APPLICATIONS OF DNA MICROARRAYS IN BIOLOGY

Roland B. Stoughton

GHC Technologies, Incorporated, La Jolla, California 92037;
email: roland_stoughton@ghctechnologies.com

Key Words expression profiling, genomics, genotype, transcription

■ **Abstract** DNA microarrays have enabled biology researchers to conduct large-scale quantitative experiments. This capacity has produced qualitative changes in the breadth of hypotheses that can be explored. In what has become the dominant mode of use, changes in the transcription rate of nearly all the genes in a genome, taking place in a particular tissue or cell type, can be measured in disease states, during development, and in response to intentional experimental perturbations, such as gene disruptions and drug treatments. The response patterns have helped illuminate mechanisms of disease and identify disease subphenotypes, predict disease progression, assign function to previously unannotated genes, group genes into functional pathways, and predict activities of new compounds. Directed at the genome sequence itself, microarrays have been used to identify novel genes, binding sites of transcription factors, changes in DNA copy number, and variations from a baseline sequence, such as in emerging strains of pathogens or complex mutations in disease-causing human genes. They also serve as a general demultiplexing tool to sort spatially the sequence-tagged products of highly parallel reactions performed in solution. A brief review of microarray platform technology options, and of the process steps involved in complete experiment workflows, is included.

CONTENTS

INTRODUCTION

The closest ancestors to DNA microarrays were the spotted arrays on nylon of
the 1980s and early 1990s (1). These held bacterial colonies carrying different
cosmid genomic inserts and later held preidentified cDNAs, with spacings down
to ∼2 mm. Hybridization of complex RNA samples to known cDNAs on these
arrays was exactly analogous to the current "expression profiling" use mode of
microarrays. The expression profiling use mode has become the dominant one due
to the wealth of functional information provided about the biological sample being
analyzed. Because full or partial transcript sequences are now available for nearly
all genes in the most commonly studied organisms, the new high-density arrays
can provide genome-wide response profiles for the changes in transcription rate
associated with drug treatments, disease states, phenotypic differences, and mu-
tations. The results of keyword searches for 1990 through 2003 show the growth
in PubMed publications significantly involving DNA microarrays (Figure 1). The
catchy concept of "DNA chips" as a confluence of infotech and biotech has con-
tributed to the sudden fame of DNA microarrays, but this fame is deserved on the
basis of rapidly widening scientific usage and results.

The term expression profiling also can apply to measurements of protein con-
centrations, which arguably are more directly related to cell function than are the
mRNA messages. But it has proven technically more difficult to achieve the equiv-
alent of genome-wide profiling of proteins (2, 3). Besides their inherent fragility
in the most accessible human sample types, such as blood cells and serum, the
range of abundances between the least abundant functional protein species and
the most abundant is several orders of magnitude greater than in the case of mR-
NAs. Technologically, proteomic profiling lags several years behind in the depth
of coverage currently achievable with DNA microarrays. Also, it is increasingly
appreciated that mRNA response profiles are not just surrogates for the corre-
sponding protein levels; rather they show the adjustment of the cell to individual
protein functional disruptions (4), even if protein levels are unchanged. Conse-
quently, the mRNA window into cell function is and should remain a very powerful
one.

Although mRNA expression profiling is the dominant use mode for DNA mi-
croarrays, several others, found by analysis of genomic DNA sequence, are notable
and are listed in Table 1. Microarrays are inherently a means of spatially sorting
molecular species so that their concentrations can be independently estimated;
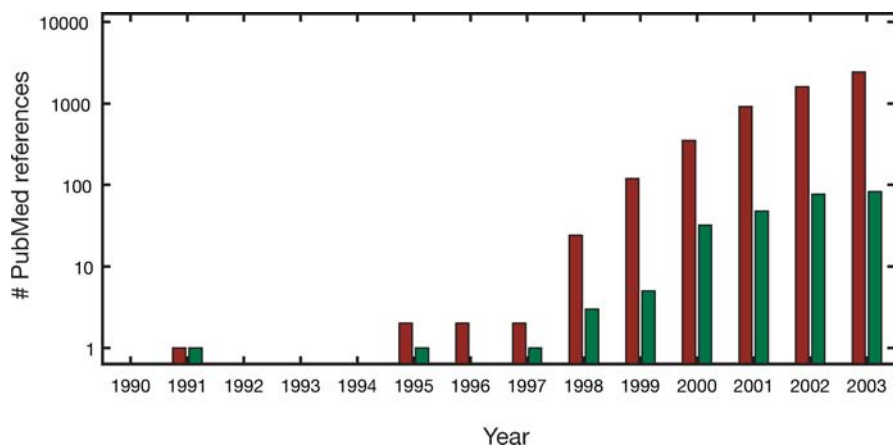
Figure 1   PubMed references found with keyword searches of the title and abstract fields. Red: (DNA microarray) or (DNA chip) or (expression array). Green: (microarray or chip) and (sequencing or resequencing or genotype). After 1996 the first search was modified by dropping the "DNA" in "DNA microarray." Before 1996 the second group in the second search was expanded with "or genome" to catch two early references. The expression profiling application quickly became dominant.

potential applications therefore include analysis of any complex reaction product. For example, large numbers of single nucleotide polymorphisms (SNPs) can be determined in parallel, using microarrays as the final stage in the process (5). Then genetically bar-coded strains of yeast are competitively grown, and their final concentrations read out by hybridization of the bar codes to an array (6).

This is intended more as a "why-to" than a "how-to" review. But because it is still essential to understand the technology in order to design experiments and interpret data effectively, the next section discusses the different types of microarray platforms and associated process steps. The applications section describes examples of expression profiling and is followed by a section containing examples of genomic DNA analysis. Finally, trends in usage and potential future applications are discussed.

MICROARRAY TECHNOLOGY PLATFORMS AND PROCESS

A complete generic process diagram for microarray experiments is shown in Figure 2. Setting up a robust complete microarray experiment process by assembling individual components is a challenge. The larger vendor corporations, such as Affymetrix, Agilent Technologies, and the GE Healthcare division of General Electric, provide suites of components, reagents, and services. For example,

TABLE 1 Modes of use for DNA microarrays

Purpose	Target sample	Multiplexed reactions	Demultiplexing probes on array
Expression profiling	mRNA or totRNA from relevant cell cultures or tissues	Amplification of all mRNAs via some combination of RT/PCR/IVT	Single- or double-stranded DNA complementary to target transcripts
Pathogen detection and characterization	Genomic DNA from microbes	Random-primed PCR, or PCR with selected primer pairs for certain target regions	Sequences complementary to preselected identification sites
Genotyping	Genomic DNA from humans or animals	Ligation/extension for particular SNP regions, and amplification	Sequences complementary to expected products
Resequencing	Genomic DNA	Amplification of selected regions	Sequences complementary to each sliding N-mer window along a baseline sequence and also to the three possible mutations at the central position
Find protein-DNA interactions	Genomic DNA	Enrichment based on transcription factor binding	Sequences complementary to intergenic regions

Affymetrix supplies predesigned arrays for human, mouse, rat, and some microbial genomes along with the reagents and protocols for performing the required amplification, hybridization, washing, and scanning steps. They also provide a public Web site with up-to-date annotations for the transcript sequences for which their array probes are designed in addition to a suite of bioinformatic tools for analyzing the data.

Microarray Manufacturing Technologies

Technical approaches to manufacture the arrays themselves are described concisely in chapters in Baldi & Hatfield (7) and in Stekel (8). See also the review by Heller (9). Figure 3 compares feature densities of the progenitor nylon spotted arrays and two modern methods. Tables 2a and 2b list the main vendors for the available array types and some of the advantages and limitations associated with these products. Table 2a covers in situ synthesis methods, while Table 2b covers methods based on deposition of presynthsized oligos and cDNAs. The light-directed synthesis
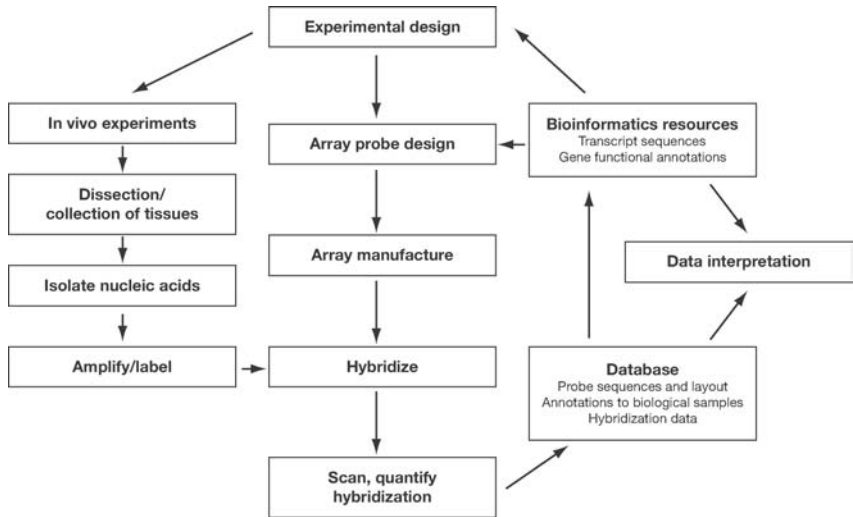
Figure 2 Process flow for microarray experiments. Note that we have used the word "probe" to refer to the reporter sequence placed at a particular position on the microarray because it interrogates the sample for the presence of its reverse complement and also because the microarray market leader, Affymetrix, has adopted this definition. Historically probe has referred instead to the biological sample.
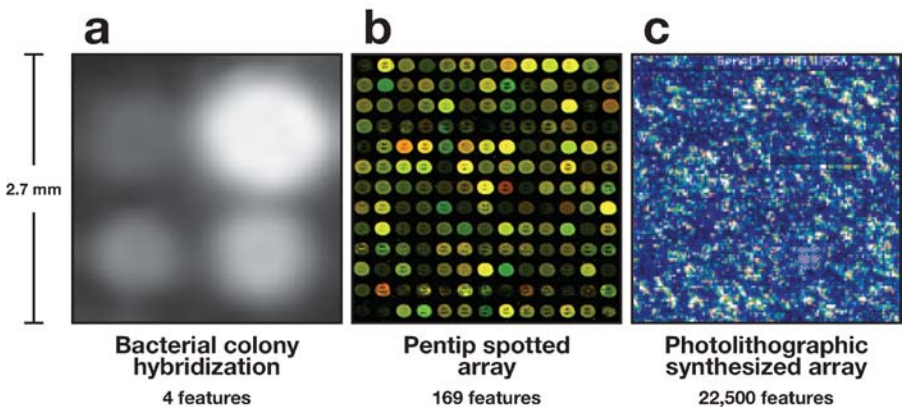

