what types of sequences to represent, how many probes per gene (or region) to include, whether replicate spots will be printed, and what controls to print on the array. Some of these issues are common to both high-density and spotted arrays, although many are issues specific to just spotted arrays. Each of these topics is covered briefly below, as well as a brief comment on available microarray probe design software. In general, unless you already have the facilities locally, we recommend that you find a microarray facility to work with and undertake a detailed conversation with them about the needs of your experimental system.

Array design considerations—Regardless of which array platform will be used, it is imperative that the researcher decides how the RNA will be labeled, as this determines whether the array should contain probes that are identical to or complementary to mRNA—termed a sense or antisense array, respectively, by Affymetrix. For example, a protocol that carries out reverse transcription to produce cDNA requires an antisense array, whereas some protocols that include an amplification step will produce DNA that is identical (sense) to the original RNA.

Probes bound to their targets should have approximately the same melting temperature (typically 50–60°C) across the array. Some oligonucleotide design software (see section below) will allow you to design probes with a range of lengths. This can be useful if more than one organism (e.g., host and virus) with different nucleotide characteristics (e.g., GC content) will be represented on the array. If you decide to design probes of different lengths, make sure that the company you are ordering the oligonucleotides from will manufacture them.

Probes need to be specific to the target of interest and sensitive enough to detect low levels of that target. Potential for secondary structure in the probe or the target can affect sensitivity; this is particularly relevant to the longer probes on spotted arrays. Some oligonucleotide design software include calculations for this potential. The specificity of a probe for its target will depend on how many mismatches there are between them, and also the location and arrangement of those mismatches along the probe sequence (Letowski et al. 2004). Gene-specific probes should have little or no sequence similarity to nontarget genes that may be present in the sample. Comprehensive studies on the effect of probe–target characteristics have provided tables of empirical results for essential design criteria for gene-specific and group-specific probes (He et al. 2005, Karaman et al. 2005). Note that some software use free energy in place of sequence identity as a measure of oligonucleotide specificity.

For bacterial arrays, where total RNA (i.e., mRNA, rRNA, and tRNA) is labeled, it is important to ensure that none of the probes are similar to ribosomal and transfer RNA sequences, as even very low labeling efficiency of these abundant RNAs is likely to mask mRNA levels. In addition, probes for bacteria and bacterial viruses should not be 3′ biased, as random hexamers, rather than polyT priming, will be used in the majority of protocols for making cDNA. Therefore, if a single probe is designed per gene it should be positioned toward the 5′ end of genes. If multiple probes are designed, as in Affymetrix arrays, we suggest these be spread across the gene, although they could also be designed toward the 5′ end of the gene.

If there is sufficient room on the array, designing probes in intergenic regions will enable identification of small unannotated genes that may have escaped annotation, as well as small noncoding RNAs that are often found in these regions (Steglich et al. 2008). The researcher should also consider whether probes for the detection of antisense RNAs will enhance the utility of the array. Perhaps rather than including probes that are antisense to all mRNAs, preliminary experiments for the detection of mRNAs, noncoding RNAs, and antisense RNAs could be carried out by Solexa- or 454-like sequencing, which would greatly inform on probe design (O. Wurtzel and R. Sorek, pers. comm.). Alternatively, as mentioned above, a tiling array could be designed across the genome on both strands.

High-density Affymetrix arrays can contain many probes on each array, making it worth considering including multiple genomes per array. However, if the genomes of two of these organisms will be investigated at the same time, as for host and viral genomes, probes with little cross-hybridization between the genomes must be designed (see "Probe design" below) and can be empirically tested with DNA from each organism. This is particularly important, as today we know that viral genomes often include host-like genes (Hughes and Friedman 2005, Moreira and Brochier-Armanet 2008, Monier et al. 2009). If this design is not feasible, and indeed whenever potential cross-hybridizing probes are being used, it is best to confirm the microarray results or carry out single-gene analysis from the outset, with a method capable of differentiating between host and viral copies of the genes—for example, using quantitative RT-PCR. Conversely, if the multiple genomes on the array will be investigated independently, i.e., separate arrays for each sample type, then there is no need to ensure that the probes do not cross-hybridize.

The types of probe sequences required on a microarray designed to study biodiversity will be different from those used on an array to study differential expression. Combinations of probes might be used to detect particular species or the presence of novel genotypes in a sample (Wang et al. 2002, 2003a; Rich et al. 2008). A recent review outlines microarray studies in microbial ecology (Gentry et al. 2006), including an overview of the types of probes included on arrays used in different types of studies. From this point in this review, we assume that the question of interest requires detection of only unique genes.

Probe design—Many computer programs are available to aid in the design of long oligonucleotide probes for microarrays. These programs usually determine specificity and the potential for cross-hybridization potential of all probes designed. Some of these programs are commercial, but many are open source and freely available for academic use (Li et al. 2002, Herold and Rasooly 2003, Nielsen et al. 2003, Rouillard et al. 2003, Wang and Seed 2003, Chou et al. 2004, Reymond et al. 2004, Chung et al. 2005, Li et al. 2005, Nordberg 2005, Stenberg et al. 2005, Schretter and Milinkovitch 2006). When looking for a software package, important considerations include on what basis it chooses probes, what type of input data it requires, what format the output data will be in, how easy it is to install, and how easy it is to use. Ideally, there will be empirical evidence available on how well the software has worked in designing probes for microarrays already in use. Another key consideration is whether the software is installed and runs on a local machine or on a remote machine (e.g., entering your data via a Web site). Running programs locally has the benefit that data are secure and private, whereas running programs remotely depends on the good will of someone else, and in essence involves passing your sequence set onto someone else's machine for processing. On the upside, the machine at the other end may be more powerful and therefore quicker (or not!) than what is available locally, and the maintenance and installation of the software is someone else's responsibility. You also need to ensure that you enter your sequence data in the appropriate direction (see probe direction section above). More detailed outlines of software considerations, as well as desirable probe characteristics, can be found in specific reviews of the topic (Millard and Tiwari 2009). As a good starting point, the authors have found that the Picky and Yoda software packages are both very user friendly. They are available from complex.gdcb.iastate.edu/download/Picky/index.html and pathport.vbi.vt.edu/YODA, respectively. In addition, many oligonucleotide software design packages are listed, and some reviewed, at nebc.nox.ac.uk/tools/bioinformatics-docs/other-bioinf/oligo-nucleotide-design.

Controls—Control spots are vital in the assessment of quality, sensitivity, and reliability of microarray experiments. Different types of controls can be used to assess various quality aspects. We strongly recommend the inclusion of spike-in labeling controls—probes for RNA that will be spiked into the sample at defined concentrations, before the labeling procedure. They can be used to estimate the minimum amount of transcript as well as the minimum change in transcript abundance detectable by the array technology being used. It is important to include a sufficient number of spike-in controls so that they can be added at varying concentrations to cover the signal intensity range of the experimental transcripts. Control probes should be placed on the array such that they appear across the spatial dimensions of the array. If sufficient controls are included, they can be used for data normalization (covered below). Affymetrix arrays include probes for detecting spike-in hybridization controls as a default, but extra probes for labeling need to be requested. Commercial control probes and their partner targets are available from companies such as Stratagene. Obviously, all spike-in controls should be for organisms other than those being investigated and should show no cross-hybridization to the experimental genomes. When working with unusual organisms, especially those for which no sequence is available in the public databases, it is worth checking with the company involved to ensure that their controls do not contain sequences similar to anything in your samples. If you don't have sequences for your own samples, or you cannot get information on control sequences from the company, you may need to run test hybridizations without the labeled control targets added to ensure that there is no cross-hybridization. The inclusion of appropriate controls at the fabrication stage is particularly pertinent for virus-focused microarrays, where virus message may or may not be present at all in the early stages of the experiment (i.e., the uninfected state of a transcriptional profiling experiment). This is discussed further in the data normalization section below. It is also useful to include empty spots, or probes for which no spike-ins will be added, which can provide an indication of nonspecific background signal. For spotted arrays, print-buffer spots are useful to check that no carryover effects occur during the printing of spotted arrays, as such artifacts may especially compromise the measurement of weakly expressed genes.

Genome annotation is often a process in flux, changing as better bioinformatics tools are developed and more experimental information becomes available. It is therefore useful to be able to change the annotation of genes on the array. It is possible for bioinformaticians to re-annotate arrays themselves for all platforms, but it is not always trivial. We have found that Affymetrix will support the need to change the appropriate files free of charge for a while, but will eventually start charging for this service; we suggest that the number of times Affymetrix will carry out this process be negotiated with them as part of the design contract. Importantly, once files with the updated annotation are included, old array data can be reanalyzed in light of the newly annotated protein coding genes, ncRNAs, or regulatory elements.

Microarray experimental design—As in any scientific endeavor, appropriate experimental design is crucial. Many aspects of planning microarray experiments are the same as for other types of experiment requiring statistical analysis of the resulting data. For those who are not comfortable with statistics, there is a solution: find a collaborator who is. This is not a glib statement, it is a serious recommendation. Planning your experiment with someone who is familiar with microarray statistics and experimental design can mean the difference between generating data that allows you to address your questions of interest or generating data providing little or no scope for meaningful analysis.

The design will include defining the type and number of samples needed, certain aspects of their preparation and replication, the number of slides to be hybridized, and what samples will be hybridized to the same slides in the case of two-color experiments. Too often, researchers perform an entire experiment and then provide a block of data to a statistician to be analyzed, having unknowingly introduced bias or without including appropriate replicates. When this occurs, it can make analysis difficult or impossible, meaning that all your time, samples, and money have just been wasted. Below we briefly outline some basic considerations for experimental design of microarray experiments.

Replicates: Different samples of the same type (e.g., samples of the organism exposed to the same treatment) are referred to as biological replicates. If you measure the same sample twice (e.g., take the same extract and put it on two microarray slides), this is a technical replicate. Common questions that arise when planning a microarray experiment are how many biological and technical replicates are needed? If your aim is to compare gene expression levels between treatments or conditions, then measurements from biological replicates are essential. We recommend that the minimum number of biological replicates considered for most situations is three, with four to six a more desirable number for spotted arrays. Note that if variability is high, even this number of biological replicates will be inadequate for certain types of analyses. Be wary of limiting the number of samples or experimental replicates too much based purely on cost or ease of obtaining samples, as this may lead to the inability to derive useful information from the experiment (and the money would have been better saved than spent on the microarrays).

Technical slide replicates inform on the signal variation due to "uninteresting" differences such as different slides, different hybridization chambers, different researchers carrying out protocols, etc. Technical replicates allow for certain types of quality control checks as well as providing greater precision for a given measurement. To be able to glean meaningful data from such replicates, it is necessary to use sophisticated statistical software to set up statistical models that take technical replication into account.

A different sort of technical replication common in microarray systems is the inclusion of replicate probes on the arrays. This is especially common with arrays where there are only a small number of unique probes, as is the case in many viral–host arrays. Replicate probe spots provide information about signal variation related to position on the array. Options for the analysis of replicate spots include taking a simple arithmetic average of the measure for these replicate spots or using an error term in the statistical model for each gene to account for the variation between spots (Smyth et al. 2005). We recommend the latter. It is important to choose software capable of taking replicate spots into consideration and knowing how such replicate spots are treated. For example, some software assumes that spots with the same name are the same probe, whereas other software assumes that spot replicates are equidistant across the array.

In broad terms, we recommend as many true biological replicates as possible, with technical replicate spots within arrays providing useful information about technical variation if the data are handled appropriately. Although technical slide replicates are important to run during the initial setting up of your array-related protocols, they can be of limited use later on. We generally do not use technical replicate slides once the system has been set up, as biological variability is usually much greater than the technical variation between measurements, providing little additional knowledge of the biological system. Therefore we suggest adding more biological replication rather than including technical replicates.

