Chromosome Research **13**: 259–274, 2005.

Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery

Paul Bertone[1], Mark Gerstein[2] & Michael Snyder[1,2]
[1]*Department of Molecular, Cellular, and Developmental Biology Yale University, New Haven, CT 06520-8103, USA; Tel: +1(203) 432-5405; Fax +1(203) 432-5175; E-mail: paul.bertone@yale.edu;*
[2]*Department of Molecular Biophysics and Biochemistry Yale University, New Haven, CT 06520-8114, USA*

Key words: chromatin immunoprecipitation, gene expression, microarrays, regulatory network, transcript mapping

Abstract

Microarrays have become a popular and important technology for surveying global patterns in gene expression and regulation. A number of innovative experiments have extended microarray applications beyond the measurement of mRNA expression levels, in order to uncover aspects of large-scale chromosome function and dynamics. This has been made possible due to the recent development of tiling arrays, where all non-repetitive DNA comprising a chromosome or locus is represented at various sequence resolutions. Since tiling arrays are designed to contain the entire DNA sequence without prior consultation of existing gene annotation, they enable the discovery of novel transcribed sequences and regulatory elements through the unbiased interrogation of genomic loci. The implementation of such methods for the global analysis of large eukaryotic genomes presents significant technical challenges. Nonetheless, tiling arrays are expected to become instrumental for the genome-wide identification and characterization of functional elements. Combined with computational methods to relate these data and map the complex interactions of transcriptional regulators, tiling array experiments can provide insight toward a more comprehensive understanding of fundamental molecular and cellular processes.

Introduction

It is widely recognized that the availability of a complete genome sequence can significantly enhance our ability to analyse biological phenomena and elucidate molecular and cellular function. Beyond the initial determination of the DNA sequence, the most valuable resource produced by genome mapping efforts entails a comprehensive catalogue of functional elements that encompass the genetic repertoire of an organism. Methods for the global analysis of gene expression include subtractive hybridisation (Hedrick *et al*. 1984), differential display (Liang & Pardee 1992), and representational difference analysis (Hubank & Schatz 1994). While these techniques are useful for characterizing differences in mRNA transcript populations, they are unable to generate comprehensive gene expression profiles.

The genome-wide identification of transcribed sequences was made possible with the development of the SAGE (serial analysis of gene expression) technique (Velculescu *et al*. 1995). SAGE enables the quantitative estimation of mRNA expression levels by sampling short (10–14mer) sequences of transcribed messages, and using these to deduce the identity of the specific transcripts from which they are derived. The advantages of

this approach are two-fold: First, it is not necessary to use a unique hybridisation probe to detect each individual transcript; second, multiple SAGE tags may be concatenated and sequenced together, providing several measurements simultaneously. A caveat inherent in the SAGE technique is that the use of relatively short sequence tags can result in ambiguous transcript identification. This deficiency can be overcome by using 200–600 nt expressed sequence tags (ESTs). Although EST methods predate SAGE technology, they afford a higher degree of specificity and can produce long stretches of transcribed sequence (Adams *et al.* 1991).

DNA microarrays are by far the most widely adopted platform for the high-throughput analysis of gene expression. The advent of cDNA (Schena *et al.* 1995, DeRisi *et al.* 1997), inkjet (Shoemaker *et al.* 2001, Hughes *et al.* 2001) and oligonucleotide (Fodor *et al.* 1993, Pease *et al.* 1994, Lockhart *et al.* 1996, Lipshutz *et al.* 1999) arrays has allowed researchers to simultaneously monitor the expression levels of thousands of genes in a single experiment. The cDNA format consists of mechanically deposited PCR products representing the entire coding sequence of annotated genes. Oligonucleotide arrays (e.g. Affymetrix GeneChips) typically contain one or more complementary oligomer sequences internal to spliced mRNA transcripts, generally positioned near the 3′ end to ensure hybridisation to incomplete cDNAs.

While all of these approaches provide the ability to measure genome-wide expression levels of annotated genes, only when a complete corpus of transcribed sequences has been defined can they be exploited to their full potential. Once an organism's complement of transcribed sequences is known, high-throughput analysis methods can be used to comprehensively investigate the dynamics of gene expression over the entire transcriptome.

Challenges in genome annotation

The early characterization of genes from prokaryotes and model eukaryotes revealed simple gene structures consisting almost entirely of protein-coding sequences. For these organisms, there usually exists a one-to-one relationship between the open reading frames (ORFs) that delineate transcribed sequences and the proteins they encode. In contrast, the genome sequences of higher eukaryotes tell a far different story. Here the predominant gene structures are often fragmented, largely due to the widespread integration of repetitive elements. The transcribed regions of larger, more complex genomes typically embody many short exons interspersed with long intron sequences. The separation of coding sequences into discrete units provides the opportunity for additional genetic variation through the mechanism of alternate splicing. Through selective exon usage, many different protein isoforms may arise from a single gene, greatly amplifying the potential coding complexity of the genome. This is particularly true in mammals, where a typical gene may consist of dozens of exons and various combinations of these might be included in spliced messages expressed in different cell types or under different environmental conditions. Thus, a single gene may give rise to a family of protein products that confer a wide range of functional roles. This mechanism is believed to account for the disproportionate increase in organismal complexity in relation to the number of genes it encodes.

Given the fragmented nature of mammalian genes, predicting coding regions from genomic DNA has proven a difficult computational challenge. Some introns may exceed tens of kilobases in length, making it difficult to aggregate the much shorter coding sequences they divide into plausible gene structures. As a result, many genomes are annotated through homology to characterised protein sequences from evolutionarily related organisms. However, this approach is inherently biased in that the putative genes identified through sequence similarity must, by definition, be related to genes that are already known. The discovery of unique or highly divergent transcribed sequences is therefore precluded by this approach. Further, the problem of identifying non-coding RNA transcripts is largely neglected by current homology-based prediction methods.

Experimental methods of determining full-length mRNA sequences usually involve the cloning and sequencing of cDNA collections (Adams *et al.* 1991, Strausberg *et al.* 1999, Kawai *et al.* 2001, Ota *et al.* 2004). Once identified, cDNAs can be mapped onto the genome based on sequence similarity to yield a preliminary annotation of

expressed gene structures. While this approach captures a wealth of information about genes transcribed under specific cellular conditions, it often fails to identify rare splice variants or messages expressed in low abundance. Additionally, 5′ ends of genes may be under-represented due to the low fidelity of the viral polymerases used to reverse-transcribe polyadenylated RNA.

Although various techniques such as primer extension and 5′ RACE can be used to more precisely map transcriptional start sites, these methods are difficult to implement in a high-throughput manner. To address this problem, Marayuma and Sagano (1994) developed a protocol for ligating a primer to the modified 5′ ends of RNA transcripts, thereby providing a template sequence from which to amplify the message for more accurate sequencing. The group went on to generate full-length cDNAs for the entire RefSeq collection (Pruitt *et al*. 2003), revising over one-third of the existing sequences (Suzuki 1997).

Empirical discovery of novel transcribed sequences

The first microarray experiments designed to address the problem of gene annotation were performed with the *E. coli* genome. Selinger *et al*. (2000) developed an oligonucleotide tiling array to represent the genome sequence at 30 bp resolution, using the array for both transcript mapping and differential expression analysis. Nearly all of the annotated sense-strand ORFs were detected as well as 3000–4000 antisense ORFs. Even though the genome of *E. coli* is among the best studied, subsequent microarray analysis by Tjaden *et al*. (2002) revealed a 25% increase in the number of transcriptional units detected beyond those previously annotated.

