Microarrays are an important technology for the high-throughput investigation of biological phenomenon. Because a number of different microarray technologies exist, and because microarrays can be used to investigate a wide array of biological phenomenon, the terminology can be cumbersome. This review focuses on nucleotide-based microarrays, by far the most popular type, and provides a simple structure for characterizing microarray technologies in terms of the technical characteristics of the array (the platform) and the biological phenomenon being probed.

Introduction
Parallelization is one of the major trends in modern biomolecular research. Assays that were previously performed for single genes or moieties are being replaced by experiments that analyze hundreds of thousands of moieties or parameters simultaneously. For example, classical genotyping is a laborious, gene-at-a-time PCR-based process. Today, so-called “SNP Chips” are providing the ability to genotype hundreds of thousands of loci simultaneously, and in the future low-cost genome-sequencing will allow for even greater parallelization.

This is the first review in a series of three that aim to introduce one of the major tools in this parallelization trend: microarrays. This first part overviews the technologies, platforms, and major types of uses: the basic nomenclature and language of microarrays. The first key question answered here is: what are microarrays? Next, we seek to classify microarray technologies based on two broad characteristics: the physical characteristics of the microarray (the “platform”) and the biological significance of the sequences on the microarray (the “probes”). In particular, I focus on avoiding some of the terminological mistakes that confound discussions of microarray technologies.

Building on this foundation, the second review in this series will describe the statistics and data-analysis aspects of different types of microarray experiments. The third review will discuss how microarrays can be used for hypothesis-testing studies, not only hypothesis-generating ones.

What are Microarrays?
A microarray is a physical substrate, such as a glass slide, on which many samples of a biological moiety are placed. Each of these “spots” on the array can be used for a simple experiment, and the outcome of each “mini-experiment” can be read simultaneously using a variety of imaging modalities. Figure 1a overviews this basic process.

Consider a practical example (Figure 1b). First, different fragments of genomic DNA are affixed to a glass slide in a series of spots. Next, genomic DNA is extracted from a tumour and labeled with a fluorescent marker. A solution containing this labeled DNA is passed over the slide, and matching (complementary) DNA sequences will bind together through Watson-Crick base-pairing. After rinsing the solution from the slide, matching spots can be identified by the fluorescent marker incorporated into the tumour DNA. This provides a simple way of looking for duplications and deletions in a cancer genome. Duplications will be evident as spots with elevated DNA binding (high fluorescent signal) while deletions will be shown as spots with reduced DNA binding (low fluorescent signal).

Given this basic framework, the specifics of microarrays can vary significantly. For example, numerous physical substrates can be used, including glass slides to 96- or 384-well plates to
quartz chips. While the vast majority of arrays have nucleic acids “spotted” as described above, proteins have also been used on microarrays. And while fluorescent detection is most common, a variety of fluorophores are in common use, and radioisotope detection has also been explored. For the remainder of this review, we focus on microarrays that involve nucleotides – either DNA or RNA – spotted to the physical substrate. These nucleotide arrays represent the vast majority of published array studies, with literally thousands of times more papers using nucleotide than protein-arrays.

The two key features of a microarray experiment are the platform used, and the biological significance of the probes. The “platform” of an array refers to the underlying technology used to construct an array and visualize its results. The “probes” of an array refer to the specific type of biological question being asked. Below, we review each of these features in turn.

Of Platforms...

Focusing on nucleotide-based arrays, array platforms can be classified based on two main features. The first is the length of the nucleotide-sequence affixed to the microarray; the second is the method used for visualization of the experimental results.

Long cDNA Arrays

One of the earliest and most commonly used platforms for microarrays are so-called “cDNA arrays”. These platforms involve glass slides to which long complementary DNA sequences (cDNAs) have been affixed. In this context, long typically means from 300 base-pairs to tens of thousands of base-pairs, depending on the probe-type, with short probes typically for assessing mRNA levels and long probes for assessing genomic amplification.