Culture infection concerns: For those studying viral systems grown in lab culture, the division between biological and technical replicates, and single and pooled samples, is often not clear. Every flask of a host–virus system is a pool of organisms. Cultures in two flasks from the same exact source sample are similar to a technical replicate: measuring gene expression in these two samples is likely to give you an idea of how technical differences (slight temperature or light variation, flask conditions) affect gene expression in each pool of virus–host in each flask. We therefore suggest that cultures be grown separately over many generations to get an indication of the biological variation possible in this host–virus system. Where possible, we also suggest carrying out biological replication at different times (for example 1 month apart) to further control for differences that may relate to the particular run of an experiment. Another consideration for virus infection experiments when host gene expression is being investigated is carrying out paired experiments where each replicate culture is divided into two subcultures. One of the subcultures is infected with the virus and the second serves as the uninfected paired control. This can help reduce potentially irrelevant variability related to culture physiology that is not related to the infection process. It is important to note that standard statistical tests commonly used in microarray analysis often include the assumption of independent, identically distributed measurement errors. This means that each measurement from a sample you consider a biological replicate should be an independent measurement from truly different

cultures and not from quasitechnical replicates, such as flasks grown from the same original culture.

Pooling: Pooling biological samples is often considered as a way to keep costs down (as a given number of samples can be measured using fewer arrays) or to reduce "noisy" variation. However, pooling biological replicates allows you to measure only a mean value for those samples. This leads to the lack of a measure of the biological variability in the system—a measure that is essential to determine the significance of changes in expression of one condition relative to another. If you do not know the inherent variability in your expression levels for a given treatment, you cannot determine when expression is significantly different or within the level of normal variability. In general, we advise against pooling, as it is likely to mask results of interest, while usually not providing any real benefit. If you decide to pool samples anyway, it is vital to take the pooling into account when interpreting your experimental results. As a rule, pool samples only if the experiment is purely exploratory, for example, providing preliminary data, and if you intend to screen all candidate genes in each biological replicate by other techniques such as quantitative RT-PCR.

Variability and confounding: Experiments should be designed in a manner that ensures that sources of technical variability are not aligned, or *confounded,* with treatment types. Examples of confounding include the use of arrays from different batches for different conditions and the use of spotted arrays printed early in a print run for one condition and those from late in the run for another condition. To circumvent these types of problems, array use should be randomized. The order of slide use can be randomized by generating a set of random numbers. An excellent place to get random numbers for this purpose is www.random.org/sequences.

Wherever possible, one researcher should carry out a particular task for all samples. If this is not feasible, it is important to build the design so that each researcher looks after equal numbers of samples from each condition—preferably for paired infected and uninfected samples where appropriate. In a similar manner, it is important to ensure that hybridization of microarrays from one condition are not hybridized on one day and those from the other on a different day. This way you will not confound the effect of the researcher or the day on which they were handled with the condition itself.

Microarray hybridization designs—A variety of experimental hybridization designs are commonly referred to in the microarray literature (Kerr and Churchill 2001). The key distinction between the design types is whether they allow direct or indirect comparison of samples. Direct comparisons refer to the hybridization of two samples to a single slide, providing a ratio indicating the relative expression levels of the two samples. Indirect comparisons refer to taking measurements from different slides and comparing them. Single-color designs, as used with the Affymetrix platform, are a type of indirect design. One sample is applied to each slide, and comparisons take place by considering biological replicate measurements of treated versus untreated samples. As such, one-color designs are relatively straightforward to devise and analyze. Two-color designs are more complicated. Thus, the rest of this section provides an overview of two-color microarray designs specific to spotted microarrays.

A common indirect design used with two-color microarray experiments is the reference design. In a reference design experiment, each sample is hybridized onto an array along with a *reference sample* (Fig. 2). This reference sample should ideally have a hybridization signal for all genes of interest during the course of your experiment, as you will be working with expression ratios. A common and generally effective choice for a reference sample is a pooled sample made from aliquots from each of the samples in the experiment (Kerr et al. 2007). Genomic DNA can also be used as a reference, although a different labeling strategy (i.e., labeling DNA, not RNA) must be used to generate such a reference sample. Despite this, it provides advantages in that every gene in your organism is represented at the same level above background, and it enables comparisons across experiments. In reference design experiments, the signal ratio is made up of your sample signal compared to a reference signal, and the ratios from different slides are then compared to one another. The indirect nature of reference design experiments makes them somewhat less efficient than direct designs, and they require more arrays (see below). However, they are relatively more straightforward and flexible than other designs.

An alternative approach for two-color arrays is direct designs (Fig. 2). These can provide a more accurate estimate of differences between samples, as each sample is hybridized to the same slide as the sample it is being compared with. Much of the technical variation is cancelled out when these ratios are taken. In this model, uninfected samples could be compared directly with infected samples taken from the same time point.

A common extension to direct designs is loop designs (sometimes extended to interwoven loop designs) (Kerr and Churchill 2001, Kerr 2003) (Fig. 3). Here again, two samples of interest are hybridized to each slide; however, these designs involve a combination of direct and indirect comparisons. By comparing two conditions through a chain of other conditions, samples can be compared directly with other samples with a multiple-pairwise methodology (Pirooznia et al. 2008). These designs have the potential to be more efficient than a standard reference design and have stronger statistical power, but are considerably more complex (see Fig. 3). We recommend ILOOP, a freely available Web-based program, as a useful tool in finding optimal loop designs for two-color microarray experiments (Pirooznia et al. 2008). Be aware, though, that not all analysis software is capable of handling data from loop designs.

For virus–host studies using two-color arrays where either virus or both host and virus gene expression is to be investigated, we generally prefer a reference design. This is because uninfected samples have no real signal from the virus probes, which makes taking a direct ratio between uninfected and

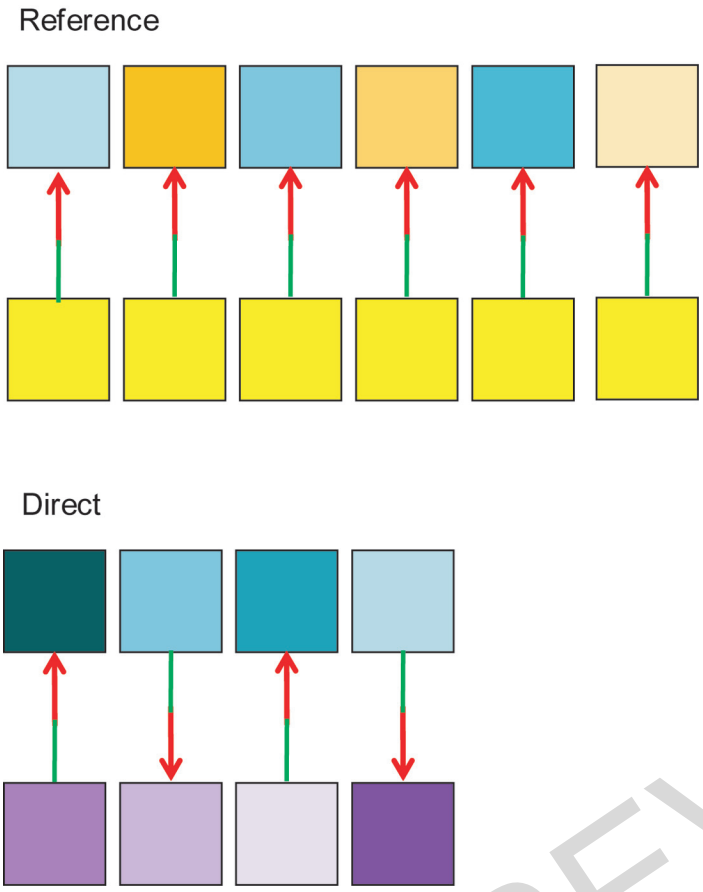

Fig. 2. Examples of reference and direct designs. Samples connected by an arrow are hybridized to the same array. Dye color can be inferred by the direction of the arrow (arrow head, Cy5 dye; tail, Cy3 dye). Here, we have also represented this by coloring the head of the arrow in red and the tail of the array in green. The reference design is shown with two experimental sample types (orange and blue), one reference sample (yellow), three biological replicates (shades of orange and blue), and no dye swaps (arrow directions) with a total of six slides. The direct design is shown with two experimental sample types (blue and purple), four biological replicates (shades of blue and purple), two pairs of dye swaps (arrow directions) with a total of four slides.

infected samples very problematic, as would be done in a direct or loop design. If, however, only host gene expression is of interest, direct or loop designs can work well. In a two-color experiment, each slide will have two samples hybridized to it. One will be labeled with Cy3 dye, and one with Cy5 dye. A dye swap involves labeling samples from a particular condition with the Cy3 on one slide, and with Cy5 on another slide. The aim of dye swaps are to avoid identifying genes as potentially interesting when in fact a relatively strong or weak signal is due to a bias associated with the dye a sample has been labeled with. A dye swap can be carried out with technical or biological replicates; however, for reasons discussed above, we recommend biological dye swaps if direct or loop designs will be used. For example, a sample pair from the treated and untreated conditions are labeled with Cy3 and

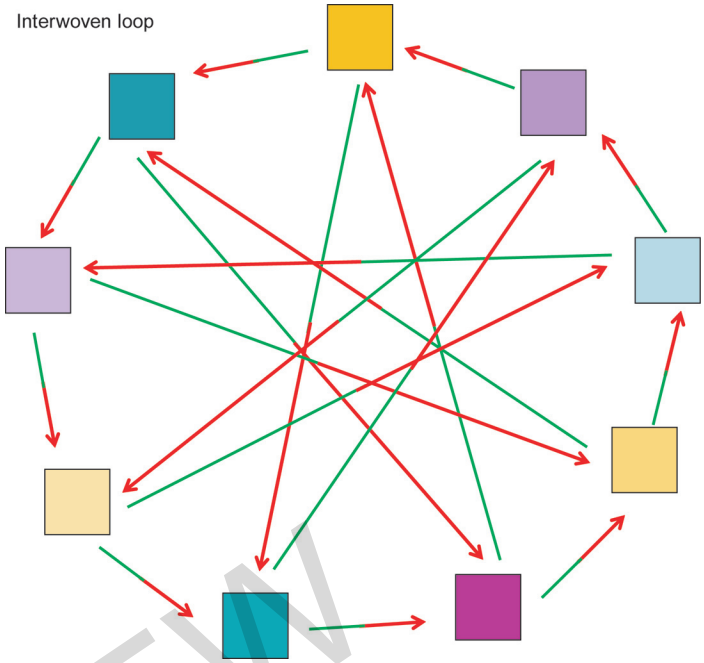

Fig. 3. Example of an interwoven loop design shown with three experimental sample types (blue, orange, and purple), three biological replicates (shades of blue, orange, and purple), each sample hybridized the same number of times and with both Cy3 and Cy5 (also the same number of times), with a total of 27 slides.

Cy5, respectively, whereas samples from another biological replicate pair (from the same treated and untreated conditions) are labeled instead with Cy5 and Cy3, respectively. For reference designs, spot ratios biased by the dye used can be avoided by labeling the reference sample with a particular dye and the experimental samples with the other dye for the whole experiment.