The level of transcriptional activity detected within unannotated regions of genomic DNA appears to increase with the size and complexity of the genome in question. Recently, the entire genome of the flowering plant *Arabidopsis thaliana* was surveyed using oligonucleotide array technology. Yamada *et al*. (2003) developed a series of 12 tiling arrays to characterize transcriptional activity in four complex tissue RNAs, producing the first comprehensive expression map of a eukaryotic genome. Many transcribed sequences were detected within intergenic regions devoid of existing gene annotation, and approximately 30% of antisense transcription was found to be coincident to sense-strand coding regions.

Tiling arrays have also been used for global expression analysis of the fruit fly, *Drosophila melanogaster*. Stolc *et al*. (2004) used maskless photolithographic DNA synthesis (Nuwaysir *et al*. 2002, Albert *et al*. 2003) to fabricate oligonucleotide arrays representing all of the predicted exons and exon splice junctions, as well as intergenic and intronic regions throughout the genome. RNA transcript levels were measured at six developmental stages in the organism's life cycle, profiling the expression levels and splice variation of known genes but also revealing the presence of novel transcribed sequences. Comparison with the *Drosophila pseudoobscura* genome indicated that transcriptionally active sequences within unannotated regions exhibit a greater degree of sequence conservation than those for which transcription was not observed.

The use of tiling arrays for human genome annotation has met considerable technical challenges, mainly due to the large size of mammalian genomes. As part of a study involving inkjet oligonucleotide arrays to survey annotated exon usage in the human transcriptome, Shoemaker *et al*. (2001) developed a tiling approach to accurately map the coding sequence of a novel transcript located within a 113 kb locus of chromosome 22. Although this analysis was carried out on a limited scale, the results clearly illustrated the value of using tiling arrays to delineate transcript boundaries, exon content and splice junctions.

The first tiling array developed to cover the sequence of an entire human chromosome was described by Kapranov *et al*. (2002). In this study, a series of oligonucleotide arrays representing all non-repetitive DNA on chromosomes 21 and 22 was interrogated with cytosolic polyadenylated RNA from 11 cell lines. Surprisingly, a roughly two-fold increase in transcribed DNA was measured over that predicted by existing gene annotation. This finding was reproduced by Rinn *et al*. (2003) using a microarray representing all non-repetitive DNA of chromosome 22 with approximately 21 000 PCR products. Transcriptional activity was measured across the chromosome in

normal placental tissue RNA, followed by strand-specific hybridisation of novel transcribed sequences to a contact-printed oligonucleotide array.

Recently, Bertone *et al.* (2004) constructed a series of 134 high-resolution oligonucleotide arrays representing both sense and antisense strands of the entire human genome, synthesising nearly 52 million 36nt probe sequences via maskless photolithography. Hybridisation to polyadenylated liver tissue RNA revealed over 10 000 new transcribed sequences and verified the transcription of nearly 13 000 predicted genes. A large fraction of novel transcripts exhibited a high degree of similarity to the mouse genome and other mammalian protein sequences, suggesting they may be functional on the basis of evolutionary conservation. Approximately 11% of the unannotated transcriptional units were found to intersect retroprocessed pseudogenic sequences identified in previous studies (Harrison *et al.* 2002, Zhang *et al.* 2003). A small number of these were not determined to be homologous to other annotated genes, decreasing the likelihood of cross-hybridisation and indicating that some of the detected pseudogenes may be transcribed. Many other novel transcribed sequences are presumed to correspond to exons retained in rare splice variants, under-represented UTRs of annotated genes, protein-coding transcripts expressed in low abundance and non-coding RNAs. All three human transcript mapping studies identified previously unannotated transcription units located distal to known genes, indicating they originate from distinct messages (Figure 1).

The studies by Rinn *et al.* and Bertone *et al.* measured differential hybridisation of RNA to sense and antisense strands of transcriptionally active regions and the entire genome, respectively. In both experiments, strand-specific transcription was detected antisense to annotated gene components, notably introns. The initial transcriptome analysis of chromosomes 21 and 22 by Kapranov *et al.* (2002) interrogated one strand of genomic DNA with double-stranded cDNA, and therefore could not discern the strand from which transcription originated. However, a subsequent study by Kampa *et al.* (2004) used end-labelled RNAs to obtain strand-specific information, finding 11% of novel transcription to occur antisense to annotated coding sequences and 50% of transcription within intron regions to originate from the antisense strand, consistent with the previous studies.

Using a computational approach to select regions where antisense transcription may occur, Yelin *et al.* (2003) conducted a microarray survey of 2667 sense–antisense sequence pairs to assay for strand-specific transcription. Hybridisation to RNA from 19 cell lines and four normal complex tissues confirmed transcription on both strands for 60% (1600) of the sequences interrogated. A subset of these were confirmed by Northern blot hybridisation to strand-specific RNAs, confirming the detection of endogenous natural antisense transcripts (NATs). Together, these findings reinforce an emerging view of widespread antisense RNA transcription throughout the human genome. The repeated identification of novel transcribed sequences by several independent research studies provides compelling evidence of a complex transcriptome encompassing novel protein-coding regulatory and structural RNAs that have previously eluded detection by conventional genetic approaches (Mattick 2003, 2004, Johnson *et al.* 2005).

Global identification of regulatory elements

Transcription factors are regulatory proteins that bind DNA to modify chromatin or recruit components of the transcriptional apparatus, ultimately manifesting or repressing the expression of their target genes. Identifying the genes regulated by an organism's complement of transcription factor proteins is central to our understanding of diverse cellular processes. It is therefore highly desirable to attain a comprehensive inventory of the *cis*-regulatory sequences that constitute the promoter elements to which a given transcription factor binds. Although the *in-vitro* DNA-binding sequences of many factors have been established to varying degrees of accuracy, *in-vivo* binding can be affected by a multitude of complex determinants. These include variations in local chromatin structure and accessibility, interaction of transcriptional activators with remote enhancer elements, and involvement of ancillary proteins. Thus, a given factor can bind to different locations *in vivo* to coordinate the transcription of different sets of genes, depending on the cellular conditions in which it is expressed.

