Beyond expression profiling: Next generation uses of high density oligonucleotide arrays

Philipp Kapranov, Victor I. Sementchenko and Thomas R. Gingeras

Date received (in revised form): 12th February 2003

Abstract
In the past decade, microarray technology has become a major tool for high-throughput comprehensive analysis of gene expression, genotyping and resequencing applications. Currently, the most widely employed application of high-density oligonucleotide arrays (HDOAs) involves monitoring changes in gene expression. This application has been carried out in a variety of organisms ranging from Escherichia coli to humans. The recent near completion of the human and mouse genome sequences, however, as well as the genomes of other model experimental species, has allowed for novel applications of HDOAs, such as: the discovery of novel transcripts, mapping functionally important genomic regions and identifying functional domains in RNA molecules. Integrating all this information will provide novel global views of the locations of RNA transcription, DNA replication and the protein nucleic acid interactions that regulate these processes.

INTRODUCTION
Biological research and medicine have entered into an era where inquiries of the entire genome of an organism can be made. The promise of this genomic era, recently hallmarked by the first drafts of the near complete genomes of human1,2 and mouse,3 along with a considerable wealth of cDNA and expressed sequence tags (ESTs)4,5 has heralded the development of new approaches in experimental biology and clinical medicine. Investigators now have a unique opportunity to ask fundamental biological questions at an unprecedented scale and depth. There are few technological tools which are available to take advantage of this explosion of genetic and sequence information. One of these tools is the high density oligonucleotide array (HDOA), by which large scale hybridisation experiments can be conducted. An extensive list of publications detailing the uses of HDOAs is available (www.affymetrix.com/community/publications/index.affx). These references indicate that the majority of the applications can be broadly categorised as RNA expression monitoring, genotyping (looking for sequence variation that has previously been characterised) or resequencing (looking for novel sequence variation). These applications have helped in the understanding of the structure and function of various genomes and they also have prompted investigators to conceive of other applications that go well beyond these early uses. This review will focus primarily on the emerging and future applications of HDOAs, as well as addressing the challenges associated with these forthcoming applications.

BRIEF SUMMARY OF HDOA SYNTHESIS
One approach to synthesising HDOAs consolidates the principles of solid-phase nucleic acid synthesis and photolithography adapted from the semiconductor manufacturing arena. This approach allows for the synthesis of oligonucleotides in situ on glass substrates.6,7 Briefly, HDOAs are synthesised in the following fashion.

Thomas R. Gingeras
Affymetrix, Inc.,
3380 Central Expressway,
Santa Clara,
CA 95051, USA
Tel: +1 408 731 5069
Fax: +1 408 481 0422
E-mail: tom_gingeras@affymetrix.com

© HENRY STEWART PUBLICATIONS 1473-9550 BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 2. NO 1. 47-56. APRIL 2003 47
Photolithography and photochemistry are the key elements for the synthesis of HDOAs.

The photoprotecting group are first attached to the silanated glass substrate (Figure 1). Each array has many predefined areas where oligonucleotide synthesis occurs. The chemical protecting groups are removed by exposure to near-ultraviolet light in the area of synthesis, directed by photolithographic masks (Figure 2). Appropriate nucleotide precursors that also are modified with a photoprotecting group are then coupled to the linker in a conventional nucleotide synthesis coupling reaction. The entire procedure is then repeated (Figure 1). Each synthesis area has a homogeneous collection of millions of oligonucleotides containing the same sequence. In reality, various optimisations of the photomask design allow for the reduction of the total synthesis steps to about 75 cycles.

Arrays are not individually manufactured: they are cut out of larger glass substrates called wafers. The wafer substrate utilised in array manufacturing is a 5 x 5 inch (12.7 x 12.7 cm) piece of quartz glass that can be diced into anything from 49 to hundreds of individual arrays. After the final chemical deprotection and dicing, individual arrays are packaged into molded plastic cartridges.