Sample labeling and microarray hybridization—Once you have designed your microarray, determined the experimental plan, and have extracted nucleic acid samples in hand (see "Case studies" in "Assessment" for in-depth examples), you are ready to proceed with the sample labeling and array hybridization. The method used for labeling your samples depends on the chemical nature of your sample (i.e., DNA or RNA) and the amount of sample available (low amounts of starting material may require an amplification of message step). Labeling of mRNA requires the use of reverse transcriptase, whereas labeling DNA requires the use of the Klenow fragment of DNA polymerase (i.e., minus the proofreading activity). Nucleotide mixes for these labeling reactions may require some optimization depending on the GC content of the systems under study. One needs to decide whether to use a direct or indirect labeling method (for spotted arrays); random, specific, or oligo dT primers; an amplification of message step. Some commercial kits using derivatives of the Eberwine (1996) method claim to quantitatively amplify message by up to a millionfold, but are very expensive. Direct labeling is the cheapest method for

spotted arrays and involves directly incorporating a florescent label conjugated to a nucleotide during polymerization of the complementary strand. The increased size of these synthetic nucleotides causes an unavoidable decrease in efficiency in the labeling reaction. Alternatively, indirect labeling involves the incorporation of an aminoallyl-modified nucleotide in the initial step followed by a second step involving the chemical addition of a fluorescent dye ester to the aminoallyl-modified nucleotide. The structural similarity of aminoallyl nucleotides to normal nucleotides circumvents the lower labeling efficiency related to direct labeling and generally gives stronger signal strength. We therefore recommend indirect labeling despite the fact that it is a more expensive method.

The precise hybridization conditions will also need to be optimized for each microarray. The optimum temperature for hybridization should be empirically tested with the temperature calculated by probe design software serving as an initial guide. Decreasing or increasing hybridization temperature (leading to a respective increase and decrease in cross-hybridization potential) can have a profound impact on the quality of the data. Volumes and sample buffer (typically 3× SSC, 0.1% SDS) can also be manipulated to optimize hybridization conditions. Once conditions have been optimized, they can be kept constant for a particular array and sample type. Indeed, it is essential that hybridization conditions are kept identical within a particular experiment. If different sample types are to be used with the same array, however, it may be necessary to change hybridization conditions between experiments. For example, for environmental samples you may wish to decrease hybridization temperature to maximize signal, or alternatively increase temperature to increase specificity of signal. It is essential for researchers to realize, however, that such differences in conditions will prevent a direct comparison between experiments.

Image acquisition and quantification—Once the sample has been hybridized, microarray signal intensities are collected via image acquisition with a microarray scanner. Software is then used to visualize the image, find features (i.e., the spots), and quantify the signal in each feature. The scanner to be used for microarray image capture depends on your platform and the available infrastructure. Affymetrix microarrays require a specialized Affymetrix scanner, and scanning is generally carried out at an array facility. For other platforms, scanning may be done with a variety of scanners in an automated, high-throughput fashion using standard settings for laser power, image position, and pixel size or on a slide-by-slide basis with parameters optimized for each microarray. If you are purchasing a scanner, your choice will depend on factors such as the resolution required (between 5 and 10 μm is usual for standard spotted microarrays), the number of microarrays you intend to analyze, and the likelihood of requiring more than the two lasers. Excellent scanners commonly used include the Axon GenePix series, PerkinElmer ScanArray series, and Agilent scanners. Image analysis software is generally supplied with the scanner, which ensures that the image is in the correct format for processing. For example, the GenePix series use GenePix Pro software, whereas PerkinElmer suggests ScanArray Express for their ScanArray series. In principle, however, any scanner can generate images which can be assessed using any microarray image analysis software. Other suitable microarray software packages include BlueGnome, ArrayVision (GE Healthcare), and ImaGene (BioDiscovery), all of which perform well. We therefore recommend that you try out a number of different software packages (demo formats can generally be downloaded from the Web) and choose according to your technical requirements and ease of use.

Data processing—Microarray data are inherently noisy and require careful processing before statistical analysis. Pre-analysis steps include quality control, background correction, and normalization. If you are working on a system with more than one probe per gene (e.g., Affymetrix), there will also be a summarization step, where a summary measure for a gene is generated from the multiple probes representing that gene. A good overview of the steps involved in preprocessing microarray data are given in chapter 1 of the 2005 Bioconductor book (Huber et al. 2005). Here we outline basic considerations, but direct the reader to other chapters of the same book that cover these topics in greater detail.

Quality assessment of the microarray data can include a variety of methods, for example, looking at regenerated images of background signal to assess spatial irregularities, plots of log signal value compared to intensity pre- and post-normalization (so-called MA plots), and box plots of slide data pre- and postnormalization and assessing spot quality flag information where this has been supplied by the image capture software. Hierarchical clustering (see "Data Analysis" below) of data pre- and postnormalization is very useful to look for whether the data are grouping according to non-interesting factors such as which day an analysis was carried out. For quality assessment methods for Affymetrix arrays, we also direct the reader to a white paper on the Affymetrix Web site (www.affymetrix.com/support/technical/whitepapers/exon_arrays_qa_whitepaper.pdf). Image capture software and analysis software manuals usually include some information on quality assessment and control. Of note is that different software programs often tag spots with indicators of measurement quality, and different analysis software will provide different ways of dealing with these tags. We tend to work only with high-quality spots, as microarray data are noisy enough without including lower-quality spots, even if downweighted, for analysis purposes.

Background correction methods aim to remove background signal from spot signal measurements. Background correction is applied to data before normalization, although it is not always obvious, as it may be included in a software choice that encompasses a number of steps, such as background correction, normalization, and summarization, under a single command.

Normalization describes a variety of methods to correct microarray data for variation introduced by experimental procedures rather than biological differences between samples. For example, differences in detection efficiency, dye labeling, fluorescence yields, and total amount of cRNA/cDNA loaded onto the array can affect the measured signal intensities. If neglected, these factors can be erroneously interpreted as changes in expression. It is therefore important to correct for such variability before further data analyses. Choosing an appropriate normalization method is a crucial step, since it has considerable influence on the results (Hoffman et al. 2002). Normalizations can be applied to data within each array, to account for intra-array issues such as dye bias, location-dependent bias, and intensity-dependent bias. They can also be applied across arrays, usually referred to as between-array normalization, to address scale differences between arrays. The aim here is to make data comparable across arrays. For example, the popular quantile normalization (Bolstad et al. 2003) shifts the distributions of signals on arrays, resulting in the same empirical distribution across arrays and across channels.

Depending on the type of microarray platform used, different normalization schemes can be applied, with within-array normalizations being applied before between-array normalizations. For one-color microarrays, the first data processing step is the calculation of the summary indices for each gene based on the corresponding probes. Two methods are widely used for this task: Microarray Suite (MAS)/GeneChip Operating Software (GCOS) by Affymetrix and Robust Multi-array Average (RMA) introduced by Irizarry and coworkers (Irizarry et al. 2003). MAS/GCOS calculates the summary indices by averaging probe signals for each array individually. In contrast, RMA simultaneously calculates summary indices for all arrays included in the experiment. Because it incorporates both probe and array effects in the calculation, it can correct for systematic difference in signal intensities between arrays and thus provides a first level of normalization. Note that this implicitly assumes that total (logged) expression intensity should be equal for the different arrays. The obtained distributions of summary indices reflecting the expression levels can be further adjusted subsequently. For MAS, scaling of the median expression to a chosen level is usually performed, if we can assume that the majority of measured genes are not differentially expressed. A common additional adjustment for RMA-processed data are quantile normalization transforming expression levels to have the same distribution for different arrays (Bolstad et al. 2003).

Several comparisons have been conducted so far between MAS and RMA processed data, but the results remain inconclusive. For example, RMA performed favorably for a benchmark dataset with a small number of spike-in controls, for which the concentrations were known (Cope et al. 2004). In another study, where a considerably larger number of spike controls was used, MAS outperformed RMA (Choe et al. 2005). The latter study, however, has been criticized for the use of a flawed design (Dabney and Storey 2006, Irizarry et al. 2006).

These contrasting results for different datasets indicate that the performance of normalization procedures can strongly depend on the dataset being processed. This is not surprising, since all of the normalization procedures are based on specific assumptions which may not hold for different datasets. When assumptions are violated, normalization might fail and can lead to erroneous results. Thus, it is of vital importance that researchers carefully check the suitability of assumptions.

To normalize two-color arrays, the signal intensities of the cohybridized samples are used. The most basic method is global normalization, which is the linear scaling of the total intensities in each channel to the same value. Nonlinear effects in the signal intensities are frequently observed, however—a phenomenon referred to as dye bias (see earlier discussion of dye swaps). This kind of artifact causes signals to be systematically larger in one channel for the low-intensity range even after balancing the average signal intensities of both channels. To cope with such bias, several intensity-dependent normalization procedures have been introduced. Some of these elaborate methods can cope with potential spatial artifacts (Yang et al. 2002, Yang and Speed 2002, Futschik and Crompton 2004). It is important to emphasize, however, that such nonlinear methods require that most of the genes assayed are nondifferentially expressed or that the differential expression is symmetrical, i.e., the overall up- and downregulation is balanced. This is not always the case for virus-derived samples—for example, if overall host genome expression declines during viral infection or if the array is dominated by virus probes only.

A solution is to scale the data to signal from reference probes that are kept constant across the experiment. The employed reference can be of either endogenous (e.g., so-called housekeeping genes) or exogenous (i.e., spike-in controls) origin. Normalization based on housekeeping genes predates most other normalization procedures for two-color arrays (DeRisi et al. 1996) and assumes that the genes chosen really do not change under the experimental conditions. If you cannot guarantee a set of expressed but nonchanging genes, we recommend you use spike-in controls. Validation, for example using quantitative RT-PCR, can be used to select an optimal normalization procedure (Lindell et al. 2007). Reference-based normalization methods seem intuitively very attractive, but careful checks are necessary to ensure they are working as desired. We have performed comparisons of several normalization schemes for microarray experiments for virus-infected *Prochlorococcus* cells, for which we expected that the majority of assayed genes underwent differential expression (Lindell et al. 2007). The microarrays were customized Affymetrix GeneChips including spike-in hybridization controls. Neither the normalization by spike-in control nor by rRNA selected as housekeeping genes yielded superior results compared with other normalization strategies when using qRT-PCR data as the gold standard in an independent comparison (Lindell et al. unpubl. results). Possible reasons for the inferior performance of spike-in controls could be their limited number and their

restricted intensity range on Affymetrix GeneChips. In the case of normalization based on rRNA, their high expression levels and their possible saturation on the array might cause difficulties in their use as reference. Having said this, we have had satisfactory results using reference-based normalizations on two-color arrays. As with all aspects of microarray experimental planning, nothing should be taken for granted. The use of spike-in controls combined with external checks of representative genes (e.g., quantitative RT-PCR), and plotting of data and controls pre- and postnormalization, are necessary to have confidence in your microarray results.

Data analysis—Clustering is a popular approach to explore large microarray data sets. The aim of clustering is to assign genes or arrays to groups based on their similarity: genes or arrays displaying similar expression profiles, where you define what *similar* means by choosing an appropriate dissimilarity measure, should be assigned to the same clusters, whereas genes or arrays displaying distinct expression profiles should be placed in different clusters. Using cluster analysis, we can detect prominent expression patterns, coexpressed genes, and similar conditions, which can be further examined for their biological meaning. Many different clustering methods have been applied to microarray data. Generally, two types of clustering exists: hierarchical and partitional (Jain and Dubes 1988). Hierarchical clustering creates a set of nested clusters, so that clusters on a higher level are composed of smaller clusters on lower levels. The resulting hierarchy of clusters are conventionally presented as a treelike structure, the so-called dendrogram. To perform hierarchical clustering, we proceed in a sequential manner. In each step, we calculate the pairwise distances between all clusters and merge the ones with the smallest distance. In contrast, in partitional clustering, all objects are simultaneously assigned to clusters. This type of clustering typically aims to optimize an objective function for a given number of clusters. A prime example of this clustering approach is *k*-means clustering, which seeks to minimize the between-cluster variation in an iterative manner. Hierarchical and partitional clustering both have their advantages and disadvantages. One strength of hierarchical clustering is that it defines relations between and within clusters. However, the sequential procedure used can be sensitive to the high noise level that is frequently contained in microarray data. Partitional clustering tends to be more robust to noise, but commonly fails to present within-cluster structures. Notably, some partitional clustering methods can reveal internal cluster structures and are still highly noise-robust. For instance, fuzzy clustering assigns genes graded membership to clusters, i.e., it can indicate how strongly a gene is associated with a cluster. Thus, genes that are tightly clustered obtain a large membership (with values close to 1), whereas genes with noisy expression patterns receive low membership values. In contrast to conventional clustering methods, fuzzy clustering even allows genes to be placed in several clusters (Futschik and Crompton 2004). A variety of software packages have been developed for clustering analysis of microarray data. Popular standalone software packages for performing and visualizing hierarchical clustering are Cluster 3.0 and Java TreeView, respectively (bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster; jtreeview.sourceforge.net). Alternatively, several Web servers enable online cluster analyses (e.g., EBI expression profiler, www.ebi.ac.uk/expressionprofiler).