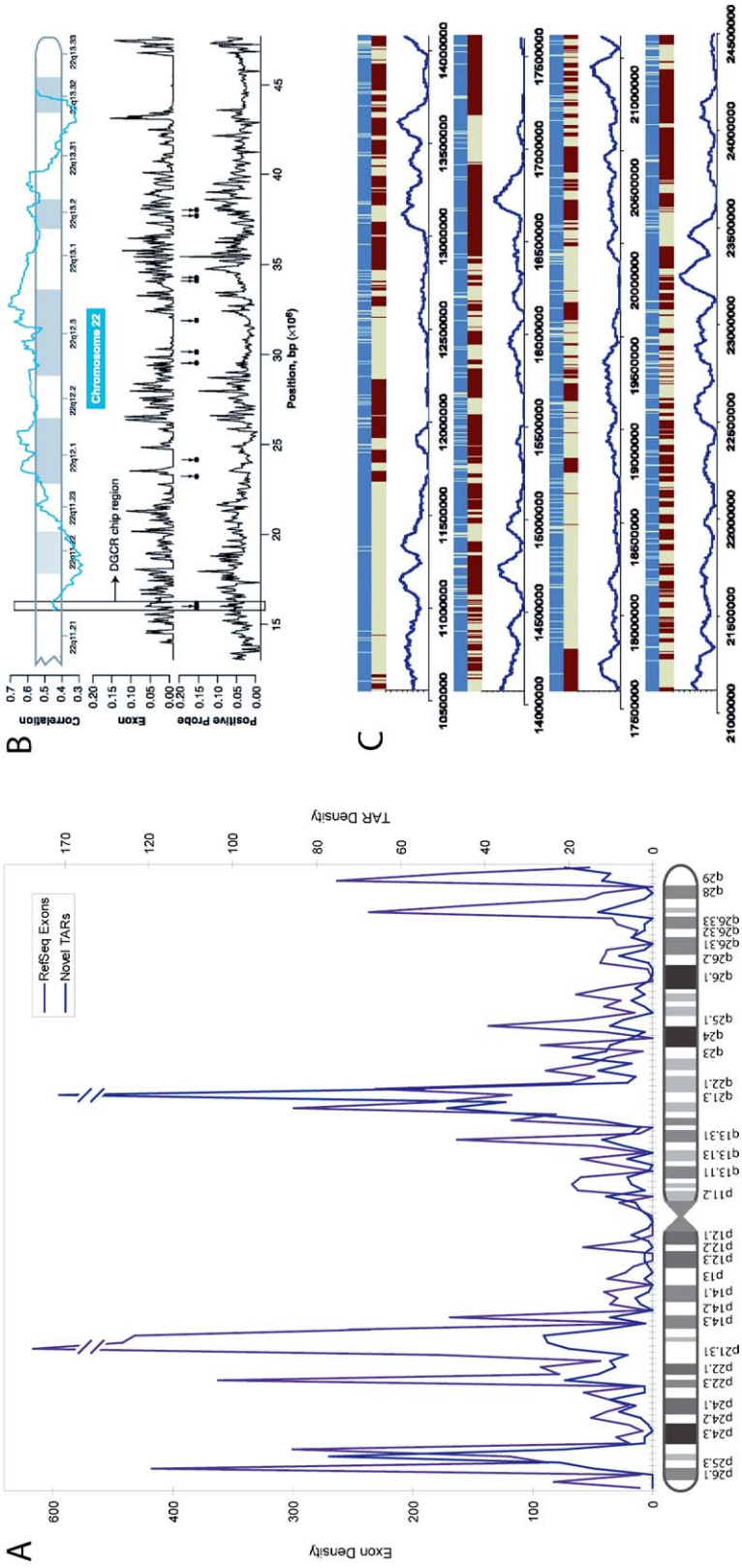

Figure 1. Transcriptional profiling of the human genome using three different tiling array platforms. (**A**) The distribution of transcriptionally active regions (TARs) is coincident with annotated exon density on a global scale, as illustrated on human chromosome 3 (Bertone *et al.* 2004). (**B**, **C**) Unbiased surveys of human chromosome 22 using oligonucleotide (**B**, Kapranov *et al.* 2002) and PCR (C, Rinn *et al.* 2003) tiling arrays reveal evidence of RNA transcription originating from previously unannotated regions as well as known genes.

Until recently, characterising the interactions of DNA-binding proteins with the genome was possible only on a single-gene basis, primarily through *in-vivo* footprinting studies.

In concert with other experimental protocols, microarrays can now be used to study the behaviour of transcriptional activators in a manner analogous to gene expression analysis. Since functional binding sites of transcription factors are expected to occur primarily within intergenic regions, microarray-based analyses of gene regulation have emerged with the tandem development of genomic tiling arrays. In shifting the selection of DNA sequences away from the exclusive representation of genes, tiling arrays facilitate the unbiased mapping of transcription factor binding sites on an unprecedented scale.

The most widely adopted procedure involves the hybridisation of chromatin immunoprecipitated (ChIP) DNA to a genomic DNA tiling array, commonly referred to as ChIP-chip (Horak & Snyder 2002, Lieb 2003). In this approach, protein–DNA interactions in cells expressing the factor of interest are fixed *in situ* with a crosslinking agent, typically formaldehyde (Solomon & Varshavsky 1985). Nuclear extracts are isolated and the transcription factor is immunoprecipitated, either with antibodies against the native protein or via an epitope tag fused to the transcription factor gene. The crosslinks are reversed with heat treatment and fluorescence-labelled samples are prepared from the transcription factor-bound DNA following the purification and sonication of the immunoselected chromatin fragments (Figure 2).

The labelled DNA is then hybridised to a microarray in parallel with a negative control sample. This can be derived from genomic DNA or consist of an identical sample precipitated either in the absence of antisera or with pre-immune sera. The resulting data can therefore be treated like those generated by a two-channel differential gene expression experiment, where fluorescence intensity ratios are computed after normalising the signals from the two channels (Figure 3A). The main analytical difference between a differential expression experiment and a ChIP-chip experiment is that, in the latter case, statistical outliers are expected to occur only in the fluorescence channel corresponding to the immunoprecipitated sample (Figure 3B). A significant increase in fluorescence intensity therefore corresponds to the enrichment of a specific population of DNA fragments in excess of those represented in the control sample, and are assumed to have hybridised to chromatin fragments containing transcription-factor-bound sequences.

Once identified, transcription-factor-bound sequence fragments can be mapped to their genomic loci and their positions compared with existing gene annotation (Figure 4). The total number of DNA fragments enriched via immunoprecipitation is usually a superset of those involved in gene regulation. Some factors recognize highly specific promoter sequences and associate with chromatin infrequently, while others may bind constitutively to many sites throughout the genome. A number of transcription factors have been observed to bind to promoter regions in clusters, such that several binding events constitute a smaller number of regulatory loci. Additionally, since the immunoprecipitated chromatin fragments are double-stranded, either strand of the denatured sample becomes available to anneal with complementary array sequences. It is therefore impossible to distinguish on which strand the factor's promoter sequence lies from this experiment alone. Instead, one must consider both strand orientations equally when analysing the data, observing the proximity of binding sites to annotated genes to determine which are likely to be involved in regulatory function.

Because transcription factor binding alone does not necessarily indicate the locations of functional promoters, evidence to support regulatory function must be accumulated by integrating other experimental data. Differential gene expression, easily observed through microarray analysis, can reveal which genes are affected in response to the stimuli under which a transcription factor is induced. This information is superimposed with binding site data to reveal where DNA binding occurs on the chromosome relative to the locations of differentially expressed genes (Figure 4B). Ultimately, careful consideration is required to interpret the results of these experiments in a biologically meaningful way.

The ChIP-chip approach was first explored in the yeast model. Ren *et al*. (2000) used a microarray of PCR products representing 6361 yeast intergenic regions to map the genome-wide binding locations

