4
Deriving Meaning from Genomic Information

RAYMOND J. CHO*

Departments of Genetics and Biochemistry, Stanford University School of Medicine, Stanford, CA 94035, U.S.A.

Introduction

As sequencing of the human genome draws to a close, the fruits of this vision have already achieved startling maturity. By leveraging DNA sequence information toward robust, new technological platforms, researchers are rapidly recharting the modern course of molecular genetics. Until now, the currencies of genomic experimentation have remained recognizable, if vastly increased in scope. We are still assaying the regulation of gene activity or linking phenotypes to genetic variation – only on a scale four or five orders of magnitude greater than before. Indeed, many in the scientific community first embraced genomics for its promise of a wealth of data traditionally generated through more painstaking means.

But large-scale technologies presage far deeper change in the very way we think about biological systems. The results of experimental genomics – noisy, sparse in context, and overwhelmingly vast in scope – resist the bounded conclusions drawn from conventional biological study. Rather, these data reflect the combinatorial complexity of cellular systems and challenge us to discern the patterns underlying biological design. Genomic approaches reveal not only discrete links that connect individual proteins and phenotypes, but also broad communications between parts of pathways, chromosomes, and cellular process. Ultimately, these studies may prove most valuable for providing answers to those questions we never set out to ask.

Divining these new sorts of conclusions is a task to which biologists find themselves largely unaccustomed. And so, as genomic data proliferates, accessing and drawing meaningful insights will soon pose as great a technological challenge as production of the data itself. In the past two years, more information regarding genetic diversity and mRNA expression has been released into the public domain than from the preceding ten. That this drastic acceleration can be explained primarily by large-scale DNA sequencing capability and the increasing popularity of DNA arrays

*Present address: InGenuity Systems. 2160 Gold Street, 2nd Floor, P.O. Box 2199, Alviso, CA 95002, U.S.A. (rcho@ingenuity.com)

Biotechnology and Genetic Engineering Reviews – Vol. 17, August 2000
0264–8725/00/17/91–107 $20.00 + $0.00 © Intercept Ltd. P.O. Box 716. Andover, Hampshire SP10 1YG, U.K.

barely diminishes the implications. Recent successes in genome-wide strategies for studying protein-protein interactions and gene disruptions predict that a new scale will soon be drawn for virtually every type of biological information (Winzeler *et al.*, 1999; Bartel *et al.*, 1996; Shoemaker *et al.*, 1996; Fromont-Racine *et al.*, 1997; Schimenti and Bucan, 1998; Ross-Macdonald *et al.*, 1999).

If the scope and growth of these data sets makes our ability to mine them appear vanishingly small, it is important to remember that the study of genomic information, like any scientific field, requires articulation of the problem before solutions can be generated. The past two years have observed tremendous advances in the large-scale study of genetic variance, transcription, and gene function, representing more than 80% of all experimental genomic data in the public domain. This emergent body of work richly illustrates the new classes of discovery made possible by genomic information. Consideration of these lessons provides a context for both the complexity and promise of the task at hand and reveals the first lights by which we navigate a new course in biological inquiry.

Constraints in the use of genomic information

VALIDATING QUALITY

The expansive scale of genomic information makes conventional peer review all but impossible. More traditional denominations of biological discovery – individual links between proteins, or between genes and phenotype – invite meticulous examination and duplicate experimentation. Genome era technologies, however, enable an individual researcher to produce thousands of data points in a single afternoon. Biological study on a genome scale, therefore, raises the odds that some component of these results may be in error, and makes it far less likely that such errors can be detected prior to public release.

Verification of these findings is further complicated by the considerable pressure on genomic researchers to make results immediately accessible to the community (Bentley, 1996). Venter and colleagues have argued that the release of non-peer reviewed data from human genome centres may seriously compromise standards of quality and completeness (Adams and Venter, 1996). Recent studies report that as much as 2% of deposited sequence data may contain some form of error, ranging from omission to incorrect assignment of sequence identity. Several reviews have documented the ease with which these mistakes can subvert both experimental decisions and the interpretation of results, supporting concerns that the public release of erroneous genomic data will incur especially steep costs in the broader biological community (Pennisi, 1999).

Sequence information, of course, benefits from direct correspondence to a discrete biological quantity. Accordingly, the quality of these data can be decisively validated through redundant sequencing. However, most genomic conclusions, like those drawn from conventional biological study, necessarily reflect subtle distinctions in experimental design, execution, and means of analysis (Lander, 1999). Nearly identical experiments at the genome level may produce conflicting conclusions about dozens, or even hundreds, of genes. This variability has led researchers to analyse far larger sample sizes for the determination of statistically significant conclusions

(Golub *et al.*, 1999; Galitski *et al.*, 1999). One solution may lie in the centralized deposition and curation of these multiple redundant data sets and the development of standardized tools and protocols for their comparison. Like inconsistent accounts of a single event related by multiple witnesses, these results, and their associated conclusions, demand judicious interpretation and consideration as a composite whole.

FINDING PATTERNS

Genomics relies heavily on the tenet that functionally related genes share quantifiable commonalities in behaviour or appearance. Even the most focused genomic experiments intimate the need for systematic methods for finding these patterns. DNA arrays may be used to find the major transcriptional targets of a disease gene, but the resulting data sets reflect the coordinate regulation of hundreds of biological pathways. Similarly, the genetic variants identified during re-sequencing of individual genes reveal the edges of vast, genome-wide patterns of allele inheritance. The identification of such patterns – whether the coordinate regulation of transcripts or the stereotyped response of gene deletion strains – now comprise a basic denomination of biological conclusion.

Genomics researchers have adopted a number of standard mathematical methods for the identification of these non-random patterns (Spellman *et al.*, 1998; Eisen *et al.*, 1998; Golub *et al.*, 1999). The full range of potential algorithms that may be applied for this purpose remains relatively untested. Certain mathematical approaches are certain to be better suited to analysis of specific genomic outputs (for example, finding patterns in transcription data as opposed to genome-wide protein-protein interaction networks), or of different stages in this analysis (for example, the discovery of patterns as opposed to redetection of these patterns in new data sets). Moreover, the application of several different algorithms may well be required to discern the full range of biologically meaningful patterns in any given data set.