The number of oligonucleotide synthesis areas on an array is determined by the photolithographic resolution of the mask. Current commercially available arrays measure 1.28 x 1.28 cm. These arrays contain >500,000 individual synthesis areas (features) of 18 x 18 µ in size. However, HDOA technology is capable of producing much smaller features and, thus, greater numbers of synthesis areas per array.

This capability to synthesise arrays of different dimensions and with different

Figure 1: Light acts as the catalyst to remove protecting groups (blocks) at defined positions. Single nucleotide washed over the array, binds where the protecting group has been removed. Through successive steps, any sequence can be built up in any position on the array. The number of steps required to build an oligonucleotide probe approximates to the probe length multiplied by 4.

Figure 2: Ultraviolet light is passed through a blocking mask with some regions open, exposing the chip in specific locations. The size of these openings delineates the size of the synthesis area at the substrate. In light-directed synthesis, chips are synthesised in combinatorial fashion, with holes in each successive mask designs (approximately four per base), determining which sequences are to be synthesised.
Entire human genome can be interrogated on one large array

Transcriptomes of higher eukaryotes are far more complex than anticipated

Table 1: Number of features as a function of array and feature dimensions

<table>
<thead>
<tr>
<th>Array dimensions (cm)</th>
<th>400 μm²</th>
<th>225 μm²</th>
<th>100 μm²</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6 × 0.6</td>
<td>9.0 × 10⁴</td>
<td>1.6 × 10⁴</td>
<td>3.6 × 10⁵</td>
</tr>
<tr>
<td>1.28 × 1.28</td>
<td>4.1 × 10⁴</td>
<td>7.3 × 10⁴</td>
<td>1.64 × 10⁵</td>
</tr>
<tr>
<td>12.7 × 12.7</td>
<td>4.0 × 10⁴</td>
<td>7.2 × 10⁴</td>
<td>1.61 × 10⁵</td>
</tr>
</tbody>
</table>

The transcriptome and whole-genome tiled arrays

The most common application for HDOAs is to estimate levels of RNA expression of the annotated transcripts in a genome. The strategy currently employed by commercial arrays to interrogate gene expression involves the use of 10–20 oligonucleotide probe pairs (25mers) per gene. Each probe pair consists of a perfect complement to some portion of the targeted transcript and a companion mismatched probe which has a destabilising mismatched nucleotide in the centre position. Such a mismatch probe serves as a negative control for cross-hybridisation. While the utility of such mismatch probes has been debated in measuring the abundance of highly expressed genes (ie >~20 copies per cell), it is generally accepted that such probe pairs are useful for the detection of transcripts present at low copy number. The estimate of the expression level of a transcript is based on the multiple measurements of the collection of probe pairs in a set. This highly redundant strategy has proven to be a successful approach and its applications have been extensively reviewed.8,9

A strategy of using HDOAs has recently been employed to discover novel transcripts. Emerging evidence indicates that the transcriptome of higher eukaryotic genomes is far larger and more complex than initially anticipated.10–14 These observations indicate that there is a need to monitor both the well-characterised portions of a genome (coding and non-coding) and unannotated parts of a genome for transcripts in an unbiased fashion. One approach to achieving this goal is to synthesise an array that contains oligonucleotide probes uniformly interrogating the entire genome of interest (Figure 3). We have recently explored this strategy for the purpose of mapping the transcriptome on human chromosomes 21 and 22. These studies indicate that there is large-scale transcriptional activity on these chromosomes that was an order of magnitude greater than could be explained by the existing annotations (well-characterised exons from Sanger, RefSeq and Genbank15). Novel transcripts could be detected far from any known transcripts, including ESTs, as well as within introns of well-
characterised genes. Many of these novel transcripts belong to the expanding family of non-coding RNAs.

The arrays used in these chromosome 21 and 22 studies were comprised of approximately 1 million probe pairs. As noted in Table 1, and later in this review, it is straightforward to construct arrays which could examine the entire human genome in an unbiased fashion at even higher resolution. It comes as no surprise that, aside from RNA transcription maps, other types of maps could be constructed using similar HDOA formats. It is these applications that represent the emerging new uses of HDOAs.