This platform was extremely popular when microarray first originated because scientists were able to exploit existing cDNA libraries for probing the levels of different mRNA transcripts. Indeed, large cDNA libraries exist for many
species whose genomes are not yet sequenced (Figure 2). As of July 31st, 2006 there were 385 completely sequenced genomes, of which only 22 are eukaryotes. By contrast, 600 organisms have more than 1000 ESTs available. Thus, cDNA arrays are one popular way of working with species whose genomic sequence is incomplete or unavailable.

The long sequences used in cDNA microarrays provide several advantages over other platforms. The long length of these sequences allows for high specificity and uniqueness within the transcriptome or genome. Even when two proteins are highly similar, regions of their mRNA sequences are often highly divergent. One good example of this phenomenon arises in the cytochrome P450 family of drug-metabolizing enzymes. One protein, Cyp1a1, is 73% identical and 85% similar to its paralog, Cyp1a2 (for a description of sequence-alignments see (1)). Distinguishing these proteins can thus be quite challenging. At the mRNA level, however, the two transcripts are only alignable along 20% of their length. Thus, distinguishing the two transcripts with long cDNA clones is quite straightforward.

On the other hand, cDNA arrays are generally unable to distinguish transcripts with uniform, high similarity. For example, an alternative splicing event that changes 25 bp of a 1000 bp transcript is extremely difficult to detect because the differential hybridization strength is so small between the two variants. As a result, cDNA arrays are often capable of detecting many variants of a transcript, even those that have not yet been identified or characterized.

Short Oligonucleotide Microarrays
In contrast to the multi-hundred base-pair sequences of cDNA microarrays, short oligonucleotide arrays use sequences of 25 to 50 base-pairs in length. Typically these arrays represent each mRNA or genomic region with multiple sequences that are “aggregated” to provide a single measure. The most common short oligonucleotide platform is that produced by Affymetrix Inc. (Santa Clara, CA), which uses between 10 and 20 sequences of 25 base-pairs in length for each gene or genomic feature. The manufacture of these arrays is quite different from that of cDNA arrays, and their construction is quite similar to that of a semi-conductor, where a photolithographic mask is used to specifically synthesis unique 25bp sequences one base at a time across the entire array. Many versions of these arrays have incorporated two versions of each sequence: one exact match (termed the Perfect Match, or PM sequence) and the other with a single base-pair mismatch in the central (13th) position (termed the Mismatch or MM sequence). As will be discussed in the next part of this review,

![Figure 2: Distribution of ESTs by Organism.](image)

Extensive EST libraries are available for many species. In total, more than 100,000 ESTs have been sequenced and deposited in GenBank for 49 distinct organisms, while in total, EST data is available for 1169 organisms. This broad coverage was one of the major driving factors behind the early popularity of cDNA-based microarray platforms for mRNA expression profiling.
several studies have indicated that the MM sequences are not a helpful control for non-specific hybridization, and many analysis methods have abandoned them.

Short oligonucleotide platforms, like the Affymetrix arrays, have three major advantages. First, the manufacturing technique used tends to produce highly reproducible results, and the array-to-array variability appears to be lower than for cDNA microarrays. Second, because each feature is represented by multiple sequences, there is some degree of internal replication that can help provide greater confidence in the experimental integrity. Finally, because the individual sequences are so short, they can be used to interrogate expression of small features, such as the expression of single exons or the mutational status of specific SNPs, thus providing finer resolution than is available using cDNA microarrays.

However, there are also disadvantages to short oligonucleotide platforms. For certain genomic features, such as transcription-factor binding sites, the 25 bp sequences might be too short to provide sufficient specificity. Indeed, there are many 25 bp sequences that occur with great frequency across mRNAs or non-coding genomic sequences. In addition, the fine resolution of these arrays means that if a splice-variant or other specific feature is not specifically known about during array design, it will be missed entirely. By contrast, cDNA platforms might be able to “aggregate” the signal from many transcript variants by virtue of their longer sequence-lengths.