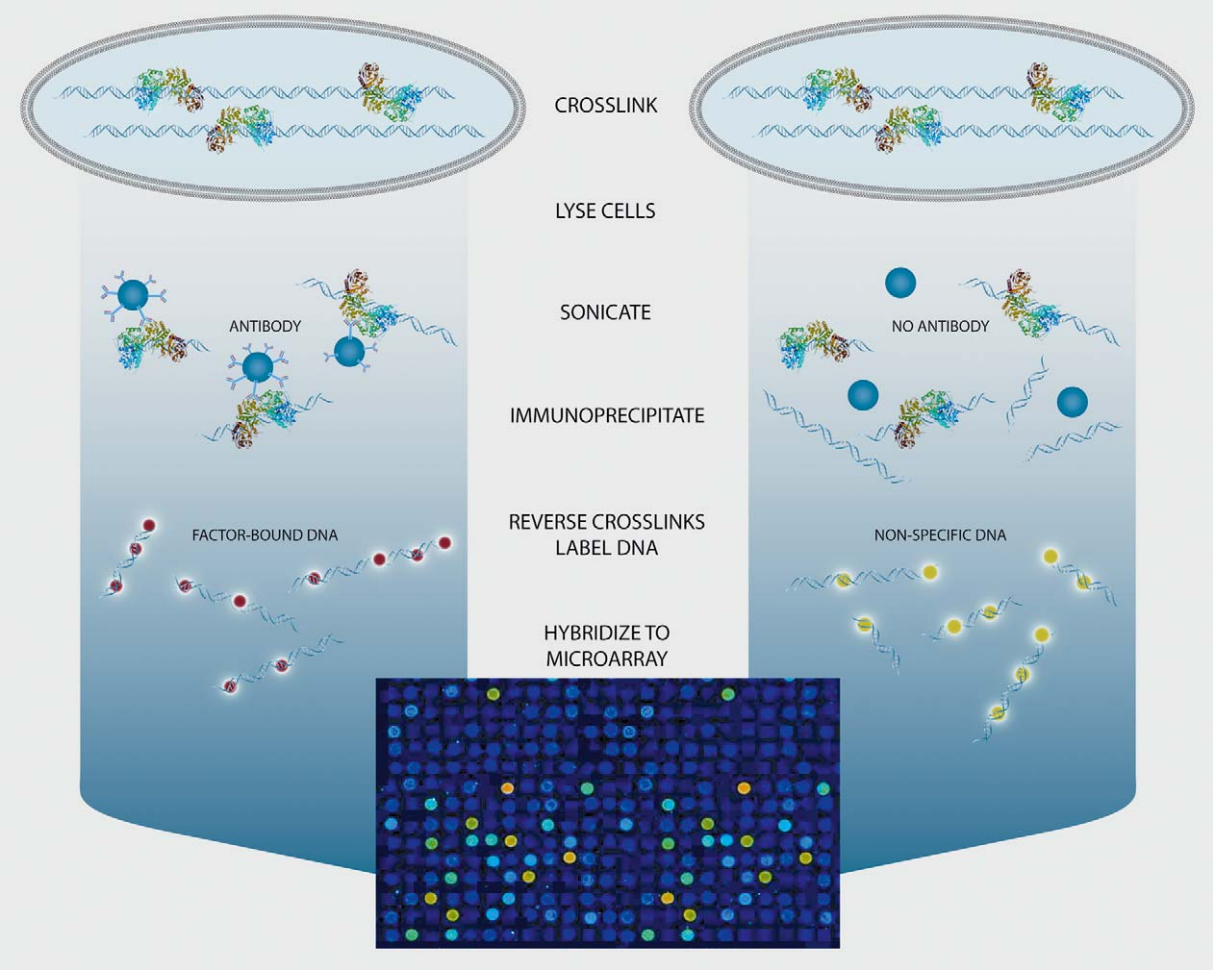

Figure 2. ChIP-chip protocol for microarray-based chromatin profiling. Protein–DNA interactions within cells expressing a transcription factor of interest are treated with formaldehyde to promote *in-vivo* crosslinking. This is followed by lysis, shearing of the genomic DNA, and immunoselection of protein–DNA complexes from nuclear extracts using antibodies against the transcription factor. The immunoprecipitated DNA is purified, fluorescence-labelled and hybridised to a tiling or intergenic microarray in parallel with a negative control sample. The control may be derived either from immunoprecipitations performed in the absence of antibodies or with control antibodies, from a deletion strain or cell line, or from genomic DNA.

of Gal4 and Ste12. Their analysis revealed 3 novel gene targets in addition to those previously known to be regulated by Gal4, and 29 genes specifically regulated by Ste12. Shortly thereafter, Iyer *et al.* (2001) developed a similar approach, constructing a PCR-product array of approximately 6700 intergenic and promoter regions to map the genome-wide binding locations of the transcription factors SBF and MBF during the G1/S transition of the mitotic cell cycle. They identified over 200 genes regulated by the factors, finding SBF and MBF to be implicated in cell wall biogenesis and DNA replication, respectively. Lieb *et al.* (2001) then used the ChIP-chip method to map the binding sites of Rap1, previously associated with telomere modification and mating-type transcriptional repression. As an essential gene, mutations to Rap1 that affect DNA binding are lethal and thus the regulatory characterization of the factor is recalcitrant to conventional genetic analyses. Binding site location analysis identified approximately 5% of yeast genes regulated by Rap1, implicating the factor in key cellular processes, such as protein biosynthesis and energy metabolism.

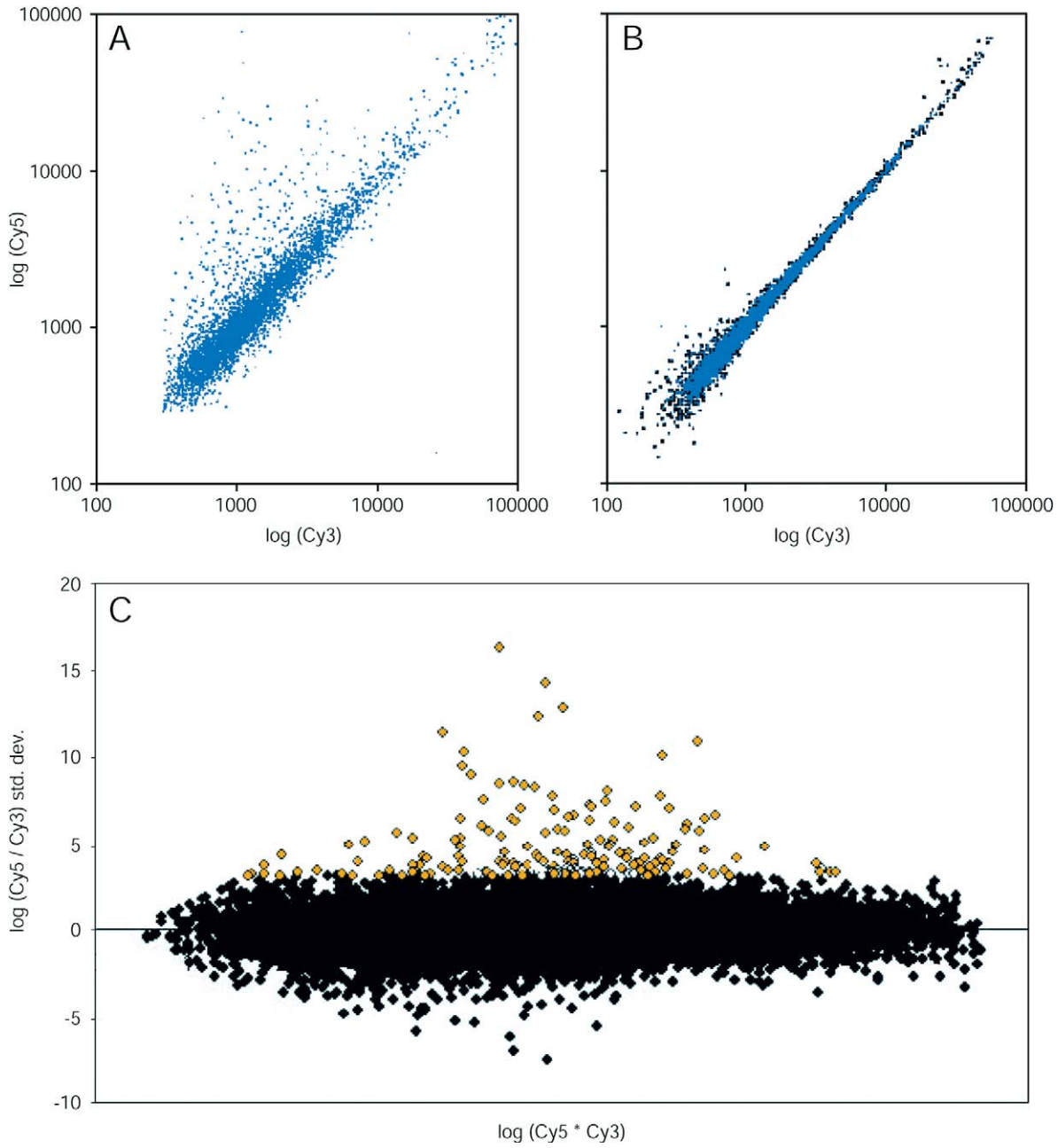

Figure 3. Scatter plots of chromatin-immunoprecipitated DNA versus a negative control sample (**A**), and a negative-versus-negative control experiment (**B**). As seen in the first example, the enrichment of transcription factor-bound DNA produces an increase in fluorescence intensity at hybridizing microarray features. (**C**) Statistical outliers are typically identified as features whose $\log_2$ intensity ratios exhibit fold change increases above a given threshold or exceed several standard deviations from the normalized intensity distribution (Quackenbush 2002, Luscombe *et al.* 2003).