Figure 3 Feature density of representative microarrays. Each image shows a 2.7 mm square subregion. (*a*) Bacterial colony spots on nylon from the 1980s. (*b*) Ink-jet in situ synthesized 60-mer oligo spots on glass. (*c*) Affymetrix human gene array with 18-μ features containing 25-mer oligos. Affymetrix chips now are available with 11-μ features.

TABLE 2a Microarray alternatives that use in situ synthesis[a]

Process	Vendors	Substrate	Density	Maturity	Advantages	Limitations
Light-directed synthesis	—	—	—	—	Potential densities approach wavelength of light	Photo-deprotection stepwise yields limit oligo length
Photolithographic mask for each of the four bases at each layer	Affymetrix	Glass	8200 features (~360 transcripts) per mm²	Very mature	Leader in density and quality control	Manufacturing costs of photolithographic masks
Digitally controlled micromirrors incorporated in chip	NimbleGen Systems	Glass	~1000 features per mm²	Early	Easy to produce small lots of custom arrays	
Digitally controlled micromirrors and microfluidics incorporated in chip	FeBit, Xeotron	Glass/silicon (semiconductor fabrication process)	~1000 features per mm²	Early	Easy to produce small lots of custom arrays	Manufacturing costs of microfluidic structures
Ink-jet printer head deposition of nucleotides layer by layer	Agilent Technologies	Glass	100 features per mm²	Mature	Easy to produce small lots of custom arrays	Feature density limited by feasible droplet sizes
Electrode-directed synthesis	CombiMatrix, Nanogen	Silicon (semiconductor fabrication process)	<100 features per mm²	Early (Nanogen is not pursuing synthesis)	Custom arrays	Fabrication process still expensive

[a]In situ synthesis has the advantage that any sequence can be created at any feature position. However, it is difficult to perform quality control on the products. In this table and Table 2b, the Advantages and Limitations columns list issues to consider and are not meant as a comprehensive evaluation of each technology.

TABLE 2b Microarray alternatives using presynthesized oligos or cDNAs[a]

Process	Vendors	Substrate	Density	Maturity	Advantages	Limitations
Pen tip deposition	—	—	—	—	Low technology investment	Deposition not highly repeatable, droplet size limits density
	Clontech Pharmingen Sigma-Genosys	Nylon or other synthetic polymer	<10 features per mm^2	Mature	Low manufacturing cost per array	Substrate properties or pen tip diameter limit feature density
	Clinical MicroSensors	Printed circuit board	<10 features per mm^2	Early	Potentially low manufacturing cost per array	Insensitive detection, substrate, chemistry difficult to control
	Clontech Pharmingen Harvard Biosciences Mergen MWG Biotech	Glass	<100 features per mm^2	Mature	Potentially very low manufacturing cost per array	Diameter of pen tip and droplet size limit density
Ink-jet deposition	GE Healthcare (CodeLink arrays)	Polyacrylamide gel on glass	100 features per mm^2	Mature	Potentially very low manufacturing cost per array	
Electro-phoretically driven deposition	Nanogen	Silicon (semi-conductor fabrication process)	100 features per mm^2	Early	Fast hybridization, relatively easy to produce custom arrays	Expensive chip fabrication process, electrophoretic process not fully developed

[a]With presynthesis, the DNA sequences can be purified and verified giving greater accuracy of hybridization. However, handling tens of thousands of individual sequence clones prior to and during deposition is difficult.

methods, both photolithographic (10) and digital micromirror-based (11), have the potential to achieve feature sizes not much larger than a wavelength of light. This should enable substantial further reductions in cost and in hybridization volume with consequent reduction in amount of biological sample required. In situ ink-jet synthesis (12) should have a valuable niche for rapid turnaround of custom arrays in small lots, unless it is overtaken by the micromirror technologies. During optimization of the ink-jet synthesis technology (12), it was shown that 60-mers allowed a much better sensitivity-specificity trade than shorter oligos. In light-directed synthesis, failure of photodeprotection at any stage terminates the oligo. The yields per stage in the Affymetrix synthesis process are such that attempts to make 60-mers would result in very few of them running to even half that length (13); Affymetrix settled on 25-mers partly for this reason. Their use of multiple probe pairs to estimate the abundance of each target transcript is driven partly by the need to make up for the performance limitations of 25-mers. Ink-jet synthesis yields are ∼98% per stage with chemical deprotection, as opposed to ∼95% for photodeprotection (13), allowing the ink-jet technology to be optimized with longer oligos and higher stringency hybridization conditions.

Presynthesis of oligos or cDNAs has the important advantage that the sequences eventually placed on the array can be exactly those desired. Pen tip spotting methods (14, 15) will continue to be a relatively low-tech but robust and affordable method for small laboratories to generate their own arrays with a moderate number of features and have the capability of spotting DNA of an unknown sequence. Ink-jet methods also can be used to print presynthesized oligos.

In a true synthesis of microelectronics and molecular biotechnology, Nanogen and CombiMatrix have incorporated electrodes in their arrays to direct synthesis. In Nanogen's case, the primary application of electrodes is to drive hybridization.

Other methods of spatial demultiplexing not listed in Table 1 are worth mention. Universal arrays adopt a fixed set of orthogonal probe sequences and conjugate their complements to problem-specific ligands (5, 16) or adopt a complete set of short N-mers and interpret all samples in terms of the hybridization pattern to these (17). Such approaches avoid the need to redesign the array for each application. The former potentially achieves better control over the hybridization step at the expense of added process complexity. Fiber optic bundles terminated with beads conjugated to specific probes can be used as arrays (18, 19) and are being produced by Illumina primarily for parallel genotyping application. Finally, individual beads conjugated to specific probes can be localized in a flow cytometry configuration and identified via unique color tags (20), as in the system being offered by Luminex.

Probe Design

Most array vendors offer standard array probe sets for some or all of the most common model organisms, such as mouse, rat, yeast, and human, and in this case, the problems of choosing target sequences and designing probes are taken out of

the user's hands. In the case of custom arrays, the customer can contribute different levels of detail to the design: a target gene list, a list of target transcript sequences, the actual probe sequences to be synthesized, or even complete specification of all probe and control sequences and their spatial layout on the array. This choice involves considering the customer's bioinformatic capabilities compared to those of the array vendor and, of course, the price of the vendor services.

Commercial software, such as ArrayDesigner (Premier Biosoft International), TILIA (Linden Biosciences), and free software (21), exist to aid in probe design. Stekel (22) summarizes sequence database resources and probe design principles. The chosen array technology and baseline amplification and hybridization protocols constrain the design as to probe length, a preferred range of melting temperature, and probe distance from the 3′ end in the case of mRNA profiling. Other considerations include avoiding repetitive sequence motifs, avoiding regions that are likely to incur mRNA secondary structure, and avoiding cross hybridization to sequences other than the target sequence. The latter calculation in particular is very computationally demanding when done for extensive probe sets for large genomes. These calculations are hampered by the absence of accurate models for binding energy; published "nearest neighbor" interaction energies result mostly from the dissociation of oligos of length ∼10 nucleotides (nt) in volume solution (23), which can be very different from the behavior of surface-bound and longer oligos owing to the electrostatic effects of the surface (24) and greater propensity for secondary structure. These models also could benefit by factoring in the rough expected abundance of all the molecular species in the tissue of interest. For example, the cross-hybridization pattern in a liver sample will likely be different compared to a brain sample, which has a different set of abundantly expressed genes. Affymetrix probes are designed in pairs: One sequence is the exact complement of the target sequence, and the other differs from the exact complement by typically two mutations near the middle of the probe. Using the difference signal between these two probes approximately cancels nonspecific binding and background contributions. However, it does not cancel cross hybridization from sequences nearly identical to the target because the mismatch probe will almost certainly differ more from the closely related sequence than does the match probe. Probe design for expression profiling increasingly will be splice-form specific as a more complete understanding of splice forms becomes available (25).

Amplification and Labeling

The protocols for isolating and amplifying the desired form of nucleic acid have many variations. In the case of mRNA profiling, amplification can proceed from purified mRNA or from total RNA, although in prokaryotes mRNA purification is problematic because the nucleic acid mostly lacks 3′ polyadenylation (26). Labeling molecules can be incorporated during synthesis of amplification products, or modified nucleotides capable of accepting label can be incorporated during synthesis and the labels coupled to these immediately afterward. Or, the labeling can

take place after hybridization as in the Affymetrix protocol. Signal amplification methods also have been employed to augment sample amplification (27–29).