A difficult question is how many clusters can be reliably retrieved from the observed expression data. The difficulty arises from the complexity of microarray data and a high noise component. Frequently, different cluster structures are apparent depending on the resolution. For example, several main clusters may exist, but each might display subclusters. Furthermore, the noise component can lead to overlapping clusters, for which a separation might be not justified. Despite these difficulties, some tools have been developed to help researchers choose the accurate number of clusters and judge the reliability of the results. Classic approaches are based on the so-called figures of merit. These measures capture a desired feature that we seek to optimize. To obtain an accurate clustering, we seek to optimize the figure of merit. An example is the Dunn index, defined as the ratio between the minimal intracluster and the maximal intercluster distance (Dunn 1974). Used as a figure of merit, we aim to minimize the Dunn index to obtain tight clusters that are well separated. A common drawback, however, is that these measures assume non-overlapping clusters, which are typically not the case for microarray data. An alternative approach is based on measuring the stability of clusters with respect to data perturbations, e.g., through resampling or addition of noise. According to this concept, reliable clusters are those that are maintained in spite of perturbation (Bittner et al. 2000, Levine and Domany 2001). For instance, we could cluster genes using only a subset of measurements and examine the resulting clusters. As reliable clusters should not depend on single measurements, they should still be detectable using partial data. Finally, the inspection of the functional composition of clusters can give us clues about reliability. This strategy assumes that genes sharing the same function tend to be coexpressed and thus should be placed into the same cluster. By optimizing the enrichment of functional gene categories, the number of clusters can be chosen (Gibbons and Roth 2002).

Despite these tools, assessing the quality of clusters remains challenging. Researchers are advised to apply several clustering approaches to their datasets, as a single method often works well with some data sets, but may perform poorly in others. In practice, we propose that clustering should be seen primarily as exploratory analysis that can then be followed up with more stringent computational and experimental examination. A good introduction to clustering as applied to microarray analysis is given in chapter 7 of Wit and McClure (2004).

Statistical significance of differential expression—A typical aim of microarray experiments is to identify genes that are differentially expressed under different conditions. Because of the large number of genes measured by microarray technologies and random fluctuations in gene expression, we expect a number of

genes to show differences in expression simply by chance. Therefore, we use stringent statistical approaches to analyze microarray data for differential expression. In classic statistical testing, we compare a test statistic calculated from our data to a distribution of that test statistic expected under the *null hypothesis* that the gene is not being differentially expressed. The comparison of our test statistic to this distribution results in a *P* value. *P* values indicate how often you would expect to see data as extreme as that you just observed if the gene is not being differentially expressed. *P* values are *not* direct indicators of the probability of the gene being or not being differentially expressed. For differential gene expression studies, a small *P* value (e.g., <0.01 for a single test, see below) indicates that we would rarely see a test statistic that extreme if the gene measurements in one condition really were from the same distribution of expressions as the condition we are comparing them to. For example, with $P = 0.01$, we would expect to see values this extreme in about 1 of every 100 samplings if the data we are working with were sampled from the same distribution as the one we are comparing it to. Another school of statistics, the Bayesian school, provides more readily interpretable probabilities, but can be harder to apply in practice. *P* values are closely related to a more readily interpretable value, the false discovery rate (Benjamini and Hochberg 1995), which is commonly used to evaluate microarray results and is discussed further below. For a good introduction to all the aspects of statistical testing for differential expression of microarray data mentioned here, we recommend chapter 8 of Wit and McClure (2004).

There are two general categories of statistical tests:

1. Parametric tests. These assume that the populations being compared can be described by particular distributions. For instance, certain tests assume that the underlying distribution is normal. For microarray analyses, the distribution being referred to is the distribution of expression values for a given gene; it is not about the distribution of expression values across genes. Typically, having sufficient biological replicates for parametric statistical analysis is a challenge. Some available statistical methods, so-called *local pooled error* methods, aim to decrease the number of replicates required to carry out reliable statistical tests by pooling the sample variation of genes with similar expression intensity; i.e., they fit the error with respect to the signal intensity based on the observed data (Baldi and Long 2001, Tusher et al. 2001, Jain et al. 2003, Smyth 2004).
2. Nonparametric tests. These do not make assumptions about the underlying distribution. Such tests commonly rank the data in value order and then carry out tests based on the order. This is very useful when we do not know the underlying distribution of what we are measuring. However, nonparametric tests have less power to detect differences and thus require more replicates to give similar confidence when interpreting your data.

Parametric tests involve comparing the test statistics calculated using your data to a standard distribution with particular parameters. If you have sufficient biological replication in your experimental design, you can instead compare your test statistic to the expected null distribution of that statistic (i.e., when a gene is not differentially expressed), which you generate through bootstrapping, or through permutation, of your own data set. For example, instead of comparing a *t*-test statistic to a standard Student *t* distribution to determine a *P* value, you would generate a distribution of the test statistic from resampling your own data set in defined ways and compare your test statistic to that distribution. Issues relating to bootstrapping and permutation are discussed in Wit and McClure (2004).

Until this point, we have really been discussing testing for differential expression of a single gene between two conditions. For a comparison yielding a *P* value of 0.01, we would expect to see data this extreme in about 1 of every 100 tests if the gene were not differentially expressed. So, if we test 10,000 genes, and 10 genes are really different between conditions, we could still expect around 110 genes with *P* values less than or equal to 0.01. Of these, 100 had "extreme" mean expression values due to chance, with 10 of them truly differentially expressed between conditions. We need to try and increase our ability to discern genes that are truly differentially expressed. In other words, we need to adjust our results to take into account that we are carrying out multiple testing. Different types of multiple testing corrections are used in microarray studies (Dudoit et al. 2003), but arguably the most popular is the false discovery rate (FDR) (Benjamini and Hochberg 1995). The FDR method allows you to define the proportion of false positives you would find tolerable in your results. It then returns the largest list of genes classified as differentially expressed that includes this specified, expected percentage of nondifferentially expressed genes. Numerous software tools can help researchers to assess the significance of differential expression. Notably, a highly powerful and flexible platform, also for other aspects of microarray data analysis, is the Bioconductor project (www.bioconductor.org). Alternatively, more specialized software solutions such as SAM (www-stat.stanford.edu/~tibs/*SAM*) or BRB Array Tools (linus.nci.nih.gov/BRB-ArrayTools.html) can be applied to calculate false discovery rates for gene expression data.

Experiment annotation and data submission—Many journals require microarray experimental data to be submitted to a public repository as a condition of publication. Indeed, some funding agencies require researchers to agree to make their data publicly available at the end of the project; for many researchers, the most sensible way to do this will be by submitting to a known public repository such as the EBI's ArrayExpress (www.ebi.ac.uk/arrayexpress) or the NCBI's GEO (www.ncbi.nlm.nih.gov/geo). Both these databases require adequate annotation of the experiment, including at least the information required by the Minimum Information about a Microarray Experiment (MIAME) standard (Brazma et al. 2001). For researchers engaging in environmental experiments, it is worth also referring to the MIAME Env extension (Morrison et al. 2006).

The importance of annotating data properly and making it publicly available has led to many new "minimum information" checklists for different domains. The Minimum Information for Biological and Biomedical Investigations (MIBBI) portal (http://mibbi.sourceforge.net) is a good place to look for standards lists. Data standards list *what* the minimum requirements are for describing a data set. *How* the data should be described is usually defined by a set of terms or an ontology. For key microarray experimental concepts, the Microarray Gene Expression Data Society (MGED) make the MGED Ontology available. Use of this ontology, sometimes in combination with other domain-specific ontologies, is highly recommended when annotating your experiments. The Open Biomedical Ontologies (OBO) consortium (Smith et al. 2007) Web site (http://obofoundry.org) is probably the best place to see if ontologies relevant to your area already exist or are under development.

Public data repositories offer tools to facilitate submission of data, and some external tools support export in a format acceptable to the public repositories. A "non-exhaustive list of possible MIAME compliant software" is held on the MGED site at http://www.mged.org/Workgroups/MIAME/miame_software.html. For researchers in the environmental sciences, the software maxdLoad2 supports annotation using the MIAME Env extension of the MIAME standard (Hancock et al. 2005).

Summary—Microarrays offer great potential, but the statistical analysis needs careful consideration from the outset. Our general recommendations are as follows:

1. If the samples required to address a particular question are too difficult to generate or collect, you should adjust the question you are asking. Your data will not miraculously provide answers to the original question if you do not have sufficient or appropriate samples.
2. If you are not experienced with statistics, then find a collaborator who is. Access to shiny software with easy-to-use menus is not the same thing as statistical knowledge.
3. Know what software you (or your collaborator) are going to use for the analysis and be sure that it is capable of analyzing data generated under a particular design. It is also a good idea to define how technical replicate spots and technical replicate slides will be handled if these are part of your design.
4. Technical replicates provide a different type of information than biological replicates and can be difficult to handle appropriately using some software. For many purposes, more biological replicates is a better option than carrying out technical replicate hybridizations.

Assessment

The marine sciences have been relatively slow to embrace microarray technology, primarily because of the historical lack of available genomic sequence. This has changed significantly over the past few years, and it is now becoming common for marine-focused researchers to develop microarrays to answer their questions of interest. The influx of molecular biology–trained researchers to the field of marine virology has served to hasten this pace further. To date, microarrays have been developed and used to study a number of aquatic viruses including cyanophages (see "Case study 1" below; Lindell et al. 2007, Millard et al. 2009), coccolithoviruses (see "Case study 2" below; Allen et al. 2006a, Allen and Wilson 2006, Allen et al. 2007), and shrimp white spot virus (Dhar et al. 2003). Although the array platforms used for these case studies are different, many aspects are relevant to any platform, especially those relating to experimentation, RNA extraction, and data analysis.

Case study 1: Affymetrix microarrays and the cyanophages—Custom-made high-density Affymetrix GeneChip arrays were designed for two marine cyanobacteria of the genus *Prochlorococcus* and two marine cyanophages. Here we describe array design considerations and microarray experiments carried out to determine the transcriptional program of the T7-like podovirus P-SSP7 when infecting the cyanobacterial host *Prochlorococcus* MED4, as well as to assess the transcriptional response of the host to infection. Transcriptional profiles for all phage genes were investigated during the latent period of infection and revealed a transcriptional program for the podovirus P-SSP7 reminiscent of that for T7. The viral genome was transcribed in three functional clusters from left to right of the genome map (Lindell et al. 2007), with the module thought to be involved in host takeover being transcribed first, then the DNA replication module, and finally the module encoding structural and packaging genes (Figs. 4 and 5). Unlike T7, however, the last three genes of the genome, including the bacterial-like transaldolase gene, were transcribed out of order. These last three genes, together with the cyanobacteria-like photosynthesis genes and a bacteria-like ribonucleotide reductase gene, were transcribed together with the phage DNA replication module, leading to the hypothesis that the products of the bacteria-like genes in the phage genome may be involved in generating energy and substrates for genome replication. Investigation of the whole-genome response of the cyanobacterial host to infection revealed that whereas the vast majority of transcripts were downregulated as infection progressed (75% of the genome), 41 protein coding genes (Lindell et al. 2007) and three ncRNAs (Steglich et al. 2008) were upregulated in two distinct expression clusters (Fig. 4). The function of these upregulated genes is still unknown, as is whether they were upregulated as a host stress response or by the phage for its own purposes. The questions raised from the results of these microarray results are active areas of current research.