An advantage of these experiments is that since protein–DNA interactions are fixed *in vivo*, the experiment can be performed under varying cellular conditions to assess regulatory activity in different environmental contexts. For example, the Ren *et al.* study (2000) measured enrichment in

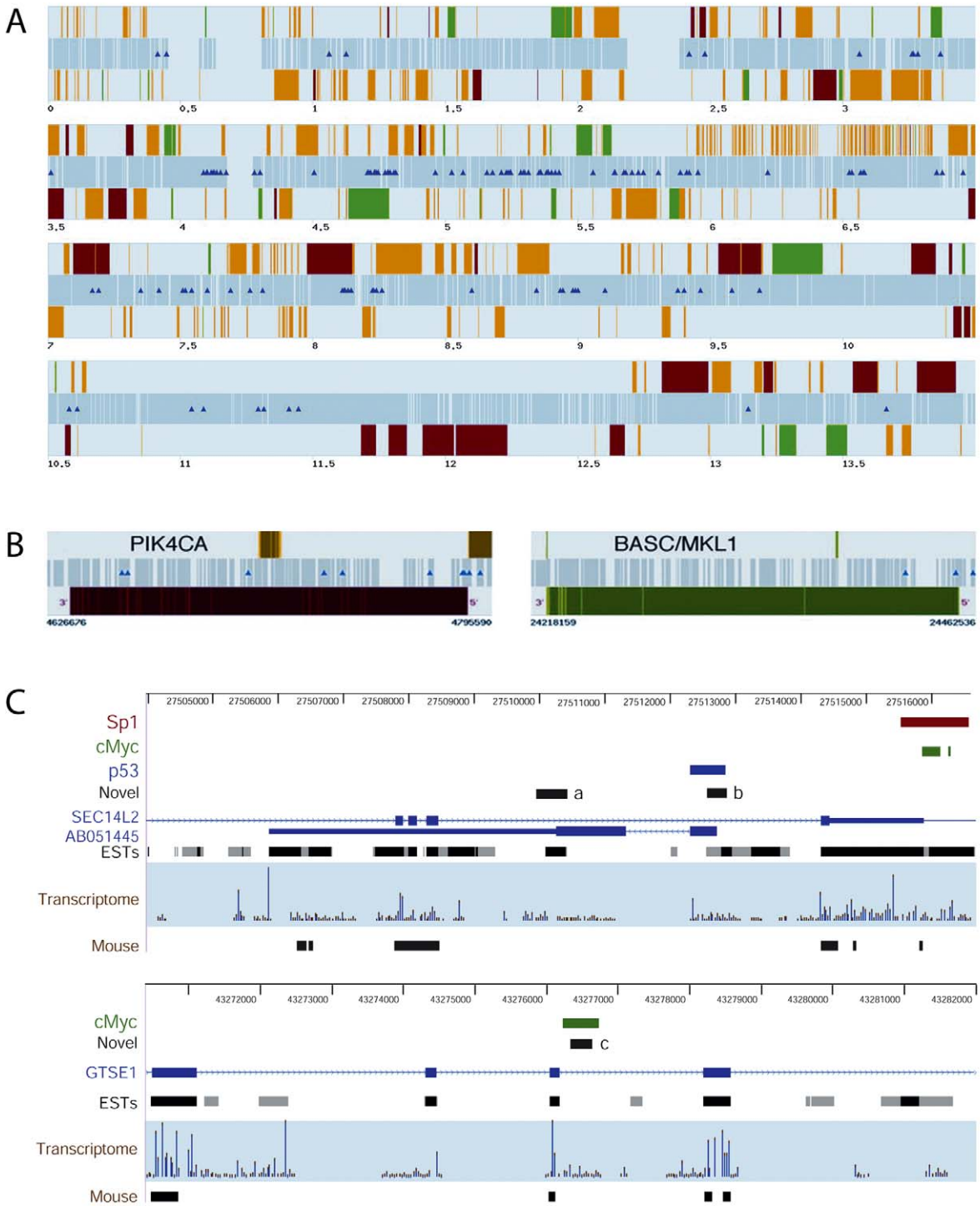

Figure 4. (**A**) Binding distribution of CREB over a segment of human chromosome 22 illustrating transcription factor binding within coding and intergenic regions as well as clusters of binding sites upstream of annotated genes (Euskirchen *et al.* 2004). Binding sites are marked as blue triangles across the chromosome; up-regulated, down-regulated, and non-differentially expressed genes appear in red, green and yellow, respectively. (**B, C**) Examples of NF-κB (**B**), Sp1, c-Myc and p53 (**C**) binding adjacent to differentially expressed genes on chromosome 22 (Martone *et al.* 2003, Cawley *et al.* 2004). Although some DNA-binding sites are located in or near canonical promoter regions 5′ of annotated genes, others lie in gene-dense regions where a single regulatory element may control the expression of multiple targets (**B**) as well as novel transcribed sequences (**C**; unannotated transcription units are labelled a–c).

transcription factor binding site occupancy in response to changes in carbon source and mating pheromone, comparing the proximity of these sites to genes whose expression levels changed under similar conditions (i.e. genes whose promoters were bound by Gal4 and induced in galactose, and those bound by Ste12 and induced by α factor).

ChIP-chip analysis was first extended to a mammalian system by Horak *et al.* (2002a). Using a PCR product microarray representing the 75-kb human β-globin locus, the binding distribution of the haemapoietic lineage-specific transcription factor GATA-1 was measured in erythroleukaemic K562 cells. Only a single region within the β-globin locus had been previously known to contain GATA-1 binding sites; however, the factor was observed to bind a region upstream of the γG gene in addition to confirming the results of previous observations. Ren *et al.* (2002) developed a promoter-proximal microarray containing PCR-amplified genomic loci directly upstream of 1444 human genes. The array was used to identify ChIP-enriched sequences bound by the transcriptional activator E2F1 during the G1/S phase transition of the cell cycle, and the repressor E2F4 during quiescence.

Subsequent to these studies, ChIP-chip has been used to survey transcription factor binding over entire human chromosomes. Martone *et al.* (2003) mapped the binding distribution of NF-κB (p65) across chromosome 22 in HeLa cells induced in the presence of tumour necrosis factor (TNF-α). Using the same microarray platform, Euskirchen *et al.* (2004) investigated CREB binding in the cAMP-inducible JEG-3 choriocarcinoma cell line. Both studies revealed a wide distribution of binding sites across the chromosome relative to annotated genes (Figure 4A). Particularly interesting was the finding that many binding sites are located proximal to 3′ ends of genes and within annotated introns, challenging the traditional view that transcription factors act exclusively in promoter regions directly upstream of transcriptional start sites. Using oligonucleotide arrays, Cawley *et al.* (2004) surveyed the binding of c-Myc, Sp1 and p53 in Jurkat and HCT1116 cells over chromosomes 21 and 22. As was reported in the NF-κB and CREB studies, transcription factor binding was observed at many locations upstream of 5′ ends, proximal to 3′ ends, and internal to genes (Figure 4C). Coincident binding of Myc and Sp1 was also found to occur at numerous locations, suggesting the possibility that some of their target genes are coregulated by the two factors.