Nucleic acid amplification can be accomplished through reverse transcription of RNA followed by linear amplification with one or more rounds of in vitro transcription (IVT) (12, 30–32), or via polymerase chain reaction (PCR) (33–36), or a combination of these (12, 37). Amplification can be 3′-biased or full-length, and the decision interacts strongly with the process of probe design. 3′-biased amplification methods take advantage of priming from polyadenylation sequences found in eukaryotic transcripts. Full-length amplification tends to employ random priming of the target molecules, either because poly(A) sequences do not exist as in the case of prokaryotic organisms, or out of a desire to amplify sequences more representative of the entire target length. The final product to be hybridized to the array can be either cDNA or cRNA. Out of these many variations a few principles have emerged. Linear and modest amplification, as well as postsynthetic incorporation of labels, generally are associated with more reproducible data. cDNA:DNA hybridizations tend to suffer less from cross hybridization, even though the actual binding energies tend to be lower than those of cRNA:DNA duplexes. 3′-biased protocols have the advantage that the untranslated sequence regions tend to be more diverse and so allow more sequence-unique and therefore noncross-hybridizing probes to be designed.

Despite one's best efforts, biases enter at all reaction steps in these workflows and can be very sequence specific, so that the final brightness of a given probe is only approximately relatable to the abundance of its corresponding target molecule. Fortunately the most meaningful results usually are contained in the ratio of abundances, referred to a reference biological sample, such as in disease versus normal studies, rather than in the absolute levels. As long as the biases are reproducible, ratios are fairly well preserved; they are not perfectly preserved due to the nonlinearity of the gains created at each stage in the process. Ratio measurements are obtained by comparing two independently hybridized samples, or two samples are independently labeled with different labels and competitively hybridized in the same hybridization reaction. Competitive hybridization is the ultimate matched control method because the variations in probe spot properties, arising from synthesis and local hybridization conditions, usually cancel each other. This permits accurate ratios to be obtained from spotted arrays even when the spots and hybridization fluidics are somewhat irregular. This is an important reason why the relatively low-budget spotting technologies have been so successful. Biases that arise from differences in chemical properties between the two dyes are significant and sequence specific, but the biases can be mostly mitigated by repeating the experiment with the reverse assignment of dye to sample and averaging the results (38, 39).

Hybridization

During hybridization, complementary sequences gradually find each other preferentially over mismatched pairings. The fundamental parameters are time, stringency, concentration, and complexity of the sample, as well as density of available

binding sites. Secondary parameters include the distribution of fragment lengths, steric effects of dye molecules (40), and surface chemistry (41). The optimization of stringency involves choosing conditions in which, for most probes, perfect match duplexes (or at least duplexes with the intended target) have a high occupancy compared to mismatch duplexes. In general, any given target sequence will pair and dissociate many times during the hybridization reaction, staying longer at high-binding-energy well-matched duplexes than in poorly matched duplexes. This annealing is therefore like musical chairs in which specificity increases over time as the reaction approaches equilibrium (42). In fact, the progression with time can be used to distinguish specific from nonspecific binding (43). In general, hybridization times of several hours or more at relatively high stringency are required to achieve the best specificity. A challenge for clinical and some biodefense applications will be to speed up this process substantially. The electric field–induced migration of target fragments toward probes of the Nanogen system increases local concentration and accelerates the accumulation of signal; it is less clear if it accelerates the progression to a certain level of specificity. Likewise, agitation, microfluidic circulation as in the Affymetrix system, surfactants, buffers (44), and acoustic sonication (45) have potential to speed and improve hybridization.

Washing off the unbound sample after hybridization is a crucial step. Stringency must be optimized here as well (46). Chemistry surprises still should be expected, such as the recent discovery of the high sensitivity of some dyes in certain protocols to the presence of atmospheric ozone (47), which seems to occur particularly during the exposure of the drying array surface to the air.

Fluorescent Scanning of Hybridized Arrays

Fluorescent labeling and detection on nonporous arrays have replaced radioactive labeling on membrane arrays in almost all labs because the sensitivity of the former now rivals or exceeds that of ^{33}P (37), and nonradioactive labeling is much easier to handle. Scanning of a fluorescent hybridization signal can be done with CCD imaging using filtered white light illumination, as in the ArrayWoRx from Applied Precision, but now it is more commonly done with laser confocal scanners (48). The laser confocal approach has fundamental geometric advantages that tend to provide better signal-to-background ratios and less photobleaching of the labels. Available options for fluorescent confocal scanners are changing even more rapidly than for arrays. In addition to the Affymetrix-specific scanners supplied by them, leading manufacturers include Agilent Technologies, Axon Instruments, Packard, and Genomic Solutions.

Most devices have lasers and filter sets compatible with common fluorescent label pairs such as Cy3 and Cy5. The leading laser confocal devices all have sensitivities high enough that background fluorescence of the array substrate and unwashed reagents, not signal faintness, generally limits performance. In other words, these scanners are as good as they need to be given the current limitations of other steps in the process. New options for brighter individual labeling units, such as quantum dots (49) and plasmon resonance particles (50, 51), may finally allow

single-molecule detection efficiency, further easing requirements on amplification and on the input biological sample amount.

Scanners come with, or the manufacturer will recommend, image processing software to reduce the raw images to spot intensities. This step is itself a challenging and complex subject (52, 53). Because experiments often are designed to detect a small number of changes among a large number of target genes, outlier errors in the image must be carefully recognized and flagged. Any compromise in the accuracy at this stage cannot be recovered downstream. The output should include pixel intensity statistics that can be used as input to measurement uncertainty estimates.

Plasmon surface resonance detection of molecular binding is potentially a way to scan microarrays without using any label (54), although this has not yet matured to any commercial systems.

Performance Metrics for the Entire System

Ultimately it is the measurement performance of the system as a whole that is important. Various metrics have been proposed to assess this performance. One key test is to spike in known amounts of a positive control into a typical complex sample. Because ratios of abundances between two samples are often most meaningful, this should be done with different spike-in amounts in two samples, comparing the observed ratios with the known input ratio. As a rule of thumb in expression profiling, spike-in levels, corresponding to roughly one copy per cell, and changes around this level, should be easily detected. It is an open question whether abundances less than one copy per cell should be considered biologically meaningful. However, most tissue samples are actually mixtures of cell types, so the effective concentration of a low-abundance species functional in a rare cell type may be much less than one copy per cell. Linearity of the intensity versus abundance relation is of interest but not as fundamentally tied to information content in the data.

Spike-in materials may depart subtly in their length, labeling, and chemical properties from the actual target fragments, making them problematic as a performance measure. One way around this is to use spike-in materials that are as similar as possible to the targets of interest. Another strategy is to identify probes that repeatedly show large ratios between two different samples, then look at the observed versus linearly predicted ratios for these probes as these samples are mixed in different fractions. The Food and Drug Administration (FDA) has adopted this method, using rat tissues in a pilot study to develop assessment methods for submissions of microarray-based drug development data (55). Another way involves generating the Receiver Operating Characteristic (ROC) curve for difference detection (56). This curve shows, for a given process flow and data analysis stream, the trade between false positives and total positives as the detection threshold is varied. It is fundamentally related to information content in the data. By analyzing nominally identical independent samples, and also samples possessing true biological differences, false positives can be defined as differences detected between nominally identical samples, whereas total positives are defined as all detected differences.

Experiment Design and Data Interpretation

Analysis of microarray data touches on most of the important issues in bioinformatics, laboratory information management, statistics, and machine learning, and of course, it must be tightly coupled to the original experiment design and intent. Commercial and free software packages have been developed to help with some or all of the analysis process, and many useful methods discussions can be found now in textbooks (57–61). The discussion here is intended to point out a few underappreciated issues and give a broad sketch of the process.

Being attractive but expensive technology, microarrays were used initially often with insufficient measurement replication, and issues of false positives under multiple hypothesis testing were somewhat slow to be confronted. It is now appreciated that every probe has its own biases and error distribution due to a particular experimental workflow and that the errors have contributions that can look additive, multiplicative (e.g., log-normal), or Poisson. Intensity transformations (62–64), or a difference detection statistic based on explicit modeling of the presence of additive and multiplicative errors (4), can remove the gross trends of error level as a function of intensity, so that analysis of variance (ANOVA) methods become very useful for analyzing factorial experiment designs (65), such as samples taken across drug, dose, time, and tissue, including replicates. But with affordable numbers of replicates, the individual probe efficiencies and error levels remain poorly determined. Individual probe efficiencies can be separated from their target species abundance using any diverse set of conditions (66). And statistical "borrowing" can be done from replicates of related experimental conditions to better determine the individual error levels. So far it has not been possible to achieve all of the desired objectives in a computationally tractable framework, which would involve ANOVA-like recovery of the effects of the multiple experiment parameters, variance stabilization, and probe- and gene-specific error properties adapting to increasing amounts of available data.

Particularly in expression profiling applications, much of the information is contained in the behavior of low-abundance transcripts whose brightness is not much above the background level. Negative control spots, having as close as possible the average sequence properties of the other probes but avoiding homology to any expected sequences in the sample, are therefore very helpful in determining and subtracting background offsets. Affymetrix match/mismatch probe pairs solve this problem well for short oligos but expend half the probes on controls. Residual errors in background subtraction result in biases in ratio measurements that are more severe for the lower-intensity probes. This behavior can be partially corrected by assuming that there is a significant subset of transcripts in the sample that are not different between the two samples being ratioed (67), but this correction is hard to integrate into the desired framework mentioned above.