Details of the procedures for this experiment can be found in the supplemental information of Lindell et al. (2007). Below, we provide a summary of the procedures used in the design and implementation of the arrays and the reasons behind their use rather than providing the details of the actual experimental procedures.

Shortly after the sequencing and annotation of two *Prochlorococcus* strains, MED4 and MIT9313 (Rocap et al.

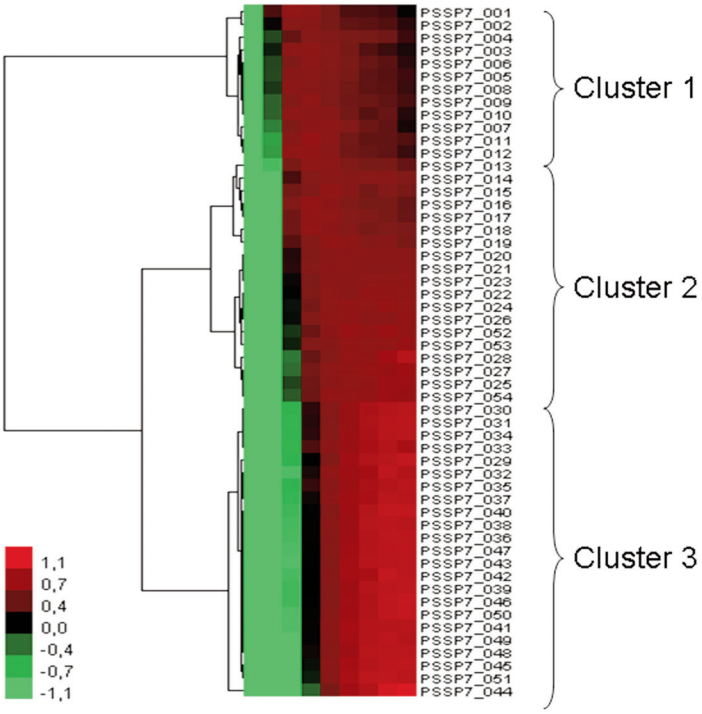

Fig. 4. Hierarchical cluster analysis of gene expression profiles of the P-SSP7 phage during infection of *Prochlorococcus* MED4 showing three distinct gene expression clusters. Transcript levels were determined by Affymetrix microarray analysis. The dendrogram appears on the left, heat map in the middle, and gene names and cluster membership on the right. In the heat map, red indicates an increase and green a decrease in expression. Time after infection appears above the heat map. Hierarchical clustering was carried out using Pearson correlation and average linkage. Input data were the average logged expression values of three biological replicates, standardized so that mean expression values for each gene equal 0 and standard deviation equals 1. Reproduced with permission from Lindell et al. (2007).

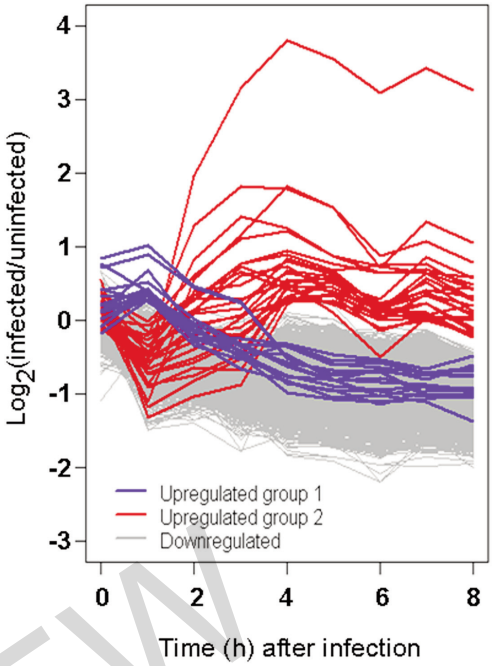

Fig. 5. Transcriptional profiles of *Prochlorococcus* MED4 genes with time after infection by the podovirus P-SSP7. Transcript levels were determined by Affymetrix microarray analysis and are presented as log2-fold change in infected cells relative to the paired uninfected cells over the 8-h latent period of infection. The results are the average of three biological replicates. Only genes whose expression levels were significant at a false discovery rate of $q < 0.05$ are shown. Blue and red indicate two significantly upregulated gene clusters. Gray indicates genes significantly downregulated at 8 h after infection. Reproduced with permission from Lindell et al. (2007).

2003), we designed the custom-made MD4-9313 high-density Affymetrix array to investigate the transcriptional response of these cyanobacteria to a variety of environmental stressors, one of which was viral infection. We therefore included the recently sequenced genomes of two cyanophages that infect *Prochlorococcus* MED4, the T7-like podovirus P-SSP7 and the T4-like myovirus P-SSM4 (Sullivan et al. 2005). Importantly, the presence of both host and viral genomes on a single array enabled the concurrent investigation of the transcriptional program of the virus during infection, together with the host's transcriptional response to this infection from the one sample on a single array. We further decided to design the array with multiple cyanobacterial and cyanophage genomes to reduce the design price per organism. In addition to the viral infection experiment described here, this array has been used successfully in numerous studies investigating the transcriptional response of *Prochlorococcus* to a variety of environmental stressors (Martiny et al. 2006, Steglich et al. 2006, Tolonen et al. 2006) over a diel cycle (Zinser et al. 2009), as well as in the investigation of small noncoding RNAs (Steglich et al. 2008).

The MD4-9313 array was designed by the researchers (D. Lindell, M. A. Wright, S. W. Chisholm, and G. M. Church) in conjunction with the Affymetrix design team. We decided on a design arrangement somewhat different from the standard for Affymetrix expression arrays, which consist of 11 probes per open reading frame (ORF), often biased toward the 3′ end. Probes for the two *Prochlorococcus* strains were not 3′ biased but rather were designed evenly along the annotated ORFs so that probes were spaced approximately every 80 bases. We further designed probes for all intergenic regions (longer than 35 bp) on both strands at intervals of about 45 bases so that probes would be present for both short protein coding genes and noncoding RNA genes not initially annotated. Probe spacing was decreased for short ORFs and intergenic regions so that 11 and four probes, respectively, were designed where possible. This same design strategy could not be used for the viral genomes, as they had only just been sequenced and had yet to be annotated at the time of design. Therefore probes were designed across the phage genomes at approximate intervals of 90 bases on both strands. Seeing as total RNA is used in bacterial arrays, care was taken to ensure that probes had no similarity to ribosomal genes to reduce the chance of nonspe-

cific binding due to their high transcript levels. Probe pairs of perfect match (identical to the sequence to be detected) and mismatch (containing a single base change at the center of the probe) were designed as per the Affymetrix standard design. Probes are designed so that their properties meet criteria determined empirically by Affymetrix. This information is proprietary and is not available to the researcher.

During analysis, an average of the signal from all the probes associated with a particular gene or intergenic region (a probe set) is used to determine the relative expression level. When using Affymetrix analysis software (MAS), the mismatch signal level is used as a measure of nonspecific hybridization. When RMA analysis, an alternative to MAS, is used, only the perfect match probes are used. Designing perfect match probes only would increase the number of genes represented on an array but should only be done if the researcher is positive they do not need mismatch probes in their analyses. This is an antisense array and is designed for protocols that generate and use labeled cDNA or cRNA sequences that are antisense to the original RNA. This array contains approximately 200,000 probe pairs over an area of 8.1 mm in a Midi array format, and each feature is 18 µm in size.

Initially, the array was tested by hybridizing labeled genomic DNA from each of the cyanobacterial strains independently. The array performed extremely well with all probe sets for predicted open reading frames, giving a reproducible signal well above background levels (generally 20-fold higher than background levels), whereas some probes sets for intergenic regions (known to be of lower quality at the time of design) did not provide a significant signal. It is important to note that the signal intensity varied up to sixfold across the probe sets even though equal amounts of DNA were present for each gene. This indicates that intrinsic differences in hybridization efficiencies exist between probe sets and highlights that absolute transcript levels cannot be determined from microarray analysis. However, standardization based on genomic hybridization signals may be useful in determining the relative levels of different transcripts if these are below saturation levels on the array. Cross-hybridization between the two *Prochlorococcus* genomes was fairly low (10% of the probes designed for one genome gave signals above background when hybridized with DNA from the other genome), indicating that it is feasible to include probe sets for multiple genomes on a single array. Indeed, no detrimental effects of the presence of the MIT9313 genome on the array were detected when running experiments with MED4.

Triplicate cultures of *Prochlorococcus* were grown under continuous light and concentrated by centrifugation to 10^8 cells mL^{-1}. Each of the three cultures was divided into two paired subcultures, one of which was infected with the podovirus P-SSP7 at a ratio of three infective viruses per cell, and the other was amended with filter-sterilized spent medium and served as an uninfected control (Lindell et al. 2007). Samples were collected by centrifugation for RNA extraction (100 mL) every hour over the 8-h latent period of infection from the infected cultures as well as their paired uninfected controls. The cell pellet was rapidly resuspended in storage buffer (200 mM sucrose, 10 mM sodium acetate, pH 5.2, 5 mM EDTA), snap-frozen in liquid nitrogen, and stored at –80°C. The Ambion product RNAlater has also been successfully used by others for storage of samples before RNA extraction. The procedure from collection of cells until freezing takes approximately 20 min. This time can be significantly reduced if the cells are harvested by filtration onto Supor-450 (Gelman-Pall) filters that are then immersed in the above storage buffer. Very high reproducibility was achieved between these three biological treatments, indicating that this number of biological replicates was sufficient and that technical replicates were not necessary when using this custom-made Affymetrix array.

We decided to grow *Prochlorococcus* under continuous light for the infection experiment (even though they grow naturally under a diel light-dark cycle) to remove the complications associated with intrinsic diel differences in expression patterns of the host (Zinser et al. 2009), although the paired treatments and controls would have controlled for such differences. Concentration by centrifugation of the cells before infection caused an unexpected problem—that of differential expression during the first 4 h after centrifugation. It was therefore fortunate that an experimental design based on paired treatments and controls was chosen, enabling us to control for transcriptional changes associated with the centrifugation at each time point. In this case, a comparison of the transcriptional response at different times after infection relative to $t = 0$ (longitudinal comparison) would have led to erroneous conclusions. A further complication with this particular host–virus system is that we manage to infect only 50% of the cells during a first round of infection irrespective of the multiplicity of infection used. This did not have any detrimental effects on determining the transcriptional program of the virus, as the cells appeared to be synchronously infected. This rather complicates elucidation of the host responses to infection, however, as uninfected cells are present in the infected treatment and could mask moderate level responses by the infected cells.

Care must be taken when extracting RNA samples for microarray analysis, as with any work using RNA, due to the high stability of RNases and the difficulty in their removal. We therefore work with nuclease-free molecular biology–grade reagents and plastics, in an RNase-free work space and with pipettes and gel boxes dedicated for the purpose. RNA was extracted using Ambion's mirVana RNA isolation kit. Immediately before extraction with this kit, the cells were thawed rapidly at 25°C and centrifuged for 2 min to exchange the resuspension buffer with the lysis buffer from the kit. Depending on the cyanobacterium/bacterium investigated, it may be necessary to add a lysozyme step to the protocol before cell lysis. After RNA extraction, contaminating DNA was removed by digestion with DNase using Ambion's Turbo DNA-free kit. The resulting RNA was then subjected to 3 M sodium acetate-ethanol precipitation

and a 70% ethanol wash to both remove the DNase reagents and nucleotides and to concentrate the sample. This yielded approximately 10–20 µg RNA after the entire procedure from approximately 10^{10} starting cells of *Prochlorococcus* MED4 (i.e., approximately 1–2 fg RNA per cell after all losses).