Unlike the yeast experiments which employed an intergenic array to assess transcription factor binding, the human chromosome studies surveyed binding over all the non-repetitive DNA in an unbiased fashion. The arrays were designed to represent both coding and intergenic regions irrespective of existing gene annotation, as was the case for the previous chromosome-wide surveys of RNA transcription. In comparing the locations of enriched ChIP fragments to annotated genes as illustrated in Figure 4A, it becomes clear that a complete representation of the genome sequence is required to fully characterize the binding distribution of a given transcription factor. Although ChIP-chip experiments performed with arrays that represent promoter-proximal regions (Ren *et al.* 2002, Li *et al.* 2003, Gao *et al.* 2004, Odom *et al.* 2004) or CpG islands (Weinmann *et al.* 2002, Mao *et al.* 2003, Wells *et al.* 2003) can provide a wealth of valuable information about transcription factor association with canonical regulatory loci, the resulting data is likely to be incomplete. A given factor may bind alternative promoters, remote enhancers or other locations that are quite distant from transcriptional start sites. This is a particularly significant issue when such experiments are applied to mammalian genomes, which exhibit an unusually small percentage of annotated coding sequence relative to the amount of intergenic DNA.

An alternative technique developed to analyse DNA binding in *Drosophila* is known as DNA adenine methyltransferase identification, or DamID (van Steensel & Henikoff 2000, van Steensel *et al.* 2001). In this approach, a transcription factor gene is fused to *E. coli* DNA adenine methyltransferase (Dam), which methylates the N6 position of the adenine nucleotide in the sequence GATC. Methylation will occur at or around these sites *in vivo*, marking the locations of transcription factor binding. The genomic DNA is then subjected to DpnI endonuclease digestion and unmethylated chromatin fragments are removed by incubating with DpnII. The remaining DNA is then amplified, labelled and hybridised to a genomic DNA tiling array. Sun *et al.* (2003) used this technique to map

the DNA-binding locations of GAF and the heterochromatin protein HP1, using a PCR-product tiling array representing approximately 3 Mb of chromosome 2 containing the *Adh–cactus* region as well as the 85 kb *82F* locus on chromosome 3.

Synthesis of transcriptional regulatory networks

Naturally, some targets of transcription factors are themselves genes that encode regulatory proteins. If the target genes of each successive transcription factor in a regulatory cascade are determined, these relationships can be linked to form a circuit whose topology describes their combined activity. Recently, graph theoretical methods have been applied to associate transcription factors with their target genes in complex regulatory networks. In this model a directed graph is produced having a scale-free topology, where transcription factors tend to localize in hubs of regulatory control (Shaw 2003). Some transcriptional regulators have been shown to modulate the expression of a disproportionately large number of genes, following power-law behaviour with respect to the number of outgoing connections originating from a given factor (Babu *et al.* 2004). Conversely, the number of genes regulated by multiple factors has been shown to decrease exponentially relative to the number of transcriptional regulators involved (Guelzim *et al.* 2002). Key transcription factors are therefore likely to be essential genes whose deletion would produce a lethal phenotype (Yu *et al.* 2004) and constituting points of vulnerability in complex regulatory systems.

Lee *et al.* (2002) explored the construction of gene regulatory networks after performing ChIP-chip analysis on 106 yeast transcription factors to determine their genome-wide binding sites using an intergenic array. By observing common patterns in the data they were able to identify several basic regulatory motifs that describe transcription factor-target relationships (Figure 5A). These include single-input, multi-input and autoregulatory motifs, feedforward loops, multicomponent loops, and regulatory chains. The binding site data produced by these experiments was later integrated with gene expression data by Bar-Joseph *et al.* (2003) to identify 106 distinct regulatory modules, based on the classification of 68 transcription factors and 655 genes. In an extension of the study by Iyer *et al.* (2002), Horak *et al.* (2003b) investigated the gene targets of nine transcription factors regulated by SBF during the G1/S cell-cycle transition using the ChIP-chip approach, using the data to build a transcription factor network (Figure 5B). Functional annotation linked to the transcription factor–target relationships revealed a complex regulatory cascade governing cell growth and differentiation.

Once derived from experimental data, transcription factor–target relationships can be combined with gene expression profiles to analyse complex functional pathways. Where transcription factor–target relationships are available, known associations can be incorporated from public databases such as TRANSFAC (Matys *et al.* 2003); others can be derived from experimental data or predicted by comparing gene expression profiles between transcription factors and putative target genes. For example, Qian *et al.* (2003) was able to use support vector machines (SVMs) to predict the regulatory targets of 36 yeast transcription factors based on gene expression data. A total of 3419 regulated genes was predicted through observation of both co-expressed and time-shifted expression profiles. Yu *et al.* (2003) integrated yeast gene expression profiles with an extensive transcriptional regulatory network constructed from ChIP-chip and other experimentally derived transcription-factor-binding data. They used the network to identify global expression patterns in the relationships between transcription factors and the genes they regulate, accounting for inverted and time-shifted behaviour. Genes belonging to the same regulatory motif were often found to be co-expressed, exhibiting higher expression levels when multiple transcription factors were involved.

Gene regulatory networks are not static entities but dynamic structures that are expected to undergo significant topological changes in response to variations in cellular physiology. Luscombe *et al.* (2004) integrated gene expression and transcription-factor-binding data from a variety of sources to construct an elaborate network comprising 7074 regulatory interactions in yeast. Examining the occurence of the motifs defined by Lee *et al.*, they further analysed expression profiles to determine which regulatory subnetworks are active under different environmental conditions such as the cell