Experiment design must consider the measurement accuracy of the microarray platform being used and the magnitude of uncontrolled variation likely to be inherent in the biological samples (39, 68, 69). For most animal-based experiments, the best expression profiling techniques now produce measurement errors

that are smaller than typical animal-to-animal variation, so there is a point of rapidly diminishing return for running multiple array replicates on the same sample. The pairing of samples in two-color competitive hybridizations is problematic because different pairings can only be created *in silico* after the experiments by taking ratios of ratios, which increases errors (56). One of the colors can be devoted to a constant, universal (70) complex reference sample, so that any pair of samples can be compared by taking a ratio of ratios; however, this uses half of all hybridization throughput capacity for the control channel. In addition, two-color competitive hybridizations usually require dye-swap hybridization pairs to be accurate.

Having obtained abundance, abundance ratio estimates, and statistical significance values associated with these estimates for the target sequences in each biological sample condition, biological interpretation can begin. The meaning of each target sequence, resident in the world of gene annotations, is accessed via its unique identifier. See Chapter 2 of Stekel (22) for a brief review of annotation information resources.

Large sets of profiles usually are explored in an unsupervised manner by clustering and other algorithms (71–73), regardless of the original experiment design. This helps uncover process artifacts in the data as well as unexpected biology. For example, circadian rhythms may confound the intended experimental variates and be diagnosed initially by the presence of a cluster of genes in which this effect dominates. Once the artifacts are characterized, model-based methods can be brought to bear to filter them out of the data. Probably the most common analysis flow involves functionally interpreting genes whose expression covaries across a set of biological conditions. Annotations of the genes and of the experimental conditions can interpret the coregulated gene sets (74).

Many experiments are now oriented toward selecting mRNA or genetic biomarkers for predetermined end points. These present the problem of classifier generation or supervised learning (75, 76). There are many algorithmic options for finding predictive genes, with little firm understanding yet as to how to choose the best algorithm for a particular biological application. There also are many pitfalls surrounding what constitutes proper and sufficient validation. The following sections include examples of more complex data sets and integrative analysis of microarray data with other information.

EXPRESSION PROFILING

mRNA measurement applications have come to dominate microarray usage because of the rich information that can be derived about the functions of genes in cells and tissues. The expression of a set of transcripts, an "expression profile," can be compared across different tissues from a given organism, across disease states and genetic backgrounds, and across experimental conditions such as drug treatments and gene disruptions. These data sets are particularly powerful when collected simultaneously with other data types from the same biological samples.

Body Maps

The baseline abundances of mRNAs in different tissues, sometimes called a "body map" of expression, are of considerable interest because the list of tissues where a gene is expressed provides a key clue as to the function of the gene and as to where compounds targeting its protein are likely to do good or harm. If two genes have similar patterns of expression across tissues, this is a clue to functional relatedness. Gene Logic (Gaithersburg, MD) sells access to such databases it created via Affymetrix array profiling of tissue samples, and there are several publicly available databases (77).

Creating these maps is of course more difficult when the tissues are less dramatically different and the samples are smaller, such as in mapping regions of the mouse brain (78). In contrast, the Allen Institute for Brain Science (http://www.brainatlas.org/) is creating an expression map of the mouse brain by performing many in situ hybridization slices through a brain for each gene. The former method should reveal functional information for known structures more efficiently, whereas the latter method should eventually reveal a new fine structure and its associated gene expression.

Case Versus Control Studies: Disease States

The most common expression profiling experiment design compares two biological conditions, such as disease state versus normal state. Genes upregulated, or possibly downregulated, in the disease state are more likely than random chance to be drug targets for that disease and offer a detailed molecular phenotype of the disease. Meaningful results are critically dependent on realistic confidence assignments for the measured differences in expression; to belabor the obvious, a 1% false positive rate yields 500 false positives out of 50,000 reporters. Also, differential expression of any given gene can arise from proximal or more distal reactive steps in the disease process; it is just one clue to function that must be combined with other evidence (79–81). Clues as to the mechanisms of disease come from looking for functional categories of genes, such as those determined from databases of functional annotations, that are overrepresented among the differentially expressed genes (82). Simple case versus control studies have given way to more powerful experiment designs to suggest targets and illuminate disease mechanisms. For example, the aftermath of stroke in a rat model was followed over time in three brain regions using contralateral samples from each animal as matched controls (83). Gene groups regulated at different times postischemia meshed with existing stroke models, and a new candidate target for therapeutic intervention was identified.

Parsing Pathways

At first, it would seem unlikely that highly parallel gene expression analysis would be an efficient way to illuminate the detailed structure of signaling pathways. However, the question of the existence of cross talk between mitogen-activated

protein kinase (MAPK) signaling pathways was successfully addressed in this way (84). Because it was not known where in the 6000-gene yeast transcriptome the evidence for cross talk might appear, development of this experimental hypothesis was not practicable without microarrays.

The existence of off-target effects of drugs turns out to be a similar kind of question. By drug-treating a cell line lacking the gene for the drug target, one can argue that any observed transcriptional changes are off-target effects that might be associated with toxicity or unexpected mechanisms of the action. This was done for the immunosuppression compound FK506 (85), and the off-target effects were confirmed by extending the experiment to cell lines deleted both for the drug target and for the suspected secondary target.

Functional Response Patterns

The power of expression profiling is most evident in experiments that explore a systematically varied set of conditions. Data redundancy is provided by sampling a smoothly varying process, and coregulation of genes across a set of biological conditions reveals functional gene groups.

In a nonmicroarray study of this type from 1998, RT-PCR measurement sets for 112 genes at various times during rat central nervous system development (86) revealed features of the regulatory cascade. The advent of microarrays, of course, enabled a much broader sampling of the genome. DeRisi et al. (87) followed essentially all the genes in the yeast *Saccharomyces cerevisiae* growing in culture through its diauxic shift. Genes with related known metabolic function showed similar expression evolution over time. The transcriptional changes observed in other genes helped flesh out knowledge of the metabolic pathways involved. Progression of expression during development was followed during early metamorphosis in *Drosophila* (88), and genes were grouped according to their pattern of expression over the different phases of development. *Caenorhabditis elegans* was profiled over a set of developmental phases, growth conditions, and genetic mutations; the diversity of these conditions yielded strong groupings of coregulated genes (89).

As in the study mentioned above (83), increasing diversity of the conditions set, up to a point, yields stronger and more informative groupings of genes by coregulation. When this point is reached is a question that is answered by considering biological complexity and using algorithms to find patterns. In any event, these groupings still are subject to the caveat that similarity of response results in a "guilt by association" inference (90) and not proof of functional relatedness. Marcotte et al. (91) were able to estimate the relative power of expression coregulation data, physical association, and sequence-based analyses to infer functions of thousands of yeast genes; comparing these predictions to accepted functional annotations indicated fairly limited accuracy of the coregulation based inferences, although this depends on the set of conditions over which the expression profiles are obtained.

Spellman et al. (92) followed the yeast *S. cerevisiae* through two cell cycles, first phasing up the cells in the culture with multiple-cycle arrest and release methods.

On the basis of cyclic expression responses, hundreds of new genes were added to those already known to be regulated in the cell cycle, and functional relationships between different phases of the cycle were suggested. Despite the visually clear and pleasing expression patterns that resulted, the detailed conclusions of this and subsequent synchronization studies with microarrays have been called into question over statistical issues and whether most cells are in fact synchronized (93). This application and its controversy illustrate the power of the technology and the challenges in analyzing these large data sets with uncertain error behavior and biology.

Hughes et al. (4) profiled a large set of different single-gene disruption mutants in yeast, comparing their transcriptional state to the wild-type strain. The resulting patterns (Figure 4) provided a visualization of major pathway groupings and provided functional inferences for previously unannotated genes. Figure 4, in which rows and columns of the expression ratio data were reordered according to agglomerative hierarchical similarity clustering, illustrates the important distinction between two modes of functional inference. Proximity of two genes in the horizontal dimension is the guilt by association mode of inference. These genes respond similarly to the disruptions of other genes. But this kind of similarity often
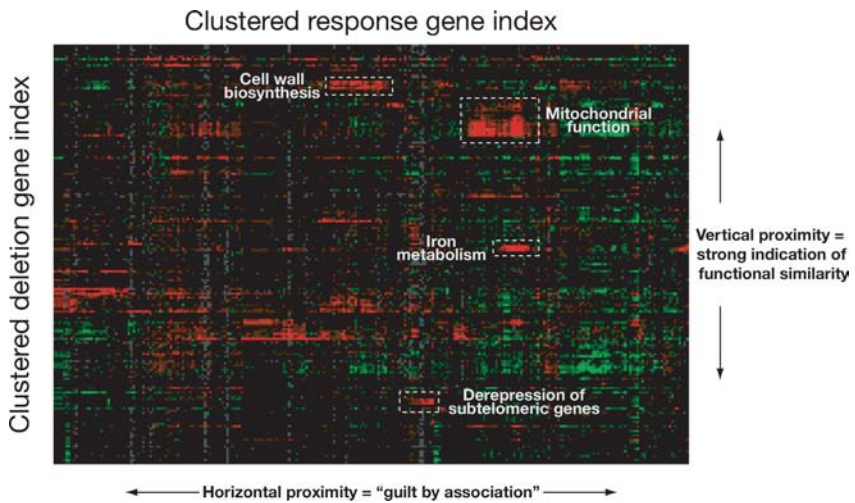
Figure 4 Expression responses to single-gene deletions in yeast. Each row represents the up- (*red*) or downregulation (*green*) of expression in response to a single-gene disruption in yeast (4). Only 300 genes (*columns*) are shown. These were the most responsive among the ∼6000 yeast genes measured in each two-color hybridization experiment. Columns, and independently the rows, have been rearranged via agglomerative hierarchical clustering to place rows with similar response patterns near each other, and columns with similar response patterns near each other. Each red or green island then represents a coordinated transcriptional response that is similar for each of a set of gene disruptions.

is the result of a fairly uninteresting downstream convergence of pathways, such as a global stress signature. Proximity on the vertical axis means two genes, when disrupted, produce similar cell responses at the molecular level. This similarity of molecular phenotype is a stronger indication of functional similarity, in analogy to conventional genetic studies, as was borne out by the biochemical confirmation rate achieved in this work for functional predictions based on the vertical axis proximity compared to those based on horizontal axis proximity (4).