The mirVana RNA extraction protocol was determined to be the most suitable for our purposes based on a combination of considerations: (a) the relative ease and speed of the procedure—although it does include a phenol-chloroform step; (b) high yield and quality of resulting RNA; and (c) the retention of RNAs as small as 50 bases, which enables the assessment of expression patterns of small noncoding RNAs (Steglich et al. 2008). Other protocols tested produced significantly lower yields (Ambion's Ribopure and Qiagen's RNeasy kit), were more labor intensive (the hot-phenol method, Lindell and Post 2001), or removed RNAs shorter than 200 bases (Ambion's RNAaqueous and Qiagen's RNeasy). However, the mirVana kit is not suitable if cells are harvested by filtration onto Supor-450 filters. In that case, it is necessary to extract the RNA using the hot phenol method, in which the filter is dissolved in the organic phenol phase and all nucleic acids are released from the cells embedded in the filter (Lindell and Post 2001, Steglich et al. 2006).

Once the RNA had been extracted and DNA removed, we determined the quantity and purity of the RNA spectrophotometrically and assessed RNA integrity on agarose gels. Smearing of the rRNA bands on gels indicates that the RNA is degraded and should not be used. We find that running agarose gels in Tris-acetate-EDTA (TAE) buffer is sufficient for this purpose, although denaturing gels are generally used for more sophisticated RNA procedures to prevent secondary structure from affecting the migration of the RNA in the gel. The amount of single-stranded RNA can be determined by measuring absorbance at 260 nm in a 1-cm quartz cuvette using the conversion factor of 40 ($A_{260} \times 40$ = concentration of RNA in ng μL^{-1}). An absorbance (A) ratio at 260 to 280 nm of 1.8–2.1 indicates that the RNA is of high quality with negligent protein contamination, and a ratio of A_{230} to A_{260} of 0.3–0.5 indicates little salt or phenol contamination. Agilent bioanalyzers can also be used for assessing the quantity, purity, and integrity of your RNA before running a microarray experiment.

RNA labeling and microarray hybridization, staining, and scanning were carried out at the Affymetrix service center situated at the BioPolymers Facility of the Department of Genetics, Harvard Medical School, Boston, MA, USA. The procedures used for bacteria are quite standard, although we found that reducing the amount of total RNA used per array from the 10 µg total RNA suggested by Affymetrix for *E. coli* to 2 µg for *Prochlorococcus* MED4 provided good results and did not compromise the quality and sensitivity of the array results. Seeing as we did not test the use of this low amount of RNA with *E. coli*, we don't know if this difference is due to an exaggerated suggestion for high amounts of RNA by Affymetrix or due to intrinsic differences between the bacteria. One potential difference is related to the high growth rate of *E. coli* (doubling every hour) relative to *Prochlorococcus* MED4 (doubling every 1–2 days), which may lead to significantly higher rRNA levels relative to mRNA in *E. coli* and therefore a requirement for greater amounts of total RNA to gain similar levels of mRNA. Therefore, if material is hard to come by, as is often the case, we suggest that you test a range of RNA concentrations for your system rather than relying on the Affymetrix suggestion.

Below is an overview of the microarray procedures we used. Details of the labeling, hybridization, staining, and scanning protocols and data analysis carried out in this study can be found in the supplementary information of Lindell et al. (2007). The standard Affymetrix protocols for bacteria can be found at their Web site: http://www.affymetrix.com/support/technical/manual/expression_manual.affx.

Briefly, 2 µg total RNA was subjected to the labeling protocol. The RNA was reverse transcribed to produce cDNA which was then fragmented to 50–200 nucleotides long, purified, and end-labeled with biotin. Biotin incorporation was verified by band shift analysis. The labeled sample was hybridized to the array in an aqueous solution, and stringency washes were performed. Staining was achieved by incubation of the array with streptavidin, which has a high affinity for biotin, and finally with a streptavidin-phycoerythrin conjugate. Scanning of the array for phycoerythrin fluorescence was carried out to determine the raw signal levels for each probe. Spike-in hybridization controls were added before hybridization. In the future, we will also include spike-in RNA labeling controls at concentrations expected to span the signal intensity range to help facilitate array normalization procedures.

Independent verification of the array results were carried out by quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR). Genes were chosen for analysis to include representatives of the different expression profiles detected, genes with a range of transcript intensities and genes of biological interest where possible (Lindell et al. 2007). Therefore, we verified the array results for a subset of the genes from each of the three phage expression clusters as well as from the last gene in the genome to represent the three genes transcribed out of order on the genome. Host genes from both upregulated clusters and the downregulated cluster were included. The qRT-PCR results for the host genes were also used for determining the appropriate normalization method. Because the high sensitivity of qRT-PCR, a more rigorous DNase treatment is necessary than that carried out for microarray analyses, and controls in the absence of reverse transcriptase (no RT controls) must be carried out to ensure that contaminating genomic DNA has been removed.

Data analyses were carried out in the statistical language R using several Bioconductor packages (Gentleman et al. 2004). The array data were normalized and probe set summaries were calculated from perfect match probe intensities using RMA analysis (Irizarry et al. 2003) from within the Bioconductor program. RMA with quantile normalization was chosen based on the validation with RT-PCR data. Compared to other normalization schemes (e.g., based on spike-in controls), it yielded

superior performance (Lindell et al. 2007). This was initially surprising, since quantile normalization assumes similar overall distribution of probe intensities in different arrays, which seemed not to be the case here. However, closer inspection revealed that this assumption still holds owing to the large number of probe sets that did not show differential expression found for intergenic regions, for the P-SSP7 phage genome, and for the additional *Prochlorococcus* and phage strain on the array. Statistical significance of differentially expressed genes between infected and control cells at each time point was determined using a Bayesian *t*-test. Hierarchical clustering was carried out to determine cluster patterns of the phage genes and upregulated host genes. The reliability of clustering was determined by repeated clustering of a random resampling of subsets of genes. For both types of genes, visual inspection indicated the existence of discrete clusters. To determine the numbers of reliable clusters, a resampling strategy was used, where hierarchical cluster analysis was performed repeatedly on randomly selected subsets of genes. Clusters were considered reliable if they occurred persistently for different random subsets of genes. This strategy was used for a range of numbers of clusters and indicated that there were three reliable phage gene expression clusters and two upregulated host gene clusters (Lindell et al. 2007).

Case study 2: Spotted microarrays and the coccolithoviruses— *Emiliania huxleyi*, a calcifying marine haptophyte with worldwide distribution, is infected by the coccolithovirus family of viruses. Whereas the study of *Emiliania huxleyi* is relatively common with international researchers owing to its fundamental importance to global ecosystem function and modeling, the study of its viruses has been restricted to a handful of groups since their discovery (Wilson et al. 2009). Nevertheless the rate of discovery for this unique and interesting viral family has been phenomenal, aided primarily by the development of microarray-based tools (Allen and Wilson 2006).

Initially, the first-generation microarray was based around the model strain *Emiliania huxleyi* virus 86, EhV-86. Designed and fabricated in tandem with an ongoing sequencing project, this microarray was initially used to help annotate a highly unique genome that contained few genes of known function and was dominated by coding sequences with no database homologs whatsoever (Wilson et al. 2005). Delays in sequencing caused by the highly repetitive nature of the EhV-86 genome (three families of repeat elements can be found in the genome [Allen et al. 2006c]) restricted the design to probes for a mere 425 of the 472 predicted genes annotated on the final EhV-86 genome. The lack of host genomic information (a dozen or so genes from *Emiliania huxleyi* were known at the time) and the relatively small number of probes required, in tandem with the local availability of an Affymetrix glass slide laser scanner, suggested that a spotted microarray would be the preferred system. Oligonucleotide probes (75mers) were designed one by one using Oligo 6 (a process that took almost a month to complete), synthesized by the commercial company Oligator, and printed using a classic contact pin printing method at the Scottish Centre for Genomic Technology and Informatics (SCGTi). Each probe was printed in triplicate in a 4 × 4 metagrid (each subgrid containing 12 × 13 features) design incorporating 2496 features in total and covering a space of approximately 18 × 20 mm. Once printed at SCGTi, microarrays were supplied to the researchers, who then had total control over experimental design, sample labeling, microarray hybridization, and data analysis. Despite the incomplete coverage of the EhV-86 genome, the first-generation coccolithovirus microarray serves as an excellent example of the versatility of this molecular tool.

Initially, genome annotation of the then-completed EhV-86 genome was aided by using the microarray to confirm the presence of transcripts of the predicted genes (Wilson et al. 2005). A severe lack of database matches due to the uniqueness of the genome created a large amount of uncertainty about what could be annotated as an ORF, coding sequence (CDS), or gene. With regard to genome annotation, not all open reading frames are coding sequences, and not all coding sequences are genes. The results from the direct labeling of total RNA using anchored oligo dT primers (exploiting the poly adenylated tail of mRNA) from cells infected with EhV-86 were used to aid the annotation of the 472 predicted CDSs, since the presence of a transcript is strong evidence that an ORF is actually a transcribed CDS. (For in-depth labeling methodology, see supplementary material of Wilson et al. [2005].) Of course, identification and characterization of the protein products is required to reclassify CDSs as genes (this has recently been done with a proteomic approach and has confirmed 28 genes), but the initial identification of 472 CDSs on a predominantly unknown and unique genome was a huge step forward in the molecular characterization of the coccolithoviruses.

The approach of directly labeling total RNA used in this type of experiment is hugely popular due to its ease and relative simplicity. However, it is fairly RNA intensive (at least 25 µg of total RNA was required), which is not always ideal when studying host–virus systems. In particular, during the early stages of infection when the ratio of virus to host RNA message is relatively low, it is often difficult to detect virus message. This problem is further exacerbated in marine systems which, in general, tend not to provide adequate biomass from manageable culture volumes. To combat this, we tried a range of protocols which would allow us to finely map the transcriptional profile of EhV-86 during the first 4 h of infection. Performing a mRNA purification from the total RNA increased the sensitivity of detection for the handful of host probes on the microarray by approximately 10-fold, yet failed to detect viral transcripts. An indirect labeling method, whereby an aminoallyl nucleotide was incorporated into the cDNA which was later cross-linked to a cyanine dye, further increased sensitivity approximately 1.5-fold. It quickly became apparent that a labeling method capable of increasing sensitivity by many more orders of magnitude would be required to detect virus transcripts during the first few hours of infection. To this end, a derivative of the Eberwine (1996) method was used. The Eberwine method was originally devel-

oped to study single-cell systems and has been phenomenally popular with microarray users. Linear amplification of mRNA message is achieved by creating cDNA using a primer containing the T7 promoter site directly adjacent to the anchored oligo dT primer. Production of cyanine-labeled cRNA using T7 polymerase can amplify the original message by up to 1000-fold. We used a commercially available derivative of this system (Roche Applied Biosciences), which boasts amplification of more than 100,000-fold by using an initial PCR-based step. Briefly, first-strand cDNA synthesis is performed using a primer (oligo dT-T7-TAS) containing a random sequence with no significant homology to any sequence in public databases (Target Amplification Sequence, TAS), a T7 promoter, and oligo dT sequence. After second-strand synthesis using a random primer coupled to TAS, a double-stranded cDNA product is made that can be amplified in a PCR reaction using TAS primers. So long as the PCR reaction is kept in the exponential phase, message can be amplified in a quantitative manner. The resulting amplified cDNA can then be transcribed into cRNA by a linear amplification step using T7 polymerase. Of course, using a PCR-based amplification strategy can create a plethora of downstream issues with data analysis and interpretation, but as long as researchers are aware of this from the beginning, most can easily be avoided. It is crucial to keep the PCR amplification step in the exponential phase; that way, transcripts should be evenly represented.