STUDYING MECHANISMS REGULATING TRANSCRIPTION
Maps of protein–DNA interactions
Chromatin immunoprecipitation (ChIP) is a technique developed to address the question of identifying in vivo occupied protein binding sites in DNA.17 Briefly, the ChIP method involves the cross linking of any bound protein to its respective nucleic acid by treating living cells with formaldehyde. The treated cells are then lysed, the chromatin randomly fragmented and the targeted DNA–protein complexes are enriched from the total cross-linked background by means of immunoprecipitation (IP). The protein component is removed by proteinase K after the cross links are reversed (by heat). The enriched nucleic acid fragments are subjected to a ligation reaction, which attaches a common set of linkers to the nucleic acid fragments. By using a common set of ligated oligonucleotide sequences, a single primer pair is used to amplify the pool or enriched fragments. The identity and genomic location of the amplified fragments can be determined by hybridising the amplified pool of DNA to an array which interrogates an entire genome in a uniform manner. Recently, a series of reports have described the assembly of binding site maps of transcription factors, of chromatin modification sites and of DNA replication...
sites using this ChIP approach in combination with microarray analysis.\textsuperscript{18–23} In one of the most impressive applications of this approach, binding sites for 106 transcription factors have been mapped along the yeast genome using arrays interrogating putative promoter regions for 6,270 yeast genes.\textsuperscript{23} In this example, each yeast strain was engineered to express specific transcription factors tagged with a $c$-myc epitope to facilitate further IP. Genome-wide analysis of the binding site data identified a number of regulatory network motifs in the yeast genome, ranging from simple transcriptional autoregulation loops when a single transcription factor regulates its own expression — also common in prokaryotes — to much more sophisticated scenarios involving several transcription factors either regulating each other or jointly regulating the expression of the same genes. Some of the regulatory configurations are novel and appear to be unique to eukaryotic genomes. Accumulation of knowledge of such network motifs will undoubtedly lead to modelling of higher order regulatory networks in the yeast genome.

Without diminishing the accomplishments of identifying the transcription factor maps within the yeast genome, it is clear that the genomic landscape of higher eukaryotes is more complex, with regulatory elements located as far as tens, and sometimes hundreds, of kilobases away from the gene of interest.\textsuperscript{24–27} These sites can reside within introns,\textsuperscript{28–30} 3' untranslated regions,\textsuperscript{31,32} and 3' flanking regions of known genes. Furthermore, the utilisation of alternative promoters\textsuperscript{33,34} is also well documented. These considerations, taken together with the emergence of an expanded and more complex transcriptional landscape in higher eukaryotic cells, present a greater challenge in deciphering the locations of regulatory regions and understanding regulatory networks. In order to identify all genome-wide binding sites for any transcription factor, it will be necessary to perform systematic interrogations of an entire genome; the examination of consensus and annotated promoter regions will not be sufficient. To achieve relative completeness, it will also be necessary to conduct this survey in a large number of developmentally diverse cell types and under a variety of physiological conditions. The final step will be to integrate these different and empirically derived maps to create a database of regulatory locations and transcripts controlled by each transcription factor.