Just as the molecular phenotypes associated with disruption of genes of known function serve as landmarks in the above example, toxicity landmarks in rat liver were produced by profiling the response to compounds of known toxicity (94, 95). The expression profiles produced by compounds under study then can be interpreted for the mechanism and likelihood of toxicity. Biological interpretation of the responding genes also gives clues to the mechanisms of toxicity. Similarly, efficacy landmarks can be provided by profiling drugs with known mechanisms of action. Expression responses to psychoactive compounds in primary human neurons in vitro were used to develop classifiers for antidepressant, antipsychotic, and opioid drug action (96). The products of such projects can be thought of either as biomarkers for particular classification decisions or as a general resource for interpreting the bioactivity of new compounds.

One of the most common experiment types in the drug discovery and diagnostics arena is the deliberate search for biomarkers of a particular human phenotypic end point. Cancer outcome prognosis is a very popular category of these because, conceivably, profitable microarray-based clinical treatment decisions are not many years off. Alizadeh et al. (97) found expression patterns indicative of survival in B-cell lymphoma patients and characteristics of two subtypes of large diffuse lymphoma B cells. In this study, the microarray probes were chosen to target genes expressed in lymphoid cells and to be relevant to immune response. A subset of the predictive markers was confirmed in follow-up validation studies with PCR. Van t' Veer et al. (98) were able to find an arithmetic function of the expression levels of 70 transcripts that predicted metastasis of breast tumors out of ∼25,000 profiled. This predictor was derived and cross-validated using a set of 98 patients, and then it was validated in a larger follow-up study of almost 300 patients (99). A recent meta-analysis of 84 microarray-based cancer outcome studies found that very few of them accomplished thorough validation and that, not surprisingly, larger cohorts and larger probe sets increased the chances of finding good biomarkers (100).

There is a close relationship in these studies between developing predictors and recognizing subphenotypes of disease. In general, the detailed molecular phenotype provided by expression profiling allows discrimination between multiple states that may at one moment have the same gross phenotype but for which the subsequent progression of events differs.

Integrative High-Throughput Studies

The exciting possibility of reconstructing biological pathways from large microarray expression data sets, hinted at in Figure 4, has proved elusive, although some

progress has been made by starting with the assumption of sparse network connectivity (101). When other data types are profiled at the same time, greater power is achieved.

By combining regulatory sequence motif findings with expression coregulation response clustering, better confidence is achieved in identifying regulatory elements and the sets of genes that should respond to them (102–105). The conclusions from these methods are checked further against direct experimental determinations (106) of transcription factor binding to the genome (107).

Building on a preliminary study, Ideker and coworkers (108, 109) obtained proteomic and mRNA response profiles for yeast cultures, corresponding to a set of galactose utilization gene mutations and growth conditions, and combined these data with physical interaction data to obtain refinements to the galactose utilization pathways. Knowing the levels of both the mRNAs and their associated proteins, including transcription factors, allows specific tests of alternative network models that are nearly impossible with mRNA data alone.

Schadt et al. (110) obtained from the same 111 mice, by microsatellite genotyping at 13-centimorgan (cM) spacing, 23,000-gene expression profiles from their livers and detailed phenotypic data. The data were combined in novel ways to find genetic loci controlling the mRNA levels and the phenotypic traits associated with common multigenic diseases like obesity. These methods promise to be an order of magnitude more efficient than conventional linkage analysis for finding causative alleles. In Figure 5, expression data for the most differentially regulated genes are displayed, organized by unsupervised clustering. The phenotypic codes on the left of the heatmap, indicating fat pad mass, show that even though the clustering was unsupervised, the results are dominated by changes associated with this phenotype. The structure evident in the horizontal direction suggests gene subsets are associated with different pathways involved in the phenotype, and the corresponding vertical structure suggests subsets of cross-bred animals that were fat for those reasons. In fact, by performing traditional quantitative trait loci analysis using subsets of animals identified in this way, linkage estimates, or log of the odds score peaks across the genome, were sharpened and intensified, and causative genes began to be associated with different subphenotypes of obesity.

There is an important distinction between integrating data types at the level of conclusions and integrating them at a deeper level in the context of a detailed biological model. Franke et al. (111) recently presented a software tool that brings together linkage, association, expression data, and functional annotations to support research objectives similar to those of Schadt et al. (110). However, this integration was at the level of inferences already drawn from the separate data sets. Schadt et al. (110) obtained the multiple data types from the same animals and used the individual mRNAs as intermediate phenotypes, implicitly placing them in a model that relates them to the causative alleles and to the downstream physiologic phenotypes. In fact, their work is being extended to reconstruct causal networks of genes. It is tempting to use the term "coherent" to refer to this deeper kind of data integration, in analogy to its usage in electrical engineering and signal processing, that leads to greater signal-to-noise ratios in detection.
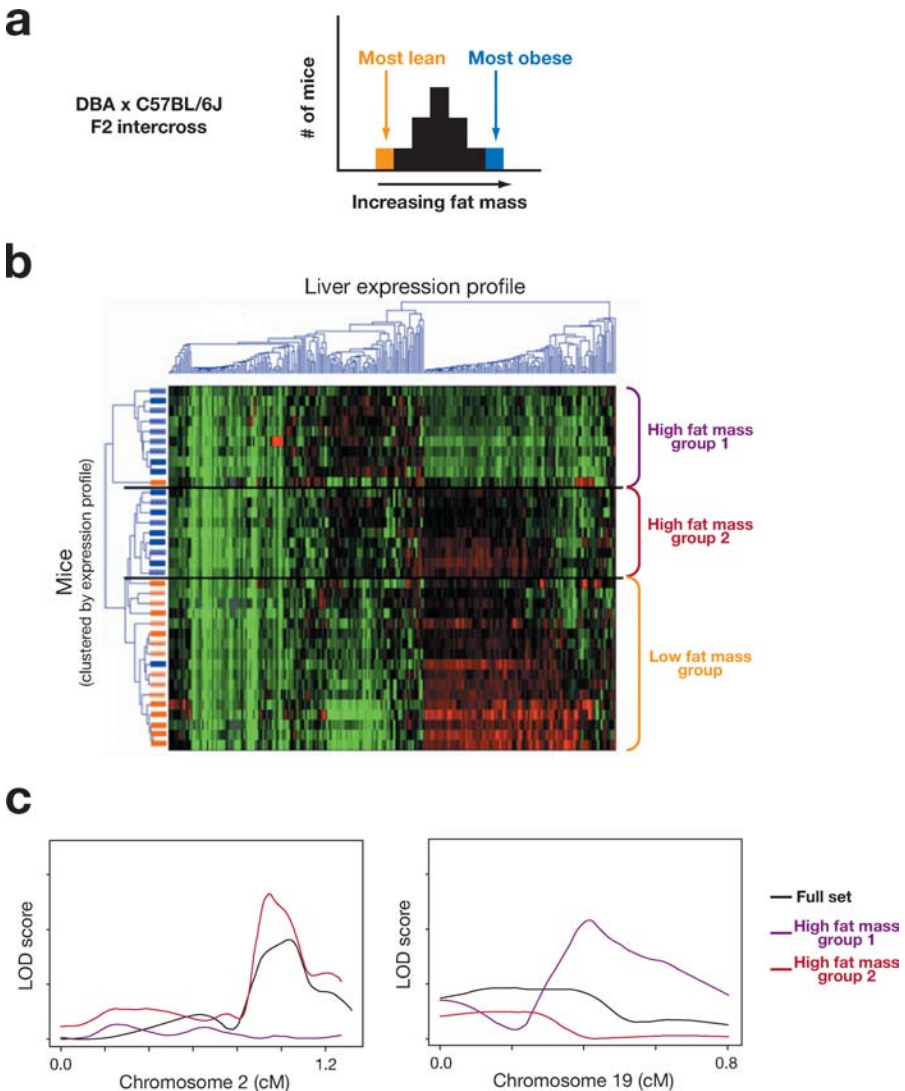
Figure 5 Coherent analysis of expression profiles, genotype scans, and phenotypic data from cross-bred mice (110). Animals with extreme phenotypic values, in this case fat pad mass, were selected for analysis (*a*). Unsupervised clustering of liver expression profiles allowed subgroups to be defined (*b*). Genetic linkages of phenotype (*c*) and of individual mRNA levels to genome locations were sharpened and strengthened by using these expression-defined phenotypic subgroups.