In the early stages of infection, the number of infected cells is low, but this increases as the experiment progresses; hence the transcriptional profile becomes "blurred" with infected cells at different stages of the infection process. This blurring of the transcriptional profile is avoided in well-studied mammalian systems by the addition of inhibitors of RNA polymerase, protein synthesis, and DNA polymerase, which allow transcripts to be assigned into classes referred to as immediate early, early, mid, and late genes.

We experimented with inhibitors such as phosphonoacetic acid (inhibitor of DNA replication) and cyclohexamide (protein synthesis) but found they failed to give reliable inhibition in our seawater-based algal culturing system. Therefore, in the case of this experimental design, we decided on a simple "when are you on?" question, which is relatively simple to address using bioinformatics methods. We did this at 0, 1, 2, and 4 h post-infection (hpi), and this allowed us to group CDSs into six generic groups (expressed 1, 2, 4 hpi; not expressed, not tested, and unconfirmed [for ambiguous results]) (Allen et al. 2006a). From these generic groups, we were able to distinguish between two major transcriptional phases during viral infection: one phase dominated by genes associated with a specific promoter and localized to a specific section of the genome and a second phase in which the remainder of the genes are transcribed.

The third use of the first-generation array was to study the genomic content of all the coccolithoviruses in our current virus collection (Allen et al. 2007). Direct labeling of genomic DNA was performed using random primers and cyanine-labeled dCTP. Single-channel hybridizations were used, with each genome hybridized to a single array. Pooling of samples to label for the alternative channel was a viable option that was discussed extensively before undertaking the experiment, but since the array was designed specifically for EhV-86, it was felt that the usefulness of the additional data provided from the additional control channel did not justify the added expense of labeling twice as many samples. In perhaps the simplest of all microarray experiments to analyze, and taking the microarray back to its Southern blot roots, an intensity value cutoff was chosen for each array whereby each spot could be considered on or off. This initially appeared to be a risky strategy, and we were worried that some genes would be mislabeled as present or absent/variable. However, the distribution of spot intensities was surprisingly reproducible between strains, and the list of probes bordering the cutoff boundary (i.e., the ambiguous spots) was found to be nearly identical on each microarray—many were even found on the control EhV-86 array, suggesting that poor labeling could be caused by some form of secondary structure in the surrounding genomic regions. An advantage to using a small microarray with fewer than 1600 features is that each can be assessed by eye relatively easily, something that is difficult to achieve with higher-density arrays. This simple approach led to the discovery that at least 70 genes of the 425 that we tested are absent or sufficiently variable in one or more of our dozen or so coccolithovirus strains (Allen et al. 2007).

Therefore, the first-generation coccolithovirus microarray proved to be a robust, useful, and versatile tool that served us well and produced and contributed to eight publications over a 3-year period (Allen and Wilson 2006).

The high degree of control we achieved by developing and modifying our own labeling techniques to answer questions specific to our work made the decision easy when it came to developing the second-generation microarray. The investment, expertise, and past success in developing and optimizing protocols for spotted microarrays suggested we should continue their use. The second-generation microarray was developed to provide greater coverage of the EhV-86 genome and to begin the task of studying the host response to infection both under laboratory conditions and in the natural environment. To this end, ESTs from *Emiliania huxleyi* were sequenced, and the partial sequence (>80%) of a second coccolithovirus, EhV-163, was obtained (Allen et al. 2006b; Kegel et al. 2007). In parallel to this work, the *E. huxleyi* genome sequencing project (led by Betsy Read) was underway; probes from this project for the ESTs with BLAST scores suggesting reasonable similarity to sequences of known function in the GenBank database were also included. In total, more than 4000 oligonucleotides were required to represent every annotated EhV-86 CDS, all unannotated ORFs >100 bp in EhV-86, the 2000+ *E. huxleyi* ESTs, all the genes on the *E. huxleyi* chloroplast and mitochondrial genomes, and for the additional and highly variable EhV-163 genes (Allen et al. 2006c). Due to the higher number of oligonucleotides required for this array and the substantial investment required in their synthesis, we chose to allow the oligonucleotide manufacturer

to design them to provide insurance should they not work well. Operon was chosen for this task, as they had both the necessary track record in oligo design and, importantly, a close working relationship with our chosen microarray printing facility. Whereas pin printing proved to be reliable and robust enough for our relatively small 2496-feature first-generation array, substantially more features were needed on this second-generation microarray. Fortunately, we have been able to take advantage of piezoelectric printing at the Liverpool Microarray Facility node of the NERC Molecular Genetics Facilities. The new second-generation microarray is now based on a 7 × 5 metagrid, with each subgrid composed of 12 × 52 features, with a total of 21,840 features. Each probe is printed five times, and the printed area is approximately 22 × 60 mm. In addition, there has also been significant investment in a tiling microarray for *E. huxleyi*; it is under construction using the Nimblegen system and will be available to the community shortly.

Other aquatic virus microarrays: White spot syndrome virus—White spot syndrome virus (WSSV) is a commercially relevant viral pathogen of the cultured shrimp (*Penaeus* sp.). Since its discovery in Japan in 1993, it has spread to shrimp farming regions throughout Asia, the Americas, Europe, and Australasia (Dhar et al. 2003). As such, it has been intensively studied over past decade by a wide range of geographically distinct and independent research groups and is one of the best studied aquatic virology systems. The inevitable consequence of this intensive yet fractured study is that a variety of WWSV microarrays have been developed in tandem and independently.

Dhar et al. (2003; California, USA), developed a glass slide–based microarray consisting of 100 probes primarily derived from the PCR amplification of EST clones from infected host cells. These researchers used a direct labeling method and hybridized fluorescently labeled first-strand cDNA to identify how shrimp genes responded to viral infection. Khadijah et al. (2003; Singapore) developed a glass slide microarray consisting of approximately 3000 amplified PCR fragments from a clone library created following the restriction digestion of purified WSSV genomic DNA. These researchers estimated that this allowed complete coverage of the WSSV genome and used the array to identify latency-related WSSV genes through the T7-based Eberwine amplification method. Liu et al. (2005) and Tsai et al. (2004) (Taiwan) created a glass slide microarray from PCR products (200–600 bp in size) using specific primers representing 532 predicted ORFs (encoding potential proteins >60 amino acids in size) of WSSV. These researchers used a direct labeling approach to create fluorescently labeled first-strand cDNA to identify immediate early genes in cyclohexamide-treated shrimp and to create a temporal profile for the expression of WSSV genes during infection. Marks et al. (2005; The Netherlands) created a glass slide microarray from PCR products (300–1000 bp in size) for 158 of the 184 annotated WSSV CDSs (encompassing two different WSSV strains). For the larger CDSs, additional probes were generated to improve coverage. PCR products were generated by using either universal primers with suitable clones from the library used to sequence the WSSV genome or specific primers with WSSV genomic DNA as template for a total of 274 probes. A postlabeling approach (initially incorporating an aminoallyl nucleotide) was used to generate fluorescently labeled cDNA. These researchers used their microarray to create a temporal expression pattern for WSSV genes. Lan et al. (2006; China) created a nylon membrane–based microarray using PCR products generated from 259 specific primer pairs (400–1000 bp in size) covering 151 of the 180 CDSs of WSSV. Using ^{32}P radiolabeled cDNA, these researchers generated a temporal transcription profile of WSSV genes during infection. In addition to these microarrays, which contain various numbers of WSSV-specific probes, many research groups have developed "shrimp probe only" microarrays to study the effects of infection by WSSV. Because their interests lie with determining how the host responds and they are not necessarily interested in what the virus is doing per se, these arrays contain only host (shrimp) probes. This approach has been successfully used by Robalino et al. (2007; South Carolina, USA) who developed a glass slide microarray using 2,469 PCR-amplified products from a shrimp EST library. Postlabeled cDNA was generated from four types of shrimp tissue to study the immune response at the transcriptional level.

Discussion

Microarrays are a very powerful and versatile tool. The platform chosen depends on a variety of reasons, technical, economic, and sometimes historical. The authors of this chapter are a great example of the diversity in approaches to using microarrays for the study of aquatic viruses: whereas D. Lindell favors the Affymetrix approach, M. J. Allen favors the spotted oligo approach. Neither author is more right or more wrong than the other: both systems are fit for their current purposes. Like evolution, when developing a microarray system you can only work with what you've got. Indeed, other, less commonly used, microarray platforms exist (such as those developed by Applied Biosystems, Eppendorf, GE Healthcare, Illumina, and Phalanx), and these may suit your needs better than the systems described in this article. Regardless of the final platform choice, most of the issues described here will be of direct relevance to your experimental requirements. This also applies to experimental design and analysis. Many options exist, and these should be considered in light of the questions you wish to address and the resources, money, and skills available to you.

In this article, we have tried to describe the technology, techniques, and rationale behind microarray experiments, with a focus on the specific issues of virus-associated systems. As the reader can see, microarrays should not be undertaken on a whim, but with proper planning and consideration they can prove to be an excellent weapon in the virologist's armory. However, it is important to note that there is no generic method that is perfect for every system. Thus, in closing, we have come to the general conclusions that if a researcher is given a blank canvas with no prior preconceptions or limita-

tions then Affymetrix may be the platform of choice for systems when a large number of experiments will be carried out, Agilent may be the platform of choice when high design flexibility is desired and few samples will be investigated, and spotted arrays may be the platform of choice for large-volume experiments when high design flexibility is required. Yet access to existing infrastructure or even a researcher's current knowledge and thinking can have a profound impact on the route and choices taken. The key is to be prepared from the outset. Although they are no easy undertaking, microarrays have the potential to drive your research in exciting and wonderful directions. It is the destination that matters most (i.e., publishable data of high quality), not the route taken to get there.

References

Allen, M. J., T. Forster, D. C. Schroeder, M. Hall, D. Roy, P. Ghazal, and W. H. Wilson. 2006a. Locus-specific gene expression pattern suggests a unique propagation strategy for a giant algal virus. J. Virol. 80:7699-7705.

———, D. C. Schroeder, A. Donkin, K. J. Crawfurd, and W. H. Wilson. 2006b. Genome comparison of two Coccolithoviruses. Virol. J. 3:15.

———, ———, and W. H. Wilson. 2006c. Preliminary characterisation of repeat families in the genome of EhV-86, a giant algal virus that infects the marine microalga *Emiliania huxleyi*. Arch. Virol. 151:525-535.

———, and W. H. Wilson. 2006. The coccolithovirus microarray: An array of uses. Brief Funct. Genomic Proteomic 5:273-279.

———, J. Martinez-Martinez, D. C. Schroeder, P. J. Somerfield, and W. H. Wilson. 2007. Use of microarrays to assess viral diversity: From genotype to phenotype. Environ. Microbiol. 9:971-982.

———, and W. H. Wilson. 2008. Aquatic virus diversity accessed through omic techniques: A route map to function. Curr. Opin. Microbiol. 11:226-232.

Alwine, J. C., D. J. Kemp, and G. R. Stark. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U. S. A. 74:5350-5354.

Baldi, P., and A. D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics 17:509-519.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Royal Stat. Soc. Ser. B 57:289-300.

Bier, F. F., M. Von Nickisch-Rosenegk, E. Ehrentreich-Forster, E. Reiss, J. Henkel, R. Strehlow, and D. Andresen. 2008. DNA microarrays. Adv. Biochem. Eng. Biotechnol. 109:433-453.

Bittner, M., and others. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536-540.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193.

Brazma, A., and others. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. 29:365-371.

Choe, S. E., M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol. 6:R16.

Chou, H. H., A. P. Hsia, D. L. Mooney, and P. S. Schnable. 2004. Picky: Oligo microarray design for large genomes. Bioinformatics 20:2893-2902.