Most of our current methods for discerning patterns originate from the study of large-scale sequence and expression data. The next five years promise far greater diversity in public domain genomic information. Maturing experimental platforms are enabling the dissection of biological pathways on levels including mRNA transcription, protein modification, and links to phenotype (Yates, 1998; Lander, 1996; Ross-Macdonald *et al.*, 1999). It follows that our analyses must soon be extended to data sets assembled from a multitude of experimental approaches. The complexity of such scenarios, of course, is dizzying. Researchers will soon find themselves bestowed with combinatorial matrices of interconnections between genes, proteins, and phenotypes. These relationships will themselves prove highly dependent on how raw data has been generated and standardized. Means for integrating this information, and the best computational methods for making discoveries from such a resource, remain to be determined.

DISCERNING MEANING

The most challenging aspects of understanding genomic data may well lie beyond development of the appropriate mathematical analyses. Identification of a non-random pattern says nothing of its biological relevance. In scoped experimental designs, of course, connections to function may be readily inferred. The ploidy-dependent transcriptional changes observed by Fink and colleagues, for example, agree well with observed alterations in cell size and mitosis (Galitski *et al.*, 1999). However, complex data sets with less obvious phenotypic implications place greater burdens on functional analysis; inspection of hundreds of genes that exhibit a common pattern of behaviour for the most convincing relationships is proving impractically tedious and arbitrary. These concerns emphasize that understanding genomic information is based largely on what we already know. That prior base of knowledge currently lies embedded in hundreds of thousands of public domain research articles that speak to genetic and biochemical function. Annotation databases help centralize this information, but only scratch the surface of our needs for more enlightened large-scale analysis. The use of legacy knowledge to interpret the functional significance of genomic results, like the identification of data patterns, may soon necessitate computational approaches. These approaches will likely require novel data architectures capable of coherently structuring diverse information about gene function.

But deriving meaning from genomic information implies more than correlation of function with pattern. Each species of these data brings to biological science a unique perspective, and logic, imperceptible from the proximity of traditional experimental thinking. Genomics not only elucidates biological processes, it also vets the mosaic of modern discovery. Recent work in genomics has challenged our beliefs about global transcriptional regulation, the multifunctionality of individual genes, and the course of human evolution. Identifying such changes, the first step in extending them, requires that we examine how this information is persistently reshaping our understanding of living systems.

Mathematical approaches to the analysis of genomic data

SEEKING SIMILARITY

Application of classical mathematical methods for pattern finding is not new to biology – DNA sequence comparisons were constructed on these very foundations. However, the sudden explosion in large-scale expression data has brought about a change to the way scientists regard such analysis. Whereas DNA sequence comparisons provide a hypothesis for detailed, mechanistic study, analyses of expression data often represent direct conclusions about a cellular process.

The direct output from DNA arrays, now the most common platform for large-scale expression study, consists of an averaged fluorescence intensity for each surveyed gene. Where multiple timepoints are profiled, these data are normalized between each sample and subjected to one of a number of mathematical algorithms that classify genes with similar expression profiles into discrete groups based on: (i) a metric of similarity; and (ii) an implementation of classification (Everitt, 1993). For instance, both the standard correlation coefficient and the Euclidean distance metric have been tested extensively in the assessment of gene similarity from expression data. Once a metric has been selected, an implementation for grouping similar profiles must be applied. Common classes of implementation include unsupervised hierarchical clustering (Eisen *et al.*, 1998), iterative *k*-means analysis (Tavazoie *et al.*, 1999), and self-organizing maps (Tamayo *et al.*, 1999), each providing a differing extent of structure and flexibility in the consequent classification. In hierarchical clustering, pairwise comparisons for similarity are performed for every profile in a data set, generating a single, rigid dendogram. Clusters of closely related transcripts are then selected arbitrarily for more detailed biological discussion. In iterative *k*-means analysis, the number of final classifications is arbitrarily determined *a priori*, and a series of assignments and recalculations of the group centre are performed to divide the total data set into final clusters. Self-organizing maps similarly utilize iterative refinements, but allow an initial mapping of nodes in *k*-dimensional space. These various methods are likely to reach similar classifications for the most closely related transcript profiles in a data set, but diverge significantly with respect to more individualistic patterns. Side-by-side comparisons of these methods are now beginning to enter the literature. Self-organizing maps, for example, may provide particular value by enabling partial structure to be imposed on a data set, as opposed to the strict classifications achieved through hierarchical clustering.

Nearly all of these mathematical analyses require subjective heuristics supplied by the researcher: for example, the discarding of genes with minimal fluctuation during the time course and more detailed study of clusters with average patterns most closely fitting canonical regulatory behaviour. As a result, the reporting of expression clusters is influenced considerably by the predispositions of the authors. Application of a standard set of analyses to all expression data sets might provide considerable value in the comparison of independent studies. It has also been noted that the output of these algorithms is dependent on the nature of the starting data – for example, numbers of duplicate samples and the reproducibility inherent to the technology used to generate the findings. The biological conclusions we draw may therefore reveal as much about the underlying information as the processes we seek to elucidate.

MODELLING CIRCUITRY

Regulatory communications within a cell have been interpreted as signalling networks (McAdams and Shapiro, 1995; Yuh *et al.*, 1998). Computational modelling and simulation of these networks offer a promising means for studying systems too complex for cognitive analysis. Iyengar and colleagues have constructed and extensively tested such kinetic models based on protein-signalling data in the public domain (Weng *et al.*, 1999; Bhalla and Iyengar, 1999). In this approach, simple quantitative models are developed for individual biological pathways, followed by iterative refinement of the model and its kinetic parameters to reach agreement with empirical data. This model is then extended sequentially to neighbouring pathways, achieving computational representations of complex regulatory networks. In addition to accurately fitting experimentally determined findings, these simulations are required to obey basic principles of mass conservation and microscopic reversibility.