Long Oligonucleotide Microarrays
The third major platform for nucleotide microarray work involves “long oligonucleotides” that range in length from 50 to 100 bp each. These longer sequences are thought to be a compromise between the high sensitivity, but poor specificity of longer cDNA clones, and the high specificity, but lower sensitivity of short-oligonucleotide arrays. Several publications have seemed to indicate they succeed in this regard (2). Long oligonucleotide arrays can be produced in several different ways, such as by spotting of clone libraries as is used for cDNA arrays or by piezoelectric spray printing, as used by Agilent Technologies Inc. (Palo Alto, CA).

In general, short-oligonucleotide arrays, and in particular Affymetrix arrays, are exclusively one-colour array experiments. Indeed, Affymetrix arrays also employ a non-Cyanine fluorophore – a protein called phycerythrin. A few groups have considered radiolabels for detection purposes (4) but in general, it has not achieved significant popularity.

The distinction between one- and two-colour microarray experiments is an important one. While one-colour experiments provide the advantage of absolute quantitation, the correlation of these absolute intensities with specific concentrations is complex (5). Two-colour studies offer no possibility of absolute quantitation, but simultaneously interrogate two experimental samples, thus providing a higher sample-throughput rate under some experimental designs (6).

…And Probes
The second key feature of a microarray experiment is the biological relevance of the sequences themselves, rather than how long they are or how they are attached to the physical substrate. With one exception – tiling arrays, dealt with immediately below – the biological question will determine which type of array is used.

Tiling Arrays
Tiling arrays are neither a specific type of platform nor a specific type or probe, but rather a general way of designing probes. In a tiling
array, probes are directly adjacent or slightly overlapping (Figure 3), and thus "tile" across the sequence of interest. These arrays have several advantages, including the ability to detect unexpected variations, such as alternative splicing events, as well as to control for cross-hybridization through "sharing" of the information across neighbouring probes. A number of major publications have employed tiling arrays for various studies over the last two years (7-10), and an excellent overview of the similarities and discrepancies across several tiling experiments has been published (11).

Transcript Arrays
The most common type of microarrays measures the levels of entire mRNA transcripts. First developed in the mid-1990s (12, 13), these arrays were applied initially to yeast, and subsequently to mammals and prokaryotes. They have been used broadly to investigate a number of biological questions, ranging from the molecular characteristics of different tumour types (14) to the identification of novel gene functions (15) to the development of predictors of toxicity (16).

These arrays have become standard parts of research, but have two key weaknesses. First, because the arrays are designed to reflect mRNA transcripts, the results are only as comprehensive as current knowledge of transcript sequences and splice-variants. Therefore, results from older arrays are often found to be discordant with those from newer arrays (17, 18). Second, because these arrays are designed to assess expression across an entire transcript, they miss the subtleties of splice-variants and other transcriptional variants. To address these issues, two types of transcript arrays have been developed: exon arrays and junction arrays.

Exon Arrays
Exon arrays are much like transcript arrays, but are designed to target the expression of only a single exon of a gene. These arrays only monitor the expression of a single exon, but probes can be designed to all the exons from a gene, and then the values for multiple exons can be aggregated into a single value using complex statistical analyses (19, 20). These arrays are relatively new, but Affymetrix has released a series of new arrays for the profiling of human, mouse, and rat exons (http://www.affymetrix.com/products/arrays/exon_application.affx). The disadvantage of exon arrays is that, as with transcript arrays, they are only as useful as current knowledge of which exons exist. Nevertheless, this is a weakness of all non-tiling transcriptomic arrays, and for well-studied species with large amounts of EST data

Figure 3: Genomic Tiling Arrays. Genomic tiling arrays represent each non-repetitive portion of the genome with multiple probes that either overlap or lie adjacent to one another. Here, four partially overlapping clones can be seen covering one stretch of DNA.
and sequenced genomes (21), there can be reasonable confidence that a large proportion of alternative transcripts have been identified.