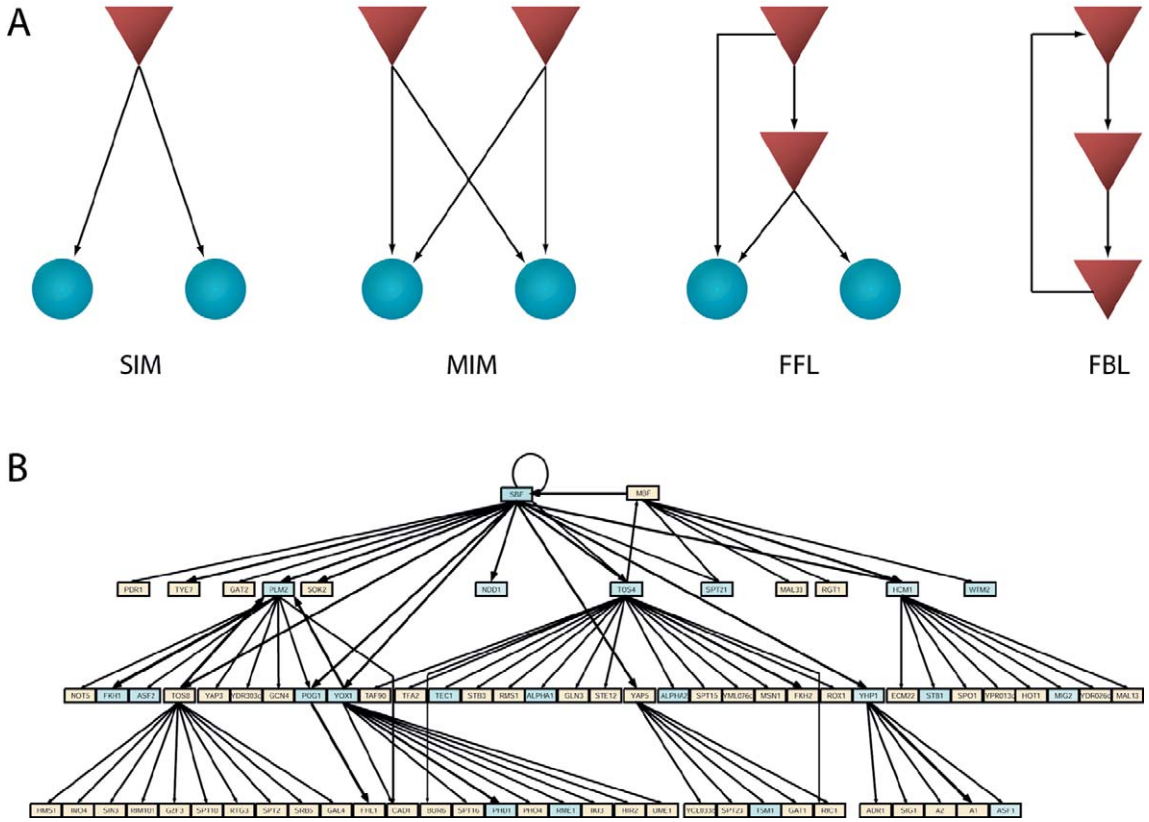

Figure 5. (**A**) Several common gene-regulatory network motifs, identified through genome-wide investigation of transcription-factor binding (after Lee *et al*. 2001). Transcriptional regulators are represented as triangles and target genes as spheres. Depicted from left to right are the single-input motif (SIM), multiple-input motif (MIM), feed-forward loop (FFL), and feedback loop (FBL). Not pictured are the autoregulatory motif and regulatory chain, which are derivative of the single-input motif. (**B**) Transcription factor network describing a cascade of regulatory control downstream of the cell-cycle regulators SBF and MBF during the G1/S transition (Horak *et al*. 2002b).

cycle, diauxic shift, sporulation, DNA damage and stress response (Figure 6A). During response to external stimuli, regulatory cascades were shown to be fairly simple and involve few feedback interactions. More complex circuitry was observed during the cell cycle and sporulation, which appear to require multiple regulatory stages involving highly interconnected transcription-factor relationships (Figure 6B). The study also characterized the influence of regulatory hubs in the system, finding that many hubs involve multifunctional transcription factors that regulate essential cellular processes. Despite their biochemical significance, the majority of regulatory hubs were observed to be transient in nature, influencing widespread transcriptional activity in some conditions but not others. Through the examination of local and global regulatory pathways at several levels of complexity, this work presents a seminal perspective of the large-scale temporal dynamics of genetic control.

Conclusion

The limited feature density of early microarray platforms led to an initial focus on gene-based sequence representation and, consequently, on comparative gene expression profiling. As array fabrication technology continues to improve, the commensurate increase in feature density has enabled the construction of microarrays able to cover large regions of eukaryotic chromosomes,

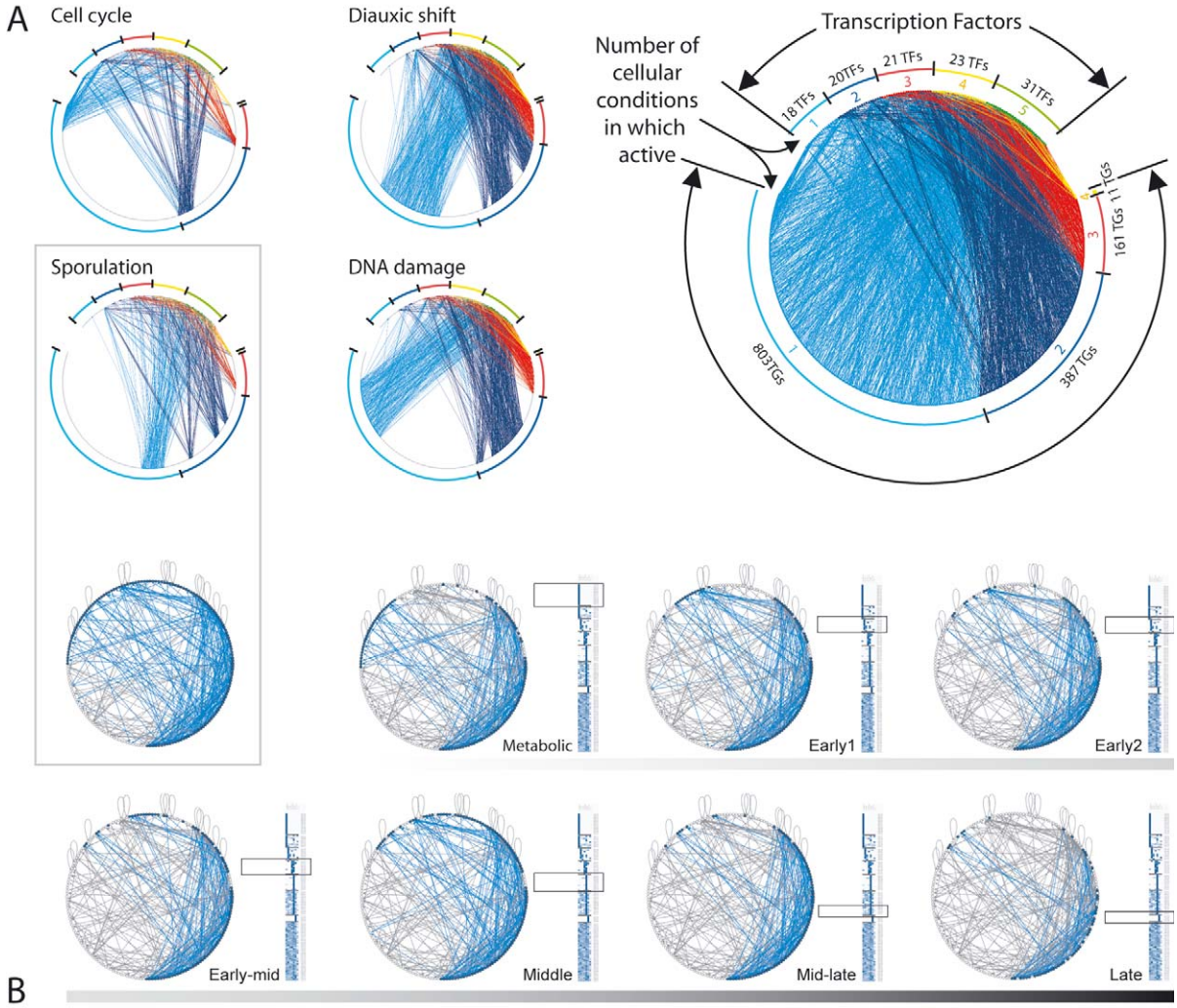

Figure 6. (**A**) Complex transcriptional regulatory networks derived from yeast ChIP-chip and gene expression data, illustrating the global static network as well as condition-specific subnetworks. Transcription factors and target genes appear as nodes on the upper and lower perimeters of each graph, respectively. Edges denote regulatory interactions and are coloured according to the number of cellular conditions in which they have been identified (adapted from Luscombe *et al.* 2004). (**B**) Detailed analysis of dynamic gene regulation during the multistage transcriptional program of sporulation (N. Luscombe, personal communication). The complete set of sporulation-associated interactions is represented in the upper leftmost graph (boxed), followed by a series of graphs highlighting the specific regulatory subnetworks activated in successive stages of the pathway. Combinatorial transcription factor usage occurs in distinct subsections of the network, as evidenced by differential gene expression patterns observed at each stage.

spanning intergenic as well as coding sequences. Novel applications of tiling arrays are constantly emerging for the large-scale characterization of chromosome dynamics. White *et al.* (2004) recently used tiling arrays to measure DNA replication timing across human chromosome 22 during S phase of the cell cycle, through differential hybridisation of early- and late-replicating chromatin from lymphoblast and fibroblast cells. The study identified 24–26 regions of early and late DNA replication, ranging in size from 100 kb to 2 Mb, and generally associated with defined cytological bands. A total of nine chromosomal regions exhibited differential replication timing between the two cell types. Additionally, a strong correlation was observed between early

replication and the expression of novel transcribed regions having low coding potential.