Identification of RNA components of ribonucleoprotein complexes

After transcription, both coding and non-coding RNAs are subject to multiple intracellular processes such as splicing, polyadenylation, 5' capping, transport from the nucleus to various intracellular localisations, translation and degradation. Regulation of gene expression happens at all of these stages and is mediated by RNA binding proteins. It is not, therefore, surprising that there are more than 200 genes encoding these proteins.\textsuperscript{2} Matching specific RNA transcripts with their cognate RNA-binding proteins, however, is woefully rudimentary. Assembling such data is the first step in understanding the regulatory networks involving RNA-binding proteins. Microarrays in combination with some IP approaches have proved useful in assembling such data. Recently, RNA-binding proteins such as Nova, FMRP, Hu–B, poly(A)-binding protein (PABP) and 5' cap-binding protein eIF-4E have been shown to interact with several dozen transcripts. The arrays used in these studies monitored only a limited number of genes, however.\textsuperscript{37–40} Despite these limitations, a unique set of RNA species has been shown to be associated with each of the latter three ribonuclear proteins (RNP) complexes. Moreover, the dynamic nature of RNP interactions was demonstrated by changes observed in the RNA profiles of Hu–B RNP complexes in response to retinoic acid-induced...
cellular differentiation. Expansion of the surveillance scope of the sequences which could be interrogated for possible RNP interactions, with concomitant increase of the biological context of these experiments, will yield a more comprehensive picture of the regulatory roles played by RNA binding proteins in eukaryotic cells.

Subcellular localisation of RNAs
As previously noted, the activities of RNAs are frequently correlated with subcellular localisation. The identification of RNA species associated with polysomes illustrates yet another interesting application of microarray technology. For example, it is often desirable to know whether observed global changes in steady-state levels of mRNAs are also accompanied by changes in the abundance of the corresponding proteins. RNA profiling of the polysome fraction can potentially yield information on which RNAs are present at the precise sites used for protein synthesis. This approach has been used in a number of studies. In one example, the polysome profiles were compared in cells with and without functional fragile X mental retardation protein (FMRP), the product of the fragile X gene. In this study, 282 mRNAs, whose association with polysomes changed in the absence of FMRP, were identified. The overall expression level of the majority of these mRNAs (251 out of 282) did not change when compared with normal cells. These observations led the authors to suggest that the pathogenesis of fragile X syndrome is related to an aberration in the translation of these messages. In another example, RNA profiling of the polysomes associated with intracellular membranes was also used to identify 285 and 275 genes in yeast and human cells, respectively, which are likely to encode secreted or membrane-bound proteins.

Probing DNA methylation
DNA methylation has been widely implicated in controlling gene expression in various eukaryotic cells. Typically, an increase in DNA methylation correlates with compacted chromatin and the reduction of gene expression. Aberrations in DNA methylation have also been linked to tumourigenesis. Several groups have successfully used microarray technology to map methylation patterns. These methods rely on either differential digestion of DNA using methylation-sensitive restriction enzymes or bisulphite sequencing. In one of these studies, genome-wide profiles of DNA methylation have been compared in wild type controls and Arabidopsis chromomethylase 3 (CMT3) and MET1 DNA defective mutants. This assay involves digestion of DNA with a methylation-sensitive tetranucleotide-recognising restriction enzyme (e.g. Msp I), purification of the fraction containing DNA fragments of <2.5 kb, followed by hybridisation to microarrays. A dual-colour labelling system was used to determine the differences between mutant and wild type DNAs. Each DNA methylase mutant had a unique, but overlapping, pattern of abnormal DNA methylation. One interesting finding of this study was the identification of retroposon-like sequences, both repetitive and single-copy, as methylation targets of CMT3. Seven out of eight regions shown to be hypomethylated in cmt3 null Arabidopsis mutants displayed sequence similarity to retroposon-like elements; one of these regions was a promoter of a known gene. Changes in the methylation patterns for the retroposon-like regions also correlated with the changes in gene expression observed in cmt3 mutants.

One possible challenge to the interpretation of such experiments is the heterogeneous nature of epigenetic events in a sampled cell population. Mosaicism of DNA methylation events has been well documented. In a complex sample, the methylation status of a DNA region will be an average of the methylation states of that region in all cells in that sample.
Thus, only the loci with a consistent methylation pattern can be readily evaluated. Another potential challenge is the allele-specific pattern of gene methylation, i.e., taking place in the imprinted loci. In this case, a methylation status of a region will represent the average of both alleles unless certain experimental procedures are undertaken to separate the input of each allele.