**Junction Arrays**
The second approach to improving transcript arrays involves designing probes directly to intron-exon junctions. These so-called junction arrays can help detect which splice-variants are expressed. It can be argued that exon arrays provide just as much information as junction arrays, but with much more complex statistical analyses. Surprisingly, after an initial paper describing a major survey of junction expression across human tissues (22) few junction-array analyses appear to have been performed.

**Promoter/CpG Island Arrays**
While transcript, exon, and junction arrays all aim to identify the level of mRNA present in a sample, promoter arrays aim to determine characteristics of DNA itself. For example, arrays encoding promoter regions of the genome (23, 24) or CpG islands (25, 26) have been combined with chromatin immuno-precipitation (ChIP) to allow identification of DNA-protein interactions. Similarly, such arrays have been used to investigate differential methylation of regions of DNA (27), and in theory could be used to search for other DNA modifications. In these arrays, the target sequences represent regions of genomic DNA that are adjacent to genes or to putative regulatory regions.

The disadvantage of promoter arrays is that they focus on regulatory regions that are in close proximity to a gene. This approach fails in two situations. First, for transcribed regions that have not yet been identified, such as novel untranslated RNAs, promoter arrays will fail to interrogate their regulation. This is a similar weakness to expression arrays, but the expression of transcripts from EST databases often out-paces the characterization of how those transcripts are coded within the genome itself (28). The solution to these two weaknesses is straight-forward: genomic tiling arrays.

**Genomic Arrays**
Arrays that cover the entire non-repetitive genome are now commercially available, and have proven to be very useful for understanding genome-wide regulatory phenomenon. For example, a major advance in our understanding of human transcription was sparked by the genome-wide survey of protein:DNA associations for the pre-initiation complex of the basal transcription machinery by Kim and coworkers (7). Another recent example is the genome-wide characterization of various transcription-factors in embryonic stem cells (29), where this survey highlighted the association of one protein with transcribed regions and another with repressed genomic regions.

Similarly, genomic arrays can also be used to identify regions of amplification and deletion across the genome, as described in the example in the first part of this review. Such experiments have frequently been used in the context of cancer because genomic alterations are a hallmark of the disease. More recently, a number of groups have sought and found reasonable correlations between genomic alterations and mRNA expression levels (30). These arrays can also search for small-scale genomic abnormalities that might arise in normal cells and might be polymorphic between individuals (so-called copy-number polymorphisms), as has been reviewed recently (31). Finally, as mentioned at the beginning of this review, genomic arrays have been used for high-throughput SNP mapping and linkage analysis (32).

**Conclusions**
The terminology behind microarray experiments is important, but relatively simple. Platforms are typically described based on the length of sequence tethered to the array and technology used to visualize the results. Probes are described based on the type of biological feature they are intended to assess, and can be described as “tiling” or “non-tiling”. The combinatorial effect of this diversity of platforms and probes makes it important to be specific in describing microarray experiments. Statements such as “we will investigate the effects of drug X using microarrays” are sorely lacking. Will the group be investigating changes in mRNA expression? Searching for genotoxic stresses with a genomic array? Understanding how the drug changes the association of specific transcription-factors with DNA? Clear communication about microarray experiments requires the careful specification of probe and platform details.
As microarray technologies progress the clear trend is towards larger arrays that represent larger portions of the transcriptome or genome. Exon arrays are dramatically more informative than simple expression arrays, but generate an order of magnitude more information. For example, the RefSeq annotation of build hg17 of the human genome on July 31st, 2006, included 24,722 transcripts, but 260,043 distinct exons – suggesting an order of magnitude increase in the information. Similarly, promoter arrays cover only about 1% of the non-repetitive genome, so whole-genome arrays for ChiP-analyses generate orders of magnitude more information than experiments performed within the last two years.

Is the analysis of these enormous datasets similar to that of traditional arrays? And how should standard promoter and transcript arrays be analyzed – is there a consensus in the field? The next review in this series will tackle exactly these questions.

References:
1. P. C. Boutros, Hypothesis 3, 26-33 (June 2005).
20. Q. Pan et al., Trends Genet 21, 73-7 (2005).