Weinstein et al. (112) combined 1376-gene expression profiles of the 60 National Cancer Institute cancer cell lines with growth inhibition data for these lines under 118 drug treatments to find clues as to molecular targets of the compounds. Correlation of growth inhibition with expression level over the 60 cell lines for each drug-gene pair yielded a 1376 $\times$ 118 table of correlation values. This table then was manipulated via clustering of rows and columns, as in Figure 4, to obtain groups of related targets and related compounds.

Splicing and Exon Discovery

Shoemaker et al. (113) designed a set of genomic tiling arrays for human chromosomes 21 and 22 to identify novel exons in an unbiased and comprehensive fashion. 60-mer probes were placed 30 nt steps across the repeat masked genomic sequence and the set of microarrays were hybridized with RNA from six different tissues. The presence of expression signals found across multiple human tissues was used to identify novel exons and to revise annotated gene structures. Figure 6 shows an example of the hybridization signal from a gene region of chromosome 21.

Multiple probes spanning several exons of a gene, hybridized to samples from multiple tissues, provide enough information to infer alternative splice forms. This was first demonstrated in rat tissues (114). Johnson and coworkers (25, 115) designed probes for $\sim$12,000 human genes that were hybridized with samples from multiple human tissues. The resulting hybridization data from the junction probes were used to identify thousands of novel splice variants along with an expression atlas showing in which tissues they were expressed.

ANALYSIS OF GENOMIC DNA

When the gene copy number is changed, there are corresponding changes in mRNA levels. Microarrays have, in this way, detected aneuploidies in yeast deletion strains (116) and have seen a mixture of regulation and copy number changes in prostate cancer cells (117). Copy number changes, of course, can be seen directly in the concentration of genomic DNA fragments from particular genome regions, and microarrays have been used to scan for cancer-related changes (118). The temporal progression of replication along the chromosome has even been tracked in this way (119).

Ren et al. (106) used microarrays to report which regulatory sites bound a certain transcription factor in yeast. A given transcription factor was allowed to bind to fragmented genomic DNA, which was then enriched for the bound regions using chromatin immunoprecipitation, amplified and labeled, and hybridized to an array of spots containing intergenic DNA. This approach was recently used to map the binding sites for three human transcription factors on chromosomes 21 and 22 (120). A similar technique was employed to map histone deacetylation in the yeast genome (121).
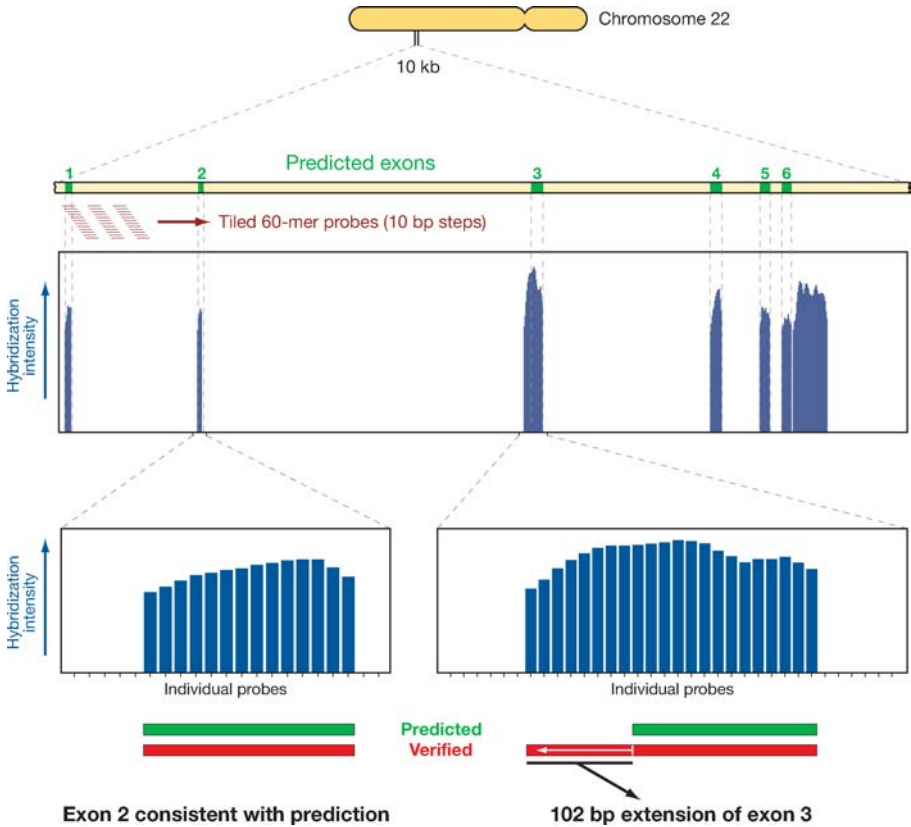
Figure 6 Tiling probes used to refine gene structures. Overlapping 60-mer oligo probes were designed to cover genomic regions where exons were computationally predicted (113), revised and reprinted with permission (http://www.nature.com/). In this display of a gene region on chromosome 21, probe intensity was truncated below the expected noise background level. For some exons, the computational prediction is confirmed. For others, it is revised by the data.

Detailed Characterization of Microbial Pathogens

The highly parallel interrogation of pathogen genomes, enabled by microarrays, promises to radically change the diagnosis of infectious disease, monitor emerging infections, and monitor the safety of food, water, and air. Probes constructed for genes in a baseline strain can be used to characterize and compare with the genomes of test strains via competitive hybridization. This approach was used to identify differences between *Mycobacterium tuberculosis* and the associated Bacillus Calmette-Guerin vaccine strain (122), to identify horizontal gene transfers causing methicillin resistance in *Staphylococcus aureus* (123), and to show

near genetic identity between strains responsible for two separate epidemics of rheumatic fever caused by Group A *Streptococcus* (124).

Tiling probes across the VP1 coding region of vaccine-derived poliovirus (VDPV) were used to detect emerging point mutations associated with vaccine virulence, whereas VDPV strain-specific probes were used to detect recombination events (125). Using PCR followed by microarray hybridization, probes targeted at specific genes are used to detect the presence of virulence factors, antigenic determinants (126) and drug resistance determinants (127) as well as to resolve closely related species of bacteria (128, 129). The advantages of microarray readout over specific PCR detection, such as one or a few TaqMan probe sets, are the number and diversity of identification sites that can be detected per amplicon and the degree of multiplexing in the PCR that can then be demultiplexed by the arrays. Several different highly multiplexed amplification strategies as front ends to microarray readout recently have been compared (130), including random primed PCR, isothermal full-genome amplification, and multiplexed specific PCR. Random primed amplification followed by microarray detection is particularly useful for detecting and identifying rapidly mutating viruses because by designing the microarray probes for conserved regions of genes, unexpected strain variants still will be detected, and a large number of virus species can be detected at once. Using conserved and strain-specific probes, this approach was demonstrated to detect and distinguish most respiratory viruses (131, 132), and the approach was used to help identify the recently emerged Severe Acute Respiratory Syndrome virus and its phylogenetic relationship to other viruses (132, 133).

Host-microbe interactions also could be studied in detail using a combination of genomic analysis of the pathogen and expression profiling of host immune cells (134, 135).

Genotyping

Rapid acquisition of genetic information was one of the original motivations for the Affymetrix microarray technology (10). Resequencing for point mutations using microarrays was demonstrated in 1996 (136) and has become an established methodology (137). The baseline method involves short probes complementary to every N-mer of the baseline target sequence and additional probes that vary the nucleotide at the putative mutation position. Each of these also can be paired with a "mismatch" probe to control for nonspecific hybridization. Chips have been designed for mutation detection in genes of particular interest to human health, including the cystic fibrosis gene CFTR (138), the breast cancer susceptibility gene BRCA1, P53 (139), and mitochondrial DNA (136, 137, 140). Studies of the performance of these devices, in the context of P53 (141) and mitochondrial DNA (140, 142), have shown the promise of these methods and also the difficulties associated with false detections when the underlying mutation rates are low.

When a specific list of known polymorphisms scattered throughout the genome need to be targeted, probe sets can be designed just for them, or microarrays can be used as a demultiplexing tool when the molecular recognition of the polymorphism takes place first in a highly multiplexed volume reaction (5, 143).

EMERGING TRENDS

Microarray technologies based on light-directed synthesis have room to follow a Moore's Law miniaturization for several more years before running up against fundamental physical limits imposed by the diffraction of light, which tend to place a lower limit of ∼1 micron on feature size. This miniaturization will allow both greater probe budgets and smaller hybridization volumes. The greater probe counts can be used to increase the redundancy of measurements and to independently report alternative splice forms for every gene in a single hybridization. Resequencing applications can tackle larger genome regions. The smaller hybridization volumes translate into less required biological input sample and/or less required amplification. With great care, laser-microdissected samples, comprising only a few cells, can be profiled currently in research settings (144, 145), and this will become commonplace. Microarray hybridizations increasingly will be run in a higher-throughput manner with sample prep done in 96-well plates (146).

Reagent use and cost will decrease with fluid volume. Eventually the costs will be low enough to make them tolerable in a routine medical point-of-care context; costs already have decreased to the point where some clinical diagnostic contexts with high reimbursable value per test, such as tests supporting choice of cancer treatment, could afford microarray measurement.