DNA replication

Until recently, data describing the origins of DNA replication were obtained from studies involving a limited number of genomic regions. A genome-wide analysis can be approached by the use of HDOAs. Studies mapping the timing of DNA replication in the yeast and fly genomes using arrays have recently been reported. In these experiments, the replicated DNA was labelled in vivo either with isotopes and separation on density gradients or with bromodeoxyuridine (BrdU) and then IP with anti-BrdU antibody. In the yeast experiments, DNA was biotinylated and hybridised to HDOAs representing every open reading frame in the yeast genome. The authors mapped 332 origins of replication and carried out extensive analysis concerning specific aspects of DNA replication, including the velocity of DNA fork migration and replication through heterochromatic and euchromatic domains. Interestingly, previously hypothesised links between levels of RNA transcription and timing of DNA replication were not experimentally confirmed. These data contrast with the study of replication in the fruit fly, where replicated and non-replicated DNAs, labelled with either Cy3 or Cy5, were hybridised to microarrays containing 6,500 cDNA probes. A strong correlation between RNA transcription and DNA replication was detected. Highly transcribed sequences had a greater chance of being replicated earlier in the S phase, suggesting a unique connection between transcription and DNA replication in higher eukaryotes.

Even though microarrays can be used for these types of genome-wide surveys of DNA replication in higher eukaryotes, such studies can be confounded by the heterogeneity of the replication events taking place even in synchronised cell populations. Regions that consistently replicate early or late in the S phase are most likely to be detected. For example, in the Drosophila study, the difference in the intensities between early or late replicated DNA was relatively small (within a factor of two) for the majority of the assayed sites in the genome. Some of the possible interpretations of this observation are: (1) a relatively weak preference in the timing of replication for a large fraction of the genome, with a few notable exceptions, i.e., certain classes of transposons; (2) there is a timing preference, but it varies in different cells; or (3) a combination of both. Such issues provide a significant challenge to the interpretation of array studies of DNA replication.

CONCLUSION

In the next few years, the complete genome sequences of multiple experimental organisms will be available. Additionally, multiple isolates of such genomes will also be sequenced in order to acquire some understanding of the common sequence variation that exists within these genomes. Initially, an informatics-based analysis of these assembled sequences will provide a basic structure for both coding and regulatory regions. However, significant advances in our understanding of how genetic information is encoded, regulated and translated in each organism will ultimately require experimental evidence. HDOA technology is an important tool that will be used to acquire this information. The evolution of this technological approach will parallel the biological questions that are being investigated.

The monitoring of RNA expression levels is an example of how HDOA...
The design of HDOAs will depend on a given genomic scale at which a biological question is being addressed.

technology has evolved to meet the needs of the biological questions posed. If a biological question is focused on specific regions of a genome (e.g. well-characterised genes and/or large EST clusters), then the number of probes of an HDOA can be reduced. The results of these experiments will be based on the fact that the interrogated sequences are already known to be transcribed regions. As the direction of investigations changes towards more complex genomic regions (singleton ESTs, splice variants of genes, etc.) or towards less annotated genomic regions, a more uniform and unbiased survey of the entire genome will become an even more important approach. HDOAs will be able to conduct $10^7$ to $10^9$ individual hybridisation experiments per array. Thus, the future of microarrays will be driven, in part, by the direction of their experimental applications.

Expanded utilisation of HDOAs, which includes monitoring the binding sites of transcription factors, the targets of RNA-binding proteins, the sites for replication origins and subcellular localisation of RNAs, indicates even greater opportunities for evolution of not only microarray technology but the front-end assays that prepare targets for array analysis. More than several hundred transcription factors and RNA-binding proteins have been identified in the human genome. Many of these transcription factors and RNA-binding proteins will be monitored in a wide spectrum of cell types and physiological conditions. Importantly, the integration of the data collected from various mapping experiments will result in a synergistic dataset which will increase our understanding of how information is transferred from the genome and how the genome itself is regulated.

References


et al. (2003), ‘Database resources of the National Center for Biotechnology’, NuclAcid Res., Vol. 31, pp. 28–33.
42. Ng, H. H. and Bird, A. (1999), ‘DNA


