The Journal of Experimental Biology 210, 1507-1517
Published by The Company of Biologists 2007
doi:10.1242/jeb.004432

Extracting biology from high-dimensional biological data

John Quackenbush

Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA and Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

e-mail: johnq@jimmy.harvard.edu

Accepted 6 March 2007

Summary

The promise of the genome project was that a complete sequence would provide us with information that would transform biology and medicine. But the 'parts list' that has emerged from the genome project is far from the 'wiring diagram' and 'circuit logic' we need to understand the link between genotype, environment and phenotype. While genomic technologies such as DNA microarrays, proteomics and metabolomics have given us new tools and new sources of data to address these problems, a number of crucial elements remain to be addressed before we can begin to close the loop and develop a predictive quantitative biology that is the stated goal of so much of current biological research, including systems biology. Our approach to this problem has largely been one of integration, bringing together a vast wealth of information to better interpret the experimental data we are generating in genomic assays and creating publicly available databases and software tools to facilitate the work of others. Recently, we have used a similar approach to trying to understand the biological networks that underlie the phenotypic responses we observe and starting us on the road to developing a predictive biology.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1507/DC1

Key words: 'omic data analysis, microarray, bioinformatics, computational biology.

Introduction

Although the first draft of the human genome sequence was published in February 2001, we remain far from the promise of a genome-inspired revolution in our understanding of human health, development and disease. We are learning that the genome itself is far more complex than we had originally imagined, that variation between individuals is greater than first estimated, and that defining a complete collection of genes requires far more than elucidating likely protein-coding regions. While each of these is profound, solving any or all of them will not solve the fundamental problem of understanding how the program stored within the genome plays itself out as the organism grows, adapts, and responds to a wide range of stimuli. As we are learning, the situation is much more complex than we may have previously imagined.

One of the emerging principles in biology is that in most cases it is not individual genes but rather biological pathways and networks that drive an organism's response to a wide range of stimuli and the development of the range of phenotypes we observe. We are coming to understand that there are many diverse, but biologically significant networks, including metabolic networks, signal transduction networks and transcriptional regulatory networks, among others. In order to fully understand organisms and the manner in which they play out their genetic programs, we must develop tools and approaches to understand not only the structure of the networks that exist, but also the rules that govern their behavior and the interactions between elements in each biological system.

We are also coming to recognize that biological systems have a stochastic component that governs the fundamental interactions between molecules within the cell. Developing a full understanding of these processes is a significant challenge given the current limitations of our experimental techniques that most commonly look at millions of cells where we see an 'ensemble average' of the behavior occurring at a cellular level; an average that can obscure the random variance that occurs within cells. In most situations, this average behavior is enough to understand the biological systems that we study. However, a full understanding of disease processes or of a physiological response, in which systems move from their steady state conditions to other cellular states, will require that we account for stochastic events that push the system away from their preferred states.

In this review, we will examine some of these developments from the perspective of analyzing gene expression data. While technologies such as those supporting proteomics and

metabolomics studies are rapidly developing, analysis of gene expression using DNA microarrays and quantitative RT-PCR is much more mature and provides a natural starting point for developing approaches that will lead us to a new understanding of the fundamental principles in which biological systems function. Much of what is presented here derives from the integrative approach we have developed for the analysis of large volumes of data generated from microarrays, combining these with other, diverse sources of available information to produce a more complete understanding of the observed biological response. These have included integration of gene expression data with genetic mapping information (Cook et al., 2004), gene functional role classification and metabolic pathway assignments (Larkin et al., 2004), phenotypic classification (Flores-Morales et al., 2001; Malek et al., 2002; Shan et al., 2002), metabolic profiling patterns, and clinical data (Bloom et al., 2004; Eschrich et al., 2005).

Biological systems as information management systems

Biological systems carry out a wide range of complex tasks, from the level of metabolic and signaling pathways to that of the cell and to the entire organism. At every level, these processes require the coordination of diverse processes and the management of complex information. One way to view biological systems (Fig. 1) is to treat them as hierarchical systems in which information is stored and exchanged through the various levels, running from the DNA messages stored in cells through RNA and proteins to pathways and networks that maintain cells and their metabolic and signaling processes, and that these, in turn, influence how organisms themselves function. Further, one can investigate how genetic variation through populations influences phenotypes and how organisms interact with their environments to form ecologies.