Chung, W. H., S. K. Rhee, X. F. Wan, J. W. Bae, Z. X. Quan, and Y. H. Park. 2005. Design of long oligonucleotide probes for functional gene detection in a microbial community. Bioinformatics [doi:10.1093/bioinformatics/bti673].

Cope, L. M., R. A. Irizarry, H. A. Jaffee, Z. J. Wu, and T. P. Speed. 2004. A benchmark for affymetrix GeneChip expression measures. Bioinformatics 20:323-331.

Coppee, J. Y. 2008. Do DNA microarrays have their future behind them? Microbes Infect. 10:1067-1071.

Dabney, A. R., and J. D. Storey. 2006. A reanalysis of a published Affymetrix GeneChip control dataset. Genome Biol. 7:401.

DeRisi, J., and others. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. 14:457-460.

Dhar, A. K., A. Dettori, M. M. Roux, K. R. Klimpel, and B. Read. 2003. Identification of differentially expressed genes in shrimp (*Penaeus stylirostris*) infected with White spot syndrome virus by cDNA microarrays. Arch. Virol. 148:2381-2396.

Dudoit, S., J. Popper Shaffer, and J. C. Boldrick. 2003. Multiple hypothesis testing in microarray experiments. Stat. Sci. 18:71-103.

Dufva, M. 2009a. Fabrication of DNA microarray. Methods Mol. Biol. 529:63-79.

———. 2009b. Introduction to microarray technology. Methods Mol. Biol. 529:1-22.

Dunn, J. 1974. Well separated clusters and optimal fuzzy partitions. J. Cybernetics 4:95-104.

Eberwine, J. 1996. Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. Biotechniques 20:584-591.

Futschik, M., and T. Crompton. 2004. Model selection and efficiency testing for normalization of cDNA microarray data. Genome Biol. 5:R60.

Gentleman, R. C., and others. 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 5:R80.

Gentry, T. J., G. S. Wickham, C. W. Schadt, Z. He, and J. Zhou. 2006. Microarray applications in microbial ecology research. Microb. Ecol. 52:159-175.

Gibbons, F. D., and F. P. Roth. 2002. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 12:1574-1581.

Gresham, D., M. J. Dunham, and D. Botstein. 2008. Comparing whole genomes using DNA microarrays. Nat. Rev. Genet. 9:291-302.

Hancock, D., and others. 2005. maxdLoad2 and maxdBrowse: Standards-compliant tools for microarray experimental annotation, data management and dissemination. BMC Bioinformatics 6:264.

He, Z., L. Wu, M. W. Fields, and J. Zhou. 2005. Use of microarrays with different probe sizes for monitoring gene expression. Appl. Environ. Microbiol. 71:5154-5162.

Herold, K. E., and A. Rasooly. 2003. Oligo Design: A computer program for development of probes for oligonucleotide microarrays. Biotechniques 35:1216-1221.

Hoffman, R., T. Seidl, and M. Dugas. 2002. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray analysis. Genome Biol. 3:research0033.1-research0033.11.

Huber, W., R. A. Irizarry, and R. Gentleman. 2005. Preprocessing overview, p. 3-12. *In* R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit [eds.], Bioinformatics and computational biology solutions using R and Bioconductor. Springer.

Hughes, A. L., and R. Friedman. 2005. Poxvirus genome evolution by gene gain and loss. Mol. Phylogenet. Evol. 35:186-195.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.

———, L. M. Cope, and Z. J. Wu. 2006. Feature-level exploration of a published Affymetrix GeneChip control dataset. Genome Biol. 7:404.

Jain, A., and R. Dubes. 1988. Algorithms for clustering data. Prentice Hall.

Jain, N., J. Thatte, T. Braciale, K. Ley, M. O'connell, and J. K. Lee. 2003. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics 19:1945-1951.

Karaman, M. W., S. Groshen, C. C. Lee, B. L. Pike, and J. G. Hacia. 2005. Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. Nucleic Acids Res. 33:e33.

Karsten, S. L., and D. H. Geshwind. 2002. Gene expression analysis using cDNA microarrays. Curr. Protoc. Neurosci. Chapter 4:Unit 4.28.

Kegel, J., M. J. Allen, K. Metfies, W. H. Wilson, D. Wolf-Gladrow, and K. Valentin. 2007. Pilot study of an EST approach of the coccolithophorid *Emiliania huxleyi* during a virus infection. Gene 406:209-216.

Kerr, K. F., K. A. Serikawa, C. Wei, M. A. Peters, and R. E. Bumgarner. 2007. What is the best reference RNA? And other questions regarding the design and analysis of two-color microarray experiments. OMICS 11:152-165.

Kerr, M. K. 2003. Design considerations for efficient and effective microarray studies. Biometrics 59:822-828.

———, and G. A. Churchill. 2001. Experimental design for gene expression microarrays. Biostatistics 2:183-201.

Khadijah, S., S. Y. Neo, M. S. Hossain, L. D. Miller, S. Mathavan, and J. Kwang. 2003. Identification of white spot syndrome virus latency-related genes in specific-pathogen-free shrimps by use of a microarray. J. Virol. 77:10162-10167.

Lan, Y., X. Xu, F. Yang, and X. Zhang. 2006. Transcriptional profile of shrimp white spot syndrome virus (WSSV) genes with DNA microarray. Arch. Virol. 151:1723-1733.

Letowski, J., R. Brousseau, and L. Masson. 2004. Designing better probes: Effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J. Microbiol. Methods 57:269-278.

Levine, E., and E. Domany. 2001. Resampling method for unsupervised estimation of cluster validity. Neural Comput. 13:2573-2593.

Li, W., J. Huang, M. Fan, and S. Wang. 2002. MProbe: Computer aided probe design for oligonucleotide microarrays. Appl. Bioinformatics 1:163-166.

Li, X., Z. He, and J. Zhou. 2005. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res. 33:6114-6123.

Lindell, D., and A. F. Post. 2001. Ecological aspects of ntcA gene expression and its use as an indicator of the nitrogen status of marine *Synechococcus* spp. Appl. Environ. Microbiol. 67:3340-3349.

———, and others. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449:83-86.

Liu, W. J., Y. S. Chang, C. H. Wang, G. H. Kou, and C. F. Lo. 2005. Microarray and RT-PCR screening for white spot syndrome virus immediate-early genes in cyclohexamide-treated shrimp. Virology 334:327-341.

Liu, X. S. 2007. Getting started in tiling microarray analysis. PLoS Comput. Biol. 3:1842-1844.

Marks, H., O. Vorst, A. M. Van Houwelingen, M. C. Van Hulten, and J. M. Vlak. 2005. Gene-expression profiling of White spot syndrome virus in vivo. J. Gen. Virol. 86 Pt 7:2081-2100.

Martiny, A. C., M. L. Coleman, and S. W. Chisholm. 2006. Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. Proc. Natl. Acad. Sci. U. S. A. 103:12552-12557.

Millard, A. D., and B. Tiwari. 2009. Oligonucleotide microarrays for bacteriophage expression studies. Methods Mol. Biol. 502:193-226.

———, K. Zwirglmaier, M. J. Downey, N. H. Mann, and D. J. Scanlan. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. Environ Microbiol. Jun 7 [Epub ahead of print].

Monier, A., A. Pagarete, C. de Vargas, M. J. Allen, B. Read, J. M. Claverie, and H. Ogata. 2009. Horizontal gene transfer of

an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. 19:1441-1449.

Moreira, D., and C. Brochier-Armanet. 2008. Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. BMC Evol. Biol. 8:12.

Morrison, N., and others. 2006. Annotation of environmental OMICS data: Application to the transcriptomics domain. Omics 10:172-178.

Nielsen, H. B., R. Wernersson, and S. Knudsen. 2003. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. Nucleic Acids Res. 31:3491-3496.

Nordberg, E. K. 2005. YODA: Selecting signature oligonucleotides. Bioinformatics 21:1365-1370.

Page, G. P., S. O. Zakharkin, K. Kim, T. Mehta, L. Chen, and K. Zhang. 2007. Microarray analysis. Methods Mol. Biol. 404:409-430.

Pirooznia, M., P. Gong, J. Y. Yang, M. Q. Yang, E. J. Perkins, and Y. Deng. 2008. ILOOP—A web application for two-channel microarray interwoven loop design. BMC Genomics (suppl. 2) 9:S11.

Reymond, N., H. Charles, L. Duret, F. Calevro, G. Beslon, and J. M. Fayard. 2004. ROSO: Optimizing oligonucleotide probes for microarrays. Bioinformatics 20:271-273.

Rich, V. I., K. Konstantinidis, and E. F. Delong. 2008. Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. Environ Microbiol 10:506-521.

Robalino, J., and others. 2007. Insights into the immune transcriptome of the shrimp *Litopenaeus vannamei*: Tissue-specific expression profiles and transcriptomic responses to immune challenge. Physiol. Genomics 29:44-56.

Rocap, G., and others. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

Rouillard, J. M., M. Zuker, and E. Gulari. 2003. OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. 31:3057-3062.

Schretter, C., and M. C. Milinkovitch. 2006. OLIGOFAKTORY: A visual tool for interactive oligonucleotide design. Bioinformatics 22:115-116.

Simon, R. 2008. Microarray-based expression profiling and informatics. Curr. Opin. Biotechnol. 19:26-29.

Sipe, C. W., and M. S. Saha. 2007. The use of microarray technology in nonmammalian vertebrate systems. Methods Mol. Biol. 382:1-16.

Smith, B., and others. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25:1251-1255.

Smyth, G. K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Applic. Genet. Mol. Biol. 3:Article 3.

———, J. Michaud, and H. S. Scott. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21:2067-2075.

Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503-517.

Steglich, C., M. Futschik, T. Rector, R. Steen, and S. W. Chisholm. 2006. Genome-wide analysis of light sensing in Prochlorococcus. J. Bacteriol. 188:7796-7806.

———, M. E. Futschik, D. Lindell, B. Voss, S. W. Chisholm, and W. R. Hess. 2008. The challenge of regulation in a minimal photoautotroph: Non-coding RNAs in Prochlorococcus. PLoS Genet. 4:e1000173.

Stenberg, J., M. Nilsson, and U. Landegren. 2005. ProbeMaker: An extensible framework for design of sets of oligonucleotide probes. BMC Bioinformatics 6:229.

Sullivan, M. B., M. L. Coleman, P. Weigele, F. Rohwer, and S. W. Chisholm. 2005. Three Prochlorococcus cyanophage genomes: Signature features and ecological interpretations. PLoS Biol. 3:790-806.

Tolonen, A. C., and others. 2006. Global gene expression of Prochlorococcus ecotypes in response to changes in nitrogen availability. Mol. Sys. Biol. 2:53.

Tsai, J. M., H. C. Wang, J. H. Leu, H. H. Hsiao, A. H. Wang, G. H. Kou, and C. F. Lo. 2004. Genomic and proteomic analysis of thirty-nine structural proteins of shrimp white spot syndrome virus. J. Virol. 78:11360-11370.

Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. 98:5116-5121.

Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Microb. Ecol. 17:1636-1647.

Wang, D., L. Coscoy, M. Zylberberg, P. C. Avila, H. A. Boushey, D. Ganem, and J. L. Derisi. 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. 99:15687-15692.

———, and others. 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1:E2.

Wang, X., and B. Seed. 2003. Selection of oligonucleotide probes for protein coding sequences. Bioinformatics 19:796-802.

Wilson, W. H., and others. 2005. Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. Science 309:1090-1092.

———, J. L. Van Etten, and M. J. Allen. 2009. The Phycodnaviridae: The story of how tiny giants rule the world. Curr. Top. Microbiol. Immunol. 328:1-42.

Wit, E., and J. McClure. 2004. Statistics for Microarrays. Wiley.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30:e15.

Zinser, E. R., and others. 2009. Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus. PLoS ONE 4:e5135.