Cost may not be the main obstacle for long. In 1996, Oncormed (Gaithersburg, MD) began using an Affymetrix P53 chip in clinical trials to stratify patients with head and neck cancer. However, as of this review, DNA chips still have not made it into clinical practice. The FDA, in July 2003, sent a strong message to Roche Molecular Diagnostics about their Affymetrix microarray-based AmpliChip™ designed to report genotypes relevant to individual human drug metabolism differences; the effect of the message was that before marketing begins the AmpliChip™ would have to pass the stringent examination expected of any device "intended for a use which is of substantial importance in preventing impairment of human health" (147). Although not unexpected, this is a reminder that years of validation testing await any new set of diagnostic markers, even those generated with impressive technology. However, the FDA is actively involved in promoting a cooperative framework to evolve guidelines for use in drug discovery, clinical trials, and public health (148).

In the near term, mRNA biomarkers identified via microarray profiling may enter the clinic instead via individual RT-PCR tests or in the form of enzyme-linked

immunosorbent assay tests for associated circulating proteins. Because serum is so accessible and information rich, proteomic profiling for biomarkers is a strong competitor (149).

As proteomic and metabolomic profiling catch up to mRNA profiling, and the informatics infrastructure is developed to interpret these three data types coherently at the level of biological pathways (150), the acquisition of detailed biological knowledge will accelerate. It will be most powerful to make these measurements simultaneously on the same tissue samples, or at least the same organism, to achieve as much coherence as possible in the measurements.

Recent advances in achieving specific gene disruptions in mammalian cells via the RNA interference mechanism should allow large sets of known specific perturbations to be executed (151) and the collection of high-dimensional response data. The increasing speed and decreasing price at which these measurements can be accomplished and analyzed suggest a different way of doing biological research. For example, a drug company wishing to develop new therapeutics for a disease could stake out an area of biology (defined by the relevant sets of gene disruptions, existing compounds, disease states, and tissues) and profile the responses to essentially all possible manipulations, achieving for their investment a new, more complete level of understanding of the relevant pathways, lists of potential drug targets, and target N-tuples for combination therapies, ranked for probable efficacy and toxicity. One caveat here is that animal experiment costs do not follow Moore's Law, but small-volume cell cultures conceivably could.

Sharing and standardization of expression data are potentially very powerful directions for this work. Of the many publicly accessible databases for gene expression (152), Gene Expression Omnibus, at the National Center for Biotechnology Information (153), and its European counterpart Array Express, at the European Bioinformatics Institute (154), are intended to become large-scale repositories and have adopted standard data structures and guidelines for minimum supporting information, expressed in Extensible Markup Language. However, typically the in vivo experiment histories are not yet documented in enough detail for truly coherent analysis across data sets, even when the probe sequences can be related to common target sequences. Given the magnitude of the combined investments being made in these measurements, this is a frustrating limitation, but one which involves the subtle variations between nominally similar biological systems as well as bioinformatics issues.

ACKNOWLEDGMENTS

The *Annual Review of Biochemistry* is online at
http://biochem.annualreviews.org

LITERATURE CITED

1. Jordan B. 2001. In *DNA Microarrays: Gene Expression Applications*, ed. B Jordan, pp. 3–12. Berlin: Springer-Verlag

2. Zhu H, Bilgin M, Snyder M. 2003. *Annu. Rev. Biochem.* 72:783–812

3. Boguski MS, McIntosh MW. 2003. *Nature* 422:233–37

4. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. *Cell* 102:109–26

5. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, et al. 2003. *Highly Parallel SNP Genotyping*. Cold Spring Harbor, NY: Cold Spring Harbor Lab.

6. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. *Nat. Genet.* 14:450–56

7. Baldi P, Hatfield GW. 2002. See Ref. 155, pp. 7–13

8. Stekel D. 2003. See Ref. 156, pp. 1–18

9. Heller MJ. 2002. *Annu. Rev. Biomed. Eng.* 4:129–53

10. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. 1991. *Science* 251:767–73

11. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, et al. 1999. *Nat. Biotechnol.* 17:974–78

12. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, et al. 2001. *Nat. Biotechnol.* 19:342–47

13. Stekel D. 2003. See Ref. 156, pp. 7–8

14. Schena M, Shalon D, Davis RW, Brown PO. 1995. *Science* 270:467–70

15. Shalon D, Smith SJ, Brown PO. 1996. *Genome Res.* 6:639–45

16. Fan JB, Yeakley JM, Bibikova M, Chudin E, Wickham E, et al. 2004. *Genome Res.* 14:878–85

17. Roth ME, Feng L, McConnell KJ, Schaffer PJ, Guerra CE, et al. 2004. *Nat. Biotechnol.* 22:418–26

18. Ferguson JA, Boles TC, Adams CP, Walt DR. 1996. *Nat. Biotechnol.* 14:1681–84

19. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, et al. 2004. *Genome Res.* 14:870–77

20. Fulton RJ, McDade RL, Smith PL, Kienker LJ, Kettman JR Jr. 1997. *Clin. Chem.* 43:1749–56

21. Li F, Stormo GD. 2001. *Bioinformatics* 17:1067–76

22. Stekel D. 2003. See Ref. 156, pp. 19–61

23. Santa Lucia J Jr, Allawi HT, Seneviratne PA. 1996. *Biochemistry* 35:3555–62

24. Vainrub A, Pettitt BM. 2003. *Biopolymers* 68:265–70

25. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. 2003. *Science* 302:2141–44

26. Baldi P, Hatfield GW. 2002. See Ref. 155, pp. 29–52

27. Nallur G, Luo CH, Fang LH, Cooley S, Dave V, et al. 2001. *Nucleic Acids Res.* 29:e118

28. Stears RL, Getts RC, Gullans SR. 2000. *Physiol. Genomics* 3:93–99

29. Karsten SL, Van Deerlin VMD, Sabatti C, Gill LH, Geschwind DH. 2002. *Nucleic Acids Res.* 30:e4

30. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. 1990. *Proc. Natl. Acad. Sci. USA* 87:1663–67

31. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, et al. 1992. *Proc. Natl. Acad. Sci. USA* 89:3010–14

32. Luo L, Salunga RC, Guo H, Bittner A, Joy KC, et al. 1999. *Nat. Med.* 5:117–22

33. Froussard P. 1993. *PCR Methods Appl.* 2:185–90

34. Zhao S, Molnar G, Zhang J, Zheng L, Averboukh L, Pardee AB. 1998. *BioTechniques* 24:842–50, 52

35. Chenchik A, Zhu YY, Diatchenko L, Li R, Hill J, Siebert PD. 1998. In *Gene Cloning and Analysis by RT-PCR*, ed. PD Siebert, JW Larrick, pp. 305–19. Westborough, MA: Eaton

36. Smith L, Underhill P, Pritchard C, Tymowska-Lalanne Z, Abdul-Hussein S, et al. 2003. *Nucleic Acids Res.* 31:e9

37. Parrish ML, Wei N, Duenwald S, Tokiwa GY, Wang Y, et al. 2004. *J. Neurosci. Methods* 132:57–68

38. Dobbin K, Shih JH, Simon R. 2003. *Bioinformatics* 19:803–10

39. Chen JJ, Delongchamp RR, Tsai CA, Hsueh HM, Sistare F, et al. 2004. *Bioinformatics* 20:1436–46

40. Naef F, Magnasco MO. 2003. *Phys. Rev. E* 68:011906

41. Hekstra D, Taussig AR, Magnasco M, Naef F. 2003. *Nucleic Acids Res.* 31:1962–68

42. Sartor M, Schwanekamp J, Halbleib D, Mohamed I, Karyala S, et al. 2004. *BioTechniques* 36:790–96

43. Dai H, Meyer M, Stepaniants S, Ziman M, Stoughton R. 2002. *Nucleic Acids Res.* 30:e86

44. Ku WC, Lau WK, Tseng YT, Tzeng CM, Chiu SK. 2004. *Biochem. Biophys. Res. Commun.* 315:30–37

45. Liu RH, Lenigk R, Druyor-Sanchez RL, Yang J, Grodzinski P. 2003. *Anal. Chem.* 75:1911–17

46. Korkola JE, Estep AL, Pejavar S, De-Vries S, Jensen R, Waldman FM. 2003. *BioTechniques* 35:828–35

47. Fare TL, Coffey EM, Dai H, He YD, Kessler DA, et al. 2003. *Anal. Chem.* 75:4672–75

48. Baldi P, Hatfield GW. 2002. See Ref. 155, p. 19

49. Mansson A, Sundberg M, Balaz M, Bunk R, Nicholls IA, et al. 2004. *Biochem. Biophys. Res. Commun.* 314:529–34

50. Schultz S, Smith DR, Mock JJ, Schultz DA. 2000. *Proc. Natl. Acad Sci. USA* 97:996–1001

51. Oldenburg SJ, Genick CC, Clark KA, Schultz DA. 2002. *Anal. Biochem.* 309:109–16

52. Kamberova G, Shah S. 2002. *DNA Array Image Analysis: Nuts & Bolts*. Salem, MA: DNA Press. 202 pp.

53. Stekel D. 2003. See Ref. 156, pp. 62–72

54. Yu F, Yao D, Knoll W. 2004. *Nucleic Acids Res.* 32:e75

55. Thompson KL. 2003. *The value of standards for regulatory product approval applications incorporating nonclinical and clinical gene expression microarray data*. http://www.cstl.nist.gov/div831/UniversalRNAStds/Thompson.pdf