With this intellectual framework, one can then envisage an integrative strategy for analyzing data obtained on such systems that builds on our long history of studying processes at all levels, from molecular biology on the smallest scale to medicine and ecology on the largest. Our work has shown that by combining information across these scales we can gain insight that would not be possible with data coming from any single domain. This use of ancillary information in analyzing gene expression data has proved essential in moving us along the path from data to knowledge and from knowledge to understanding.

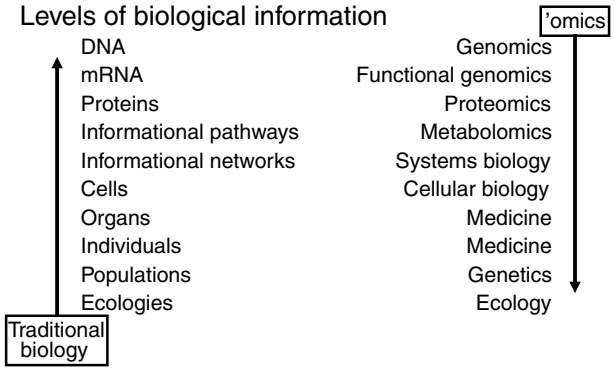

Fig. 1. Biological systems can be thought of as 'information management' systems with multiple levels of organization that interact and influence each other. The study of biological systems has a long history and although high-throughput ''omics' approaches are expanding their range of applications, integrating information from various levels can provide powerful insights for the interpretation of high-throughput data.

A natural starting point for organizing and integrating this information was the products of genome sequencing projects, including complete genome sequences that are being produced at an increasing rate for a wide range of eukaryotic species. The degree to which these genomes are truly complete varies across species, however, and missing genes, found in other mammals and amplifiable from genomic DNA, can even be identified within the now 'finished' human genome sequence. While these completed genomes represent a tremendous technological achievement, and while the process of completing the sequence and identifying and annotating the genes remains an ongoing process, we must remember that these sequences do not represent absolute and final truth. Rather, the complete genomes represent hypotheses that must be tested and validated. Similarly for the catalogues of genes that have emerged from genome sequencing projects; additional protein-coding genes remain to be found within the sequence, alternate splice forms are woefully under-identified, and non-coding and other functional RNA transcripts that may well confer important phenotypes are greatly undercounted.

The first step: associating probes, genes and annotation

As noted previously, our focus here will be on DNA microarray assays, although much of what will be presented can be generalized to other applications. As with any assay, the starting point here is to understand precisely what is being measured. On a DNA microarray, each individual element represents a distinct transcript and knowing which genes – and which other information about those genes – map to each element is essential for any real analysis of the data. Although the proliferation of robust commercial arrays makes it convenient to rely on the annotation for the probes that are provided by the manufacturer, even a 'catalog array' may have probes whose identity may still be in flux.

To address this problem, we have organized information from genome sequencing projects, the sequencing of expressed sequence tags (ESTs), and other information, into The Gene Index databases (Lee et al., 2005; Liang et al., 2000) (TGI; http://biocomp.dfci.harvard.edu/tgi), a collection of more than 100 species-specific databases representing likely transcripts in a wide range of eukaryotic species. Each of these freely available databases is constructed using open-source software tools (Pertea et al., 2003) using nearly identical protocols. EST and gene sequences from a species of interest are obtained from public repositories such as GenBank, cleaned to remove

contaminating vector and other sequences, and placed into groups based on shared sequence similarities. The resulting clusters are assembled at high stringency as elements of a 'transcriptome sequencing project'. The resulting Tentative Consensus (TC) sequences, largely similar to the transcriptional units described elsewhere in this issue by Piero Carninci (Carninci, 2007), are searched against public databases and subjected to a number of other analyses aimed at providing extensive annotation and identifying relevant biological features, including putative functional assignments, mapping to Gene Ontology terms (Ashburner et al., 2000) and KEGG pathways (Ogata et al., 1999), links to relevant records in PubMed, and identification of potential transcribed single nucleotide polymorphisms (SNPs).

Wherever possible, the TCs are mapped to available genome sequences, allowing identification of genes and splice variants – often supported by data from multiple species – that have not yet been annotated in the 'official' genome representations such