The surprising success of such approaches in approximating the kinetic behaviour of multipathway networks suggests their application to quantitative genomic information. Although little modelling has yet been performed on such information, it has been noted that large-scale expression data sets represent propitious territory for extension of these computational paradigms (Huang, 1999). Genomic data promise more consistent inputs than findings generated by independent laboratories under varying experimental conditions. Furthermore, transcriptional circuitry may be modelled without knowledge of function for any given gene. Computational simulations may help solve one of the central conundrums of genomics – the observation of large numbers of small to moderate changes in cellular activity. These alterations may collectively influence phenotype, but like multigenic contributions to phenotype, their effects are often difficult to measure individually. Theoretically, modelling these inputs in a quantitative manner may allow the detection of complex, aggregate effects – transcriptional and otherwise – on molecular and clinical phenotypes.

Major classes of genomic data and their translation to biological meaning

MINING GENETIC VARIANCE

As illustrated by the recent proliferation of single nucleotide polymorphisms (SNPs) in public databases, genome sequence is abundantly plural. Application of DNA array and high-throughput sequencing technologies have radically accelerated the characterization of nucleotide diversity in a range of organisms (Wang *et al.*, 1998; Hacia *et al.*, 1999; Kwok *et al.*, 1996; Buetow *et al.*, 1999). Several groups have completed the contiguous, *de novo* sequencing of human genes, including the angiotensin converting enzyme *ACE* and the lipoprotein lipase *LPL* (between 10–30 kB of a single gene from 10–70 individuals) (Nickerson *et al.*, 1998; Rieder *et al.*, 1999). In a complementary approach, groups led by Lander and Chakravarti have used oligonucleotide arrays to scan shorter segments of genes related to human disease (~ 190 kB distributed over 75–100 genes, in an average of 75 individuals) (Cargill *et al.*, 1999; Halushka *et al.*, 1999).

The excitement surrounding these advances arises from two principal applications of genome-scale SNP data. First, such a catalogue represents a revolutionary tool for finding genes responsible for heritable traits. SNPs comprise the densest set of genetic markers in eukaryotic genomes and may underlie a significant proportion of phenotypic variation. Information regarding the frequency of these variants in targeted populations may therefore associate a gene and phenotype, either directly or indirectly, with far greater statistical power than possible with conventional linkage mapping (Risch and Merikangas, 1996). Second, global surveys of genetic variance should reveal global patterns of mutation across the genome, enabling a new perspective on molecular and organismal evolution (Chakravarti, 1999; Nickerson *et al.*, 1998).

The requirements for both direct and indirect association studies are considerable at the genome level. SNPs display significantly lower heterozygosity than microsatellite polymorphisms, reducing the relative informativeness of any given marker. Moreover, the lower heterozygosity of SNPs predisposes to ambiguous haplotyping, incurring further loss of discrimination (Hodge *et al.*, 1999). The power of these studies depends not only on the identification of larger numbers of SNPs, but also the rarest ones. Recent disease mutations may segregate with newer (and therefore less frequent) alleles. With regard to direct association studies, the Lander and Chakravarti studies confirm that non-synonymous coding polymorphisms, which can give rise to phenotypic variation by altering protein structure, display significantly lower minor allele frequencies in the human population (Cargill *et al.*, 1999). Consequently, a vast catalogue of SNPs – perhaps well over a half million – might be required to test association for every gene in a genome. In the interim, as denser biallelic maps are assembled in natural populations, genome-scale SNP data have been leveraged toward the acceleration of traditional linkage mapping in model organisms, where most variants are maximally informative. Recent studies in yeast and *A. thaliana* confirm the feasibility of this approach and indicate that more than half of all SNPs may eventually prove amenable to high-throughput genotyping with oligonucleotide arrays (Cho *et al.*, 1999; Winzeler *et al.*, 1998).

Nucleotide diversity also bears a profound and informative imprint of evolutionary selection. For instance, polymorphism rates vary widely from gene to gene (Cargill *et al.*, 1999; Halushka *et al.*, 1999). Some of this variability may be attributed to physical proximity to: (i) repeat sequences and pseudogenes, which may significantly increase nearby mutation rates; and (ii) loci that experience strong functional selection for or against diversity (Nickerson *et al.*, 1998). However, the relative variance in a gene is also certain to reflect selective pressures related to its own biological function. Methods to differentiate these effects have been described, and will become increasingly relevant as greater nucleotide diversity is characterized for individual genes. Significant differences in nucleotide diversity have also been detected between non-coding, degenerate, and non-degenerate positions within genes. These findings are ostensibly explained by strong selection against mutations that affect protein structure. Unexpectedly, perigenic non-coding regions apparently contain lower average diversity than coding sequence, consistent with the hypothesis that changes in regulatory and splicing regions are more likely to affect the phenotype than are mutations to affect protein structure (Cargill *et al.*, 1999).

Genomic approaches also enable the characterization of genetic diversity on a gross scale. Arrays have been utilized, for example, to detect differences in gene content between strains of *S. cerevisiae*, between *Mycobacterium* species, and loss of heterozygosity in tumour samples (Lashkari *et al.*, 1997; Pollack *et al.*, 1999; Behr *et al.*, 1999). These differences have been correlated with attendant phenotypic differences. The macroscopic nature and lower frequency of such variants facilitates their evaluation for functional significance, especially in organisms amenable to genetic complementation. As these data sets accumulate, centralized databases for their curation will prove critical for the mapping of loci to phenotype, particularly in the case of large clinical studies of loss of heterozygosity.

SURVEYING CELLULAR ACTIVITY

The simplest class of experimental genomic information describes a single cellular activity; for example, protein abundance or localization in the absence of overt external stimuli (Burns *et al.*, 1994; Yates, 1998; Blackstock and Weir, 1999). In these broad surveys of the genome, internal comparison of a single data type across thousands of genes provides the opportunity to draw global conclusions regarding biological systems. Because these approaches do not attempt to profile dynamic process, initial interpretations can be reached at relatively low computational cost. The most comprehensive of these surveys has been performed with respect to the regulation of mRNA expression (*Figure 4.1*). Several groups have taken independent

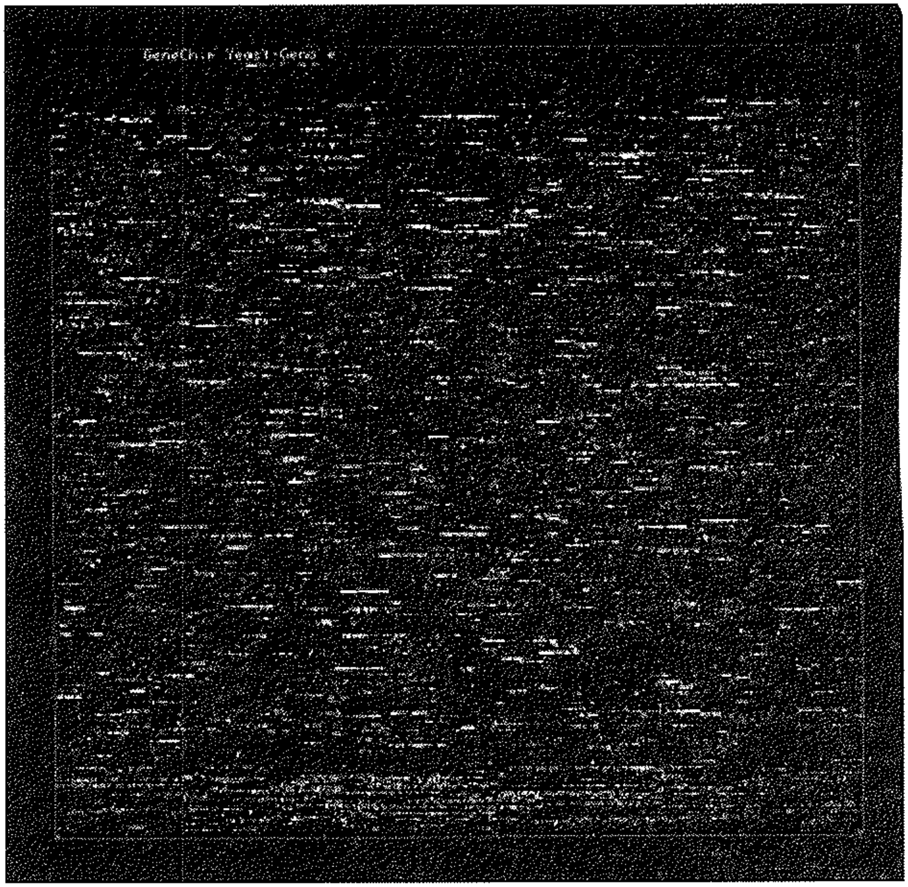

Figure 4.1. Fluorescence image of an Affymetrix transcript array containing oligonucleotide probe sets for all ORFs in the yeast genome, following hybridization of 10 ug of biotin-labelled, fragmented yeast genomic DNA. Courtesy of David J. Lockhart, Elizabeth A. Winzeler and Dan Giang, Novartis Institute for Functional Genomics, San Diego, CA, U.S.A.

approaches to quantitate thousands of transcript levels in log phase *S. cerevisiae* cells, characterizing the so-called yeast transcriptome (Wodicka *et al.*, 1997; Velculescu *et al.*, 1997). Collectively, these results establish a standardized, genome-wide histogram of absolute transcript abundance, reveal gross chromosomal effects on transcription (for example, telomeric silencing) and identify open-reading frames (ORFs) not previously annotated in public databases.

Genome-scale surveys provide a unique opportunity for observing interrelationships between cellular activities that are usually examined separately. For example, low concordance has been established between genome-wide transcript and protein abundance in *S. cerevisiae*, with differences in ratio as high as 30-fold (Burns *et al.*, 1994). These comparisons represent the first steps toward dissecting the determinants of protein level for every gene in a genome. Global assessments may also be made at the functional level. Genes displaying the lowest levels of transcription in these studies are significantly enriched for both open-reading frames with no known biological role and non-essential genes (Winzeler *et al.*, 1999). One explanation may be that functionally redundant genes, as a class, tolerate lower levels of expression. Alternatively, expression level may correlate positively with the likelihood that a phenotype may be observed for a gene. Identification of conditions that elevate expression levels of some uncharacterized genes may therefore facilitate elucidation of their function.

If patterns of cellular activity truly underlie phenotype, it follows that different cell types may be distinguished on the basis of these data. For instance, certain transcripts are likely to display consistent differences in abundance between, for example, distinct tissues and genetic backgrounds. Lander and colleagues have identified a set of such transcripts through large-scale transcriptional surveys of numerous acute lymphoblastic leukaemia and acute myelogenic leukaemia samples, and successfully used these differences as a basis for the classification of new samples (Golub *et al.*, 1999). Other groups have begun focusing on transcriptional differences in tumour samples that may be predictive of phenotype, with regard to both clinical course and therapeutic responsiveness (Perou *et al.*, 1999). These comparative approaches promise powerful basic research and biomedical applications without the requirement for characterization of gene function.

Surveys of cellular activity also reveal a general need for quantitative assessments to reduce the noise inherent to genomic data. For instance, a considerable proportion of transcripts in large-scale expression data display insufficient abundance for meaningful fold-change comparisons, and should be excluded from further analysis (Lockhart *et al.*, 1996). In addition, the number of genes assayed in these studies necessitates statistical tests to differentiate significant changes from random fluctuation (Galitski *et al.*, 1999). Quantitative metrics can also directly predicate biological relevance: Legrain and colleagues have demonstrated in large-scale two-hybrid screens that the repeated identification of single clones, and identification of separate inserts for the same gene, are predictive of functional interaction (Fromont-Racine *et al.*, 1997). In the sizeable data generated by these approaches, redundance may represent a key signature of signal over noise.

PROFILING DYNAMIC PROCESS

DNA array technologies are distinguished from current genome-wide approaches to protein interaction and localization in their ability to easily characterize gene activity (transcript abundance) as a function of cellular state (Lockhart *et al.*, 1996; Schena *et al.*, 1995). Expression studies therefore represent the first genomic method useful for the study of dynamic biological processes. Researchers have already profiled mRNA expression levels during classic processes such as diauxic shift, mitosis, meiosis, and the serum response (DeRisi *et al.*, 1997; Cho *et al.*, 1998; Chu *et al.*, 1998; Spellman *et al.*, 1998; Iyer *et al.*, 1999). The novel complexity of the resulting data sets has proved somewhat of a *tabula rasa* from which researchers have drawn four major classes of biological conclusion.

First, the occurrence of a gene in an expression cluster with classical kinetics of induction has been interpreted as indication of function in the profiled process. For example, Chu and colleagues have observed the up-regulation of numerous genes related to vesicle fusion and membrane formation during the profiling of gametogenesis in *S. cerevisiae*, leading to the conclusion that these genes may facilitate spore formation. Similarly, Iyer and colleagues have reported induction of genes implicated in clot formation and remodelling during the serum response, concluding that increased activity of these transcripts may represent an important functional consequence of serum exposure. These inferences are not based purely on kinetics of induction, but also on the observation that these kinetics are shared by other genes previously established to play a central role in these processes; what Chu and colleagues have termed 'guilt by association'. Second, an apparent critical mass of genes of common function in an expression cluster has been used to link that function and the profiled biological process. In the study by Iyer *et al.*, clot remodelling itself was proposed to play an important role in the response to serum. More recently, the differential expression of metabolism-related genes in ageing mice has been read as an indication that these pathways may be causal for some of the manifestations of senescence (Lee *et al.*, 1999a).

Third, expression data have been used to dissect regulatory networks. In these studies, global transcriptional levels are quantitated before and after chromosomal deletion or controlled repression of a putative regulatory gene (Holstege *et al.*, 1998; Lee *et al.*, 1999b; Harkin *et al.*, 1999; Wyrick *et al.*, 1999). The consequent differences provide long-awaited verification of the proposed physiological relevance of specific transcription factors and identify major targets of genes whose function is still under investigation. In some cases, these data demand reconsideration of accepted hypotheses regarding the effects of specific regulatory molecules (Holstege *et al.*, 1998). Finally, profiling studies have been used to explore the regulatory architecture of the genome: re-evaluation of known transcriptional regulatory sequences, identification of new elements, and examination of the chromosomal organization of coordinately regulated genes (Cho *et al.*, 1998; Chu *et al.*, 1998; Spellman *et al.*, 1998; Tavazoie *et al.*, 1999; Zhang, 1999).

The caveats to the first two interpretations of profiling data, naturally, do not lie in the findings themselves, which reflect the observational nature of genomics, but in: (i) the extremely subjective, and therefore variable, context in which these conclusions are drawn; and (ii) the general lack of mechanistic validation. With regard to standardization of functional interpretation, recent advances in systematic analysis indicate potential solutions. Church and colleagues have interpreted the statistical overrepresentation of genes of common function within a set of coordinately regulated transcripts as activation of a biological pathway (Tavazoie *et al.*, 1999). In this approach, each gene was assigned a non-unique function based on the publicly accessible MIPS database. The application of this analysis to genome-wide transcription data in *S. cerevisiae* has enabled the detection of up-regulation of, for example, chromosome segregation and DNA replication pathways during cell division, with no *a priori* assumptions about mitotic functions. Such methods advantageously define a standard interpretation of function, albeit a simple one, which may be rigorously applied by multiple laboratories to independent data sets.

With regard to questions of mechanism, increasing numbers of published studies have reported validation of the phenotypic significance of expression findings through traditional genetic or biochemical means (Lee *et al.*, 1999b). These detailed studies are of great interest to the larger biological community, especially in light of recent studies that call into question the thesis that transcripts are generally up-regulated during conditions that necessitate their function (Winzeler *et al.*, 1999). In the long run, however, only a small proportion of inferences made on the genome level can be evaluated mechanistically before publication. The data sets are simply too large and the demand for them too great. Development of complementary genomic technologies should enable the generation of multiple lines of evidence for the function of a gene, providing leads of higher quality for the traditional biologist.

FINDING FUNCTION

If genomic information has been perceived solely as numeric networks of data devoid of biological significance, the emergence of large-scale efforts for the functional characterization of genes has emphatically proved otherwise. Using variations on the theme of genetic disruption, researchers are systematically evaluating the functions of sequenced genes in organisms ranging from yeast to mouse (Shoemaker *et al.*, 1996; Winzeler *et al.*, 1999; Davis and Justice, 1998; Schimenti and Bucan, 1998; Giaever *et al.*, 1999; Smith *et al.*, 1996). Information from these projects represents the critical link to phenotype from virtually every other type of genomic information.

In a sense, these efforts present a mirror image of projects that characterize genome-wide sequence diversity. Rather than elucidating the full spectrum of genes that may contribute to an individual phenotype, functional genomics first seeks the complement of biological roles fulfilled by any one gene. Information from large-scale functional analysis projects currently ranges from one-to-many links between individual gene disruptions and gross phenotype (in the case of systematic deletions in vertebrate organisms), to more quantitative evaluations of the behaviour of yeast strains deficient for the activity of specific genes. However, these projects also promise differentiation of the phenotypes resulting from heterozygous, homozygous, and partial disruptions, as well as descriptions of animals containing disruptions of multiple genes (Fields *et al.*, 1999; Spradling *et al.*, 1999; Giaever *et al.*, 1999).

Scaling of functional studies to the genome provides researchers with the ability to interpret the effects of gene disruptions in a standardized context. Knockouts have long been known to exhibit phenotypes in a background-dependent fashion, complicating assessment of the full genetic requirements for expression of a trait. Coordinated projects, such as efforts in mouse at the Jackson Laboratory and in *S. cerevisiae* through the International Deletion Consortium, will construct a complete catalogue of gene disruptions in a defined set of strains (*Figure 4.2*). Furthermore, these projects seek to assay the phenotypes of these disruptions under standardized conditions. Therefore, the information generated from these projects should display relatively low variation from controllable sources. This degree of standardization also heralds the introduction of true quantitative measures to the study of gene disruption. Phenotypes, traditionally compared on a gross level, should now be distinguishable on criteria such as the rate of drop-out of a deletion strain from a deletion pool, or changes in blood gas levels in a knockout mouse.

Some of the most illuminating results regarding gene disruption have come from the set of tagged transpositions and gene disruptions in *S. cerevisiae* generated by Snyder and colleagues (Ross-Macdonald *et al.*, 1999). The potential for multiple transposition events in a single gene enables a higher degree of resolution in functional characterization. For example, a series of insertions in the transcription

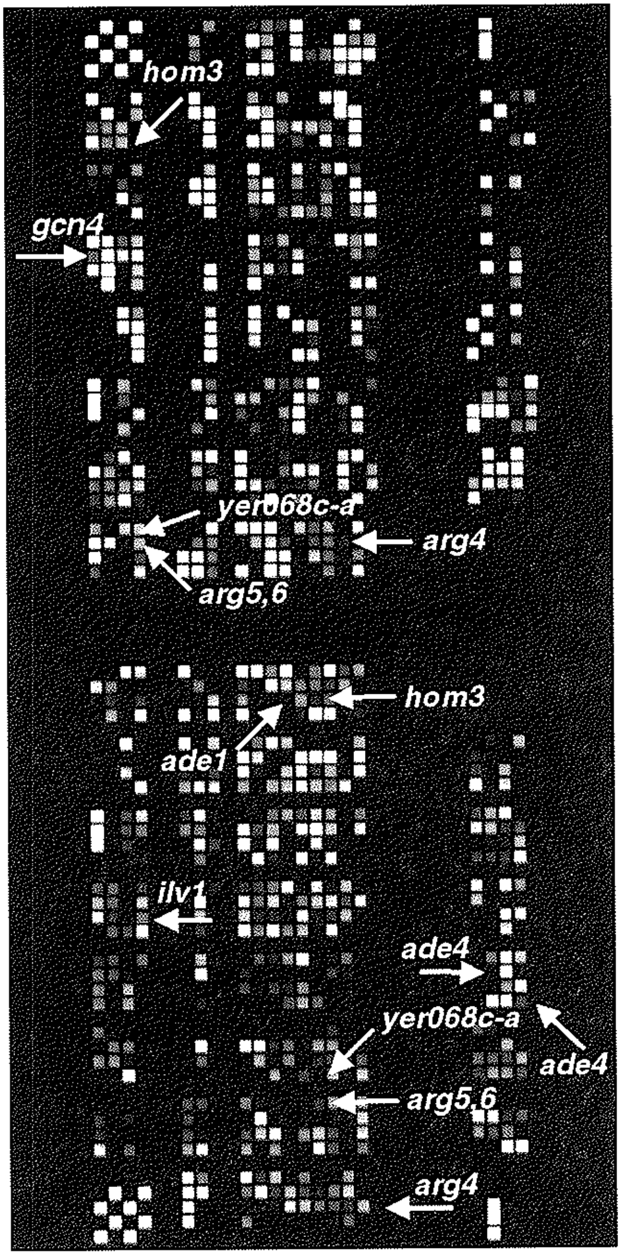

Figure 4.2. Scanned images of high-density oligonucleotide arrays following hybridization with fluorescently labelled DNA 'barcodes' amplified from 558 diploid yeast strains grown for 0 and 6 hours in minimal medium. Each strain carries a homozygous knockout of a single gene replaced with a cognate 20-mer oligonucleotide barcode corresponding to a feature on the array. Samples from the two timepoints were labelled with distinct fluorophores, allowing detection of differential fluorescence intensities at a given feature, and therefore, strains that exhibit a growth defect in minimal medium. Features matching a strain that shows a growth defect in minimal medium are labelled and marked with an arrow. Only a portion of the total array is shown. Courtesy of Elizabeth A. Winzeler and David J. Lockhart, Novartis Institute for Functional Genomics, San Diego, CA, U.S.A.

factor *IMP2* distinguishes its requirements in cell wall biogenesis and sugar utilization. These initial results suggest a significant but uncharacterized degree of pleiotropy in eukaryotic genes. Furthermore, Snyder and colleagues have demonstrated the first applications of pattern finding to large-scale studies of biological phenotype. This approach may be utilized to group genes responsible for a similar phenotype or, alternatively, to segregate growth conditions that induce phenotypes from a common set of disruptions. Expansion of this model to a genome scale should facilitate not only understanding of the biological role of individual genes, but provide a framework for perceiving function-related patterns in transcription and proteomic data.

Of course, these mass characterizations of gene disruption comprise surface clues – initial, standardized observations that serve as a prelude to the detailed analyses necessary to confirm specific biological hypotheses. Impaired survival of a deletion stain in minimal media may implicate the disrupted gene in a metabolic pathway. However, this observation may also indicate a more complex phenotype enhanced by reduced ATP levels. Assaying each disruption under a large number of differentiating conditions will help distinguish these possibilities, but will require continued development of information systems sufficiently expressive to represent a broad range of phenotypic findings and experimental conditions.

Moving forward

Any discussion of the current state of genomic information risks almost instant obsolescence. The imminent potential of these data are so high, and our ability to represent and analyse complex systems so rudimentary, that this field stands ripe for revolutionary change. These advances must necessarily emerge from different directions. Researchers will require new paradigms, and technologies, that allow translation of disparate genomic data types and traditional biological information into the same language. These more coherent data architectures must themselves be examined for intricate patterns revealing global biological design. To this end, the computational methods currently used for the analysis and comparison of genetic sequence and transcriptional information must rapidly beget algorithms capable of mining more diverse data types. Moreover, the very quantitative parameters of these systems appear amenable to computational simulation, allowing the evolution of genomics into a true information science. That so little has been determined about the use of data in the post-genome era intimates that the landscape, and vision of molecular life science, is up for grabs. If only one thing is clear, it is that this race is on.

References

ADAMS, M.D. AND VENTER, J.C. (1996). Should non-peer-reviewed raw DNA sequence data release be forced on the scientific community? *Science* **274**, 534–536.

BARTEL, P.L., ROECKLEIN, J.A., SENGUPTA, D. AND FIELDS, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. *Nature Genetics* **12**, 72–77.

BEHR, M.A., WILSON, M.A., GILL, W.P., SALAMON, H., SCHOOLNIK, G.K., RANE, S. AND SMALL, P.M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523.

BENTLEY, D.R. (1996). Genomic sequence information should be released immediately and freely in the public domain. *Science* **274**, 533–534.

BHALLA, U.S. AND IYENGAR, R. (1999). Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387.

BLACKSTOCK, W.P. AND WEIR, M.P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology* **17**, 121–127.

BUETOW, K.H., EDMONSON, M.N. AND CASSIDY, A.B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genetics* **21**, 323–325.

BURNS, N., GRIMWADE, B., ROSS-MACDONALD, P.B., CHOI, E.Y., FINBERG, K., ROEDER, G.S. AND SNYDER, M. (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in Saccharomyces cerevisiae. *Genes & Development* **8**, 1087–1105.

CARGILL, M., ALTSHULER, D., IRELAND, J., SKLAR, P., ARDLIE, K., PATIL, N., LANE, C.R., LIM, E.P., KALAYANARAMAN, N., NEMESH, J., ZIAUGRA, L., FRIEDLAND, L., ROLFE, A., WARRINGTON, J., LIPSHUTZ, R., DALEY, G.Q. AND LANDER, E.S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**, 231–238.

CHAKRAVARTI, A. (1999). Population genetics – making sense out of sequence. *Nature Genetics* **21**, 56–60.

CHO, R.J., CAMPBELL, M.J., WINZELER, E.A., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T.G., GABRIELIAN, A.E., LANDSMAN, D., LOCKHART, D.J. AND DAVIS, R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

CHO, R.J., MINDRINOS, M., RICHARDS, D.R., SAPOLSKY, R.J., ANDERSON, M., DRENKARD, E., DEWDNEY, J., REUBER, T.L., STAMMERS, M., FEDERSPIEL, N., THEOLOGIS, A., YANG, W.H., HUBBELL, E., AU, M., CHUNG, E.Y., LASHKARI, D., LEMIEUX, B., DEAN, C., LIPSHUTZ, R.J., AUSUBEL, F.M., DAVIS, R.W. AND OEFNER, P.J. (1999). Genome-wide mapping with biallelic markers in Arabidopsis thaliana. *Nature Genetics* **23**, 203–207.

CHU, S., DERISI, J., EISEN, M., MULHOLLAND, J., BOTSTEIN, D., BROWN, P.O. AND HERSKOWITZ, I. (1998). The transcriptional program of sporulation in budding yeast [published erratum appears in *Science* 1998 Nov 20; **282** (5393), 1421]. *Science* **282**, 699–705.

DAVIS, A.P. AND JUSTICE, M.J. (1998). Mouse alleles: if you've seen one, you haven't seen them all. *Trends in Genetics* **14**, 438–441.

DERISI, J.L., IYER, V.R. AND BROWN, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O. AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

EVERITT, B.S. (1993). *Cluster Analysis.* Oxford: Oxford University Press.

FIELDS, S., KOHARA, Y. AND LOCKHART, D.J. (1999). Functional genomics. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 8825–8826.

FROMONT-RACINE, M., RAIN, J.C. AND LEGRAIN, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two- hybrid screens. *Nature Genetics* **16**, 277–282.

GALITSKI, T., SALDANHA, A.J., STYLES, C.A., LANDER, E.S. AND FINK, G.R. (1999). Ploidy regulation of gene expression. *Science* **285**, 251–254.

GIAEVER, G., SHOEMAKER, D.D., JONES, T.W., LIANG, H., WINZELER, E.A., ASTROMOFF, A. AND DAVIS, R.W. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics* **21**, 278–283.

GOLUB, T.R., SLONIM, D.K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J.P., COLLER, H., LOH, M.L., DOWNING, J.R., CALIGIURI, M.A., BLOOMFIELD, C.D. AND LANDER, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

HACIA, J.G., FAN, J.B., RYDER, O., JIN, L., EDGEMON, K., GHANDOUR, G., MAYER, R.A., SUN, B., HSIE, L., ROBBINS, C.M., BRODY, L.C., WANG, D., LANDER, E.S., LIPSHUTZ, R., FODOR, S.P. AND COLLINS, F.S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics* **22**, 164–167.

HALUSHKA, M.K., FAN, J.B., BENTLEY, K., HSIE, L., SHEN, N., WEDER, A., COOPER, R., LIPSHUTZ, R. AND CHAKRAVARTI, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* **22**, 239–247.

HARKIN, D.P., BEAN, J.M., MIKLOS, D., SONG, Y.H., TRUONG, V.B., ENGLERT, C., CHRISTIANS, F.C., ELLISEN, L.W., MAHESWARAN, S., OLINER, J.D. AND HABER, D.A. (1999). Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* **97**, 575–586.

HODGE, S.E., BOEHNKE, M. AND SPENCE, M.A. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nature Genetics* **21**, 360–361.

HOLSTEGE, F.C., JENNINGS, E.G., WYRICK, J.J., LEE, T.I., HENGARTNER, C.J., GREEN, M.R., GOLUB, T.R., LANDER, E.S. AND YOUNG, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728.

HUANG, S. (1999). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine* **77**, 469–480.

IYER, V.R., EISEN, M.B., ROSS, D.T., SCHULER, G., MOORE, T., LEE, J.C.F., TRENT, J.M., STAUDT, L.M., HUDSON, J., JR., BOGUSKI, M.S., LASHKARI, D., SHALON, D., BOTSTEIN, D. AND BROWN, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87.

KWOK, P.Y., DENG, Q., ZAKERI, H., TAYLOR, S.L. AND NICKERSON, D.A. (1996). Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. *Genomics* **31**, 123–126.

LANDER, E.S. (1996). The new genomics: global views of biology. *Science* **274**, 536–539.

LANDER, E.S. (1999). Array of hope. *Nature Genetics* **21**, 3–4.

LASHKARI, D.A., DERISI, J.L., MCCUSKER, J.H., NAMATH, A.F., GENTILE, C., HWANG, S.Y., BROWN, P.O. AND DAVIS, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 13057–13062.

LEE, C.K., KLOPP, R.G., WEINDRUCH, R. AND PROLLA, T.A. (1999a). Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390–1393.

LEE, S.B., HUANG, K., PALMER, R., TRUONG, V.B., HERZLINGER, D., KOLQUIST, K.A., WONG, J., PAULDING, C., YOON, S.K., GERALD, W., OLINER, J.D. AND HABER, D.A. (1999b). The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. *Cell* **98**, 663–673.

LOCKHART, D.J., DONG, H., BYRNE, M.C., FOLLETTIE, M.T., GALLO, M.V., CHEE, M.S., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H. AND BROWN, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680.

MCADAMS, H.H. AND SHAPIRO, L. (1995). Circuit simulation of genetic networks. *Science* **269**, 650–656.

NICKERSON, D.A., TAYLOR, S.L., WEISS, K.M., CLARK, A.G., HUTCHINSON, R.G., STENGARD, J., SALOMAA, V., VARTIAINEN, E., BOERWINKLE, E. AND SING, C.F. (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics* **19**, 233–240.

PENNISI, E. (1999). Keeping genome databases clean and up to date. *Science* **286b**, 447-450.

PEROU, C.M., JEFFREY, S.S., VAN DE RIJN, M., REES, C.A., EISEN, M.B., ROSS, D.T., PERGAMENSCHIKOV, A., WILLIAMS, C.F., ZHU, S.X., LEE, J.C., LASHKARI, D., SHALON, D., BROWN, P.O. AND BOTSTEIN, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9212–9217.

POLLACK, J.R., PEROU, C.M., ALIZADEH, A.A., EISEN, M.B., PERGAMENSCHIKOV, A., WILLIAMS, C.F., JEFFREY, S.S., BOTSTEIN, D. AND BROWN, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41-46.

RIEDER, M.J., TAYLOR, S.L., CLARK, A.G. AND NICKERSON, D.A. (1999). Sequence variation in the human angiotensin converting enzyme. *Nature Genetics* **22**, 59–62.

RISCH, N. AND MERIKANGAS, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

ROSS-MACDONALD, P., COELHO, P.S.R., ROEMER, R., AGARWAL, S., KUMAR, A., JANSEN, R., CHEUNG, K., SHEEHAN, A., SYMONIATIS, D., UMANSKY, L., HEIDTMAN, M., NELSON, F.K., IWASAKI, H., HAGER, K., GERSTEIN, M., MILLER, P., ROEDER, G.S. AND SNYDER, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418.

SCHENA, M., SHALON, D., DAVIS, R.W. AND BROWN, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

SCHIMENTI, J. AND BUCAN, M. (1998). Functional genomics in the mouse: phenotype-based mutagenesis screens. *Genome Research* **8**, 698–710.

SHOEMAKER, D.D., LASHKARI, D.A., MORRIS, D., MITTMANN, M. AND DAVIS, R.W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genetics* **14**, 450–456.

SMITH, V., CHOU, K.N., LASHKARI, D., BOTSTEIN, D. AND BROWN, P.O. (1996). Functional ana-lysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.

SPELLMAN, P.T., SHERLOCK, G., ZHANG, M.Q., IYER, V.R., ANDERS, K., EISEN, M.B., BROWN, P.O., BOTSTEIN, D. AND FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

SPRADLING, A.C., STERN, D., BEATON, A., RHEM, E.J., LAVERTY, T., MOZDEN, N., MISRA, S. AND RUBIN, G.M. (1999). The berkeley drosophila genome project gene disruption project. Single P-element insertions mutating 25% of vital drosophila genes. *Genetics* **153**, 135–177.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E.S. AND GOLUB, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

TAVAZOIE, S., HUGHES, J.D., CAMPBELL, M.J., CHO, R.J. AND CHURCH, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

VELCULESCU, V.E., ZHANG, L., ZHOU, W., VOGELSTEIN, J., BASRAI, M.A., BASSETT, D.E., JR., HIETER, P., VOGELSTEIN, B. AND KINZLER, K.W. (1997). Characterization of the yeast transcriptome. *Cell* **88**, 243–251.

WANG, D.G., FAN, J.B., SIAO, C.J., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLYAK, L., STEIN, L., HSIE, L., TOPALOGLOU, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M.S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T.J. AND LANDER, E.S. (1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

WENG, G., BHALLA, U.S. AND IYENGAR, R. (1999). Complexity in biological signaling systems. *Science* **284**, 92–96.

WINZELER, E.A., RICHARDS, D.R., CONWAY, A.R., GOLDSTEIN, A.L., KALMAN, S., MCCULLOUGH, M.J., MCCUSKER, J.H., STEVENS, D.A., WODICKA, L., LOCKHART, D.J. AND DAVIS, R.W. (1998). Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197.

WINZELER, E.A., SHOEMAKER, D.D., ASTROMOFF, A., LIANG, H., ANDERSON, K., ANDRE, B., BANGHAM, R., BENITO, R., BOEKE, J.D., BUSSEY, H., CHU, A.M., CONNELLY, C., DAVIS, K., DIETRICH, F., DOW, S.W., EL BAKKOURY, M., FOURY, F., FRIEND, S.H., GENTALEN, E., GIAEVER, G., HEGEMANN, J.H., JONES, T., LAUB, M., LIAO, H. AND DAVIS, R.W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901–906.

WODICKA, L., DONG, H., MITTMANN, M., HO, M.H. AND LOCKHART, D.J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnology* **15**, 1359–1367.

WYRICK, J.J., HOLSTEGE, F.C.P., JENNINGS, E.G., CAUSTON, H.C., SHORE, D., GRUNSTEIN, M., LANDER, E.S. AND YOUNG, R.A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402**, 418–421.

YATES, J.R., 3RD (1998). Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* **33**, 1–19.

YUH, C.H., BOLOURI, H. AND DAVIDSON, E.H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902.

ZHANG, M.Q. (1999). Large-scale gene expression data analysis: A new challenge to computational biologists. *Genome Research* **9**, 681–688.