56. He YD, Dai H, Schadt EE, Cavet G, Edwards SW, et al. 2003. *Bioinformatics* 19:956–65

57. Stekel D. 2003. See Ref. 156, pp. 62–252

58. Kohane IS, Kho AT, Butte AJ. 2003. *Microarrays for an Integrative Genomics*. Cambridge, MA: MIT Press. 306 pp.

59. Speed T, ed. 2003. *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton, FL: Chapman & Hall/CRC. 222 pp.

60. Knudsen S. 2004. *Guide to Analysis of DNA Microarray Data*. Hoboken, NJ: Wiley. 168 pp. 2nd ed.

61. Baldi P, Hatfield GW. 2002. See Ref. 155, pp. 29–176

62. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. 2002. *Bioinformatics* 18 (Suppl. 1):S105–10

63. Durbin BP, Rocke DM. 2004. *Bioinformatics* 20:660–67

64. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. 2002. *Bioinformatics* 18(Suppl. 1):S96–104

65. Kerr MK, Martin M, Churchill GA. 2000. *J. Comput. Biol.* 7:819–37

66. Li C, Wong WH. 2001. *Proc. Natl. Acad. Sci. USA* 98:31–36

67. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. 2001. *Nucleic Acids Res.* 29:2549–57

68. Churchill GA. 2002. *Nat. Genet.* 32 (Suppl.):490–95

69. Speed T, Yang YH. 2003. See Ref. 59, pp. 35–92

70. Novoradovskaya N, Whitfield ML, Basehore LS, Novoradovsky A, Pesich R, et al. 2004. *BMC Genomics* 5:20

71. Stekel D. 2003. See Ref. 156, pp. 139–182

72. Chipman H, Hatie T, Tibshirani R. 2003. See Ref. 59, pp. 159–200

73. Valafar F. 2002. *Ann. NY Acad. Sci.* 980:41–64

74. Rougemont J, Hingamp P. 2003. *BMC Bioinformatics* 4:15

75. Stekel D. 2003. See Ref. 156, pp. 183–210

76. Dudoit S, Fridlyand J. 2003. See Ref. 59, pp. 93–158

77. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. 2004. *Proc. Natl. Acad. Sci. USA* 101:6062–67

78. Del Rio JA, Barlow C. 2002. *Prog. Brain Res.* 135:149–60

79. Lock C, Hermans G, Pedotti R, Brendolan A, Schadt E, et al. 2002. *Nat. Med.* 8:500–8

80. Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, et al. 2002. *Nat. Genet.* 32(Suppl.):509–14

81. Miklos GL, Maleszka R. 2004. *Nat. Biotechnol.* 22:615–21

82. Breitling R, Amtmann A, Herzyk P. 2004. *BMC Bioinformatics* 5:34

83. Roth A, Gill R, Certa U. 2003. *Mol. Cell. Neurosci.* 22:353–64

84. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, et al. 2000. *Science* 287:873–80

85. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, et al. 1998. *Nat. Med.* 4:1293–301

86. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, et al. 1998. *Proc. Natl. Acad. Sci. USA* 95:334–39

87. DeRisi JL, Iyer VR, Brown PO. 1997. *Science* 278:680–86

88. White KP, Rifkin SA, Hurban P, Hogness DS. 1999. *Science* 286:2179–84

89. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, et al. 2001. *Science* 293:2087–92

90. Clare A, King RD. 2002. *In Silico Biol.* 2:511–22

91. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. *Nature* 402:83–86

92. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. 1998. *Mol. Biol. Cell* 9:3273–97

93. Cooper S, Shedden K. 2003. *Cell Chromosome* 2:1

94. Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, et al. 2001. *Toxicol. Appl. Pharmacol.* 175:28–42

95. McMillian M, Nie AY, Parker JB, Leone A, Kemmerer M, et al. 2004. *Biochem. Pharmacol.* 67:2141–65

96. Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. 2003. *Proc. Natl. Acad. Sci. USA* 100:9608–13

97. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. 2000. *Nature* 403:503–11

98. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. 2002. *Nature* 415:530–36

99. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. 2002. *N. Engl. J. Med.* 347:1999–2009

100. Ntzani EE, Ioannidis JP. 2003. *Lancet* 362:1439–44

101. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. 2003. *Science* 301:102–5

102. Sinha S, Tompa M. 2000. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8:344–54

103. GuhaThakurta D, Palomar L, Stormo GD, Tedesco P, Johnson TE, et al. 2002. *Genome Res.* 12:701–12

104. Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS. 2003. *Nat. Biotechnol.* 21:435–39

105. Keles S, Van Der Laan MJ, Vulpe C. 2004. *Bioinformatics* 20(16):2799–811

106. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. 2000. *Science* 290:2306–9

107. Gao F, Foat BC, Bussemaker HJ. 2004. *BMC Bioinformatics* 5:31

108. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, et al. 2002. *Mol. Cell. Proteomics* 1:323–33

109. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. 2001. *Science* 292:929–34

110. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. 2003. *Nature* 422:297–302

111. Franke L, van Bakel H, Diosdado B, van Belzen M, Wapenaar M, Wijmenga C. 2004. *Eur. J. Hum. Genet.* 12(8):633–38

112. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, et al. 1997. *Science* 275:343–49

113. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, et al. 2001. *Nature* 409:922–27

114. Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, et al. 2001. *Genome Res.* 11:1237–45

115. Castle J, Garrett-Engele P, Armour CD, Duenwald SJ, Loerch PM, et al. 2003. *Genome Biol.* 4:R66

116. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, et al. 2000. *Nat. Genet.* 25:333–37

117. Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, et al. 2001. *Cancer Res.* 61:8143–49

118. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, et al. 2003. *Genome Res.* 13:2291–305

119. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, et al. 2001. *Science* 294:115–21

120. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. 2004. *Cell* 116:499–509

121. Robyr D, Suka Y, Xenarios I, Kurdistani SK, Wang A, et al. 2002. *Cell* 109:437–46

122. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, et al. 1999. *Science* 284:1520–23

123. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM. 2001. *Proc. Natl. Acad. Sci. USA* 98:8821–26

124. Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, et al. 2002. *Proc. Natl. Acad. Sci. USA* 99:4668–73

125. Cherkasova E, Laassri M, Chizhikov V, Korotkova E, Dragunsky E, et al. 2003. *Proc. Natl. Acad. Sci. USA* 100:9398–403

126. Chizhikov V, Rasooly A, Chumakov K, Levy DD. 2001. *Appl. Environ. Microbiol.* 67:3258–63

127. Volokhov D, Chizhikov V, Chumakov K, Rasooly A. 2003. *J. Appl. Microbiol.* 95:787–98

128. Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL. 2002. *Mol. Cell. Probes* 16:119–27

129. Volokhov D, Chizhikov V, Chumakov K, Rasooly A. 2003. *J. Clin. Microbiol.* 41:4071–80

130. Vora GJ, Meador CE, Stenger DA, Andreadis JD. 2004. *Appl. Environ. Microbiol.* 70:3047–54

131. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. 2002. *Proc. Natl. Acad. Sci. USA* 99:15687–92

132. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, et al. 2003. *PLoS Biol.* 1:E2

133. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, et al. 2003. *Science* 300:1394–99

134. Cummings CA, Relman DA. 2000. *Emerg. Infect. Dis.* 6:513–25

135. Bryant PA, Venter D, Robins-Browne R, Curtis N. 2004. *Lancet Infect. Dis.* 4:100–11

136. Chee M, Yang R, Hubbell E, Berno A, Huang XC, et al. 1996. *Science* 274:610–14

137. Hacia JG. 1999. *Nat. Genet.* 21:42–47

138. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. 1996. *Hum. Mutation* 7:244–55

139. Ahrendt SA, Halachmi S, Chow JT, Wu L, Halachmi N, et al. 1999. *Proc. Natl. Acad. Sci. USA* 96:7382–87

140. Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, et al. 2001. *Genome Res.* 11:1913–25

141. Wikman FP, Lu ML, Thykjaer T, Olesen SH, Andersen LD, et al. 2000. *Clin. Chem.* 46:1555–61

142. Maitra A, Cohen Y, Gillespie SE, Mambo E, Fukushima N, et al. 2004. *Genome Res.* 14:812–19

143. Fan JB, Chen X, Halushka MK, Berno A, Huang X, et al. 2000. *Genome Res.* 10:853–60

144. Kamme F, Zhu J, Luo L, Yu J, Tran DT, et al. 2004. *Methods Mol. Med.* 99:215–24

145. Glanzer JG, Eberwine JH. 2004. *Br. J. Cancer* 90:1111–14

146. Zarrinkar PP, Mainquist JK, Zamora M, Stern D, Welsh JB, et al. 2001. *Genome Res.* 11:1256–61

147. Kling J. 2003. *Nat. Biotechnol.* 21:959–60

148. Petricoin EF, Hackett JL, Lesko LJ, Puri RK, Gutman SI, et al. 2002. *Nat. Genet.* 32(Suppl.):474–79

149. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, et al. 2002. *Lancet* 359:572–77

150. Neumann E, Thomas J. 2002. *Drug Discov. Today* 7:S160–62

151. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, et al. 2004. *Nature* 428:431–37

152. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. 2003. *Nucleic Acids Res.* 31:68–71

153. Edgar R, Domrachev M, Lash AE. 2002. *Nucleic Acids Res.* 30:207–10

154. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, et al. 2003. *C. R. Biol.* 326:1075–78

155. Baldi P, Hatfield GW, eds. 2002. *DNA Microarrays and Gene Expression: from Experiments to Data Analysis and Modeling.* Cambridge, UK: Cambridge Univ. Press. 213 pp.

156. Stekel D, ed. 2003. *Microarray Bioinformatics.* Cambridge, UK: Cambridge Univ. Press. 263 pp.

$\mathbb{A}\mathbb{R}$ *Annual Review of Biochemistry*
Volume 74, 2005

CONTENTS