Unlike gene-directed approaches, tiling array experiments enable the discovery of novel genetic elements. In particular, they are becoming increasingly important for the identification of previously unannotated transcribed sequences and the large-scale analysis of gene regulation via the unbiased interrogation of the genome. Various tiling array platforms have recently been adopted as primary discovery tools by the ENCODE Project Consortium, in an effort to provide an in-depth transcriptional and regulatory characterization of 44 select regions of the human genome (Feingold *et al*. 2004).

Computational methods to relate gene expression with transcription-factor binding can produce complex networks from which higher-order regulatory mechanisms may be derived. It is expected that our ability to elucidate functional relationships from high-throughput genomic data will be enhanced through the combination of these experimental and computational techniques. This integrated approach represents a powerful analysis methodology, able to generate an unprecedented view of the transcriptional regulatory program of the cell.

Acknowledgements

This work was supported by NIH grant P50 HG02357.

References

Adams MD, Kelley JM, Gocayne JD *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651–1656.

Albert TJ, Norton J, Ott M *et al*. (2003) Light-directed 5′→3′ synthesis of complex oligonucleotide microarrays. *Nucl Acids Res* **31**: e35.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14**: 283–291.

Bar-Joseph Z, Gerber GK, Lee TI *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**: 1337–1342.

Bertone P, Stolc V, Royce TE *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242–2246.

Cawley S, Bekiranov S, Ng HH *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499–509.

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680–686.

Euskirchen G, Royce TE, Bertone P *et al*. (2004) CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* **24**: 3804–3814.

Feingold EA, Good PJ, Guyer MS *et al*. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306**: 636–640.

Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL (1993) Multiplexed biochemical assays with biological chips. *Nature* **364**: 555–556.

Gao Y, Li J, Strickland E *et al*. (2004) An *Arabidopsis* promoter microarray and its initial usage in the identification of HY5 binding targets *in vitro*. *Plant Mol Biol* **54**: 683–699.

Guelzim N, Bottani S, Bourgine P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31**: 60–63.

Harrison PM, Hegyi H, Balasubramanian S *et al*. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res* **12**: 272–280.

Hedrick SM, Cohen DI, Nielsen EA, Davis MM (1984) Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* **308**: 149–153.

Horak CE, Snyder M (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Meth Enzymol* **350**: 469–483.

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M (2002a) GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci USA* **99**: 2924–2929.

Horak CE, Luscombe NM, Qian J *et al*. (2002b) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16**: 3017–3033.

Hubank M, Schatz DG (1994) Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res* **22**: 5640–5648.

Hughes TR, Mao M, Jones AR *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**: 342–347.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.

Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**: 93–102.

Kampa D, Cheng J, Kapranov P *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**: 331–342.

Kapranov P, Cawley SE, Drenkow J *et al*. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916–919.

Kawai J, Shinagawa A, Shibata K *et al*. (2001) Functional annotation of a full-length mouse cDNAs collection. *Nature* **409**: 685–690.

Lee TI, Rinaldi NJ, Robert F *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**: 799–804.

Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci USA* **100**: 8164–8169.

Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**: 967–971.

Lieb JD (2003) Genome-wide mapping of protein–DNA interactions by chromatin immunoprecipitation and DNA microarray hybridisation. *Meth Mol Biol* **224**: 99–109.

Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**: 327–334.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nat Genet* **21**: 20–24.

Lockhart DJ, Dong H, Byrne MC *et al.* (1996) Expression monitoring by hybridisation to high-density oligonucleotide arrays. *Nat Biotechnol* **14**: 1675–1680.

Luscombe NM, Royce TE, Bertone P *et al.* (2003) Express yourself: A modular platform for processing and visualizing microarray data. *Nucl Acids Res* **31**: 3477–3482.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**: 308–312.

Mao DY, Watson JD, Yan PS *et al.* (2003) Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* **13**: 882–886.

Martone R, Euskirchen G, Bertone P *et al.* (2003) Distribution of NF-κB binding sites across human chromosome 22. *Proc Natl Acad Sci USA* **100**: 12247–12252.

Maruyama K, Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**: 171–174.

Mattick JS (2003) Challenging the dogma: the hidden layer of nonprotein-coding RNAs in complex organisms. *Bioessays* **25**: 930–939.

Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet* **5**: 316–323.

Matys V, Fricke E, Geffers R *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucl Acids Res* **31**: 374–378.

Nuwaysir EF, Huang W, Albert TJ *et al.* (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* **12**: 1749–1755.

Odom DT, Zizlsperger N, Gordon DB *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**: 1378–1381.

Ota T, Suzuki Y, Nishikawa T *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**: 40–45.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* **91**: 5022–5026.

Pruitt KD, Tatusova T, Maglott DR (2003) NCBI Reference Sequence project: update and current status. *Nucl Acids Res* **31**: 34–37.

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19**: 1917–1926.

Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32** (Suppl): 496–501.

Ren B, Robert F, Wyrick JJ *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.

Ren B, Cam H, Takahashi Y *et al.* (2002) E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. *Genes Dev* **16**: 245–256.

Rinn JL, Euskirchen G, Bertone P *et al.* (2003) The transcriptional activity of human chromosome 22. *Genes Dev* **17**: 529–540.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.

Selinger DW, Cheung KJ, Mei R *et al.* (2000) RNA expression analysis using a 30-base pair resolution *Escherichia coli* genome array. *Nat Biotechnol* **18**: 1262–1268.

Shaw S (2003) Evidence of scale-free topology and dynamics in gene regulatory networks. *Proceedings of the ISCA 12th International Conference on Intelligent and Adaptive Systems and Software Engineering*. pp. 37–40.

Shoemaker DD, Schadt EE, Armour CD, *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*. **409**: 922–927.

Solomon MJ, Varshavsky A (1985) Formaldehyde-mediated DNA–protein crosslinking: a probe for *in vivo* chromatin structures. *Proc Natl Acad Sci USA* **82**: 6470–6474.

Stolc V, Gauhar Z, Mason C, *et al.* (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*. **306**: 655–660.

Strausberg RL, Feingold EA, Klausner RD, Collins FS (1999) The mammalian gene collection. *Science* **286**: 455–457.

Sun LV, Chen L, Greil F *et al.* (2003) Protein–DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc Natl Acad Sci USA* **100**: 9428–9433.

Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200**: 149–156.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucl Acids Res* **30**: 3732–3738.

van Steensel B, Henikoff S (2000) Identification of *in-vivo* DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* **18**: 424–428.

van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet* **27**: 304–308.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* **270**: 484–487.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ (2002) Isolating human transcription factor targets by

coupling chromatin immunoprecipitation and CpG island microarry analysis. *Genes Dev* **16**: 235–244.

Wells J, Yan PS, Cechvala M, Huang T, Farnham PJ (2003) Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sties during S phase. *Oncogene* **22**: 1445–1460.

White EJ, Emanuelsson O, Scalzo D *et al*. (2004) DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc Natl Acad Sci USA* **101**: 17771–17776.

Yamada K, Lim J, Dale JM *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842–846.

Yelin R, Dahary D, Sorek R *et al*. (2003) Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21**: 379–386.

Yu H, Luscombe NM, Qian J, Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* **19**: 422–427.

Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends Genet* **20**: 227–231.

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**: 2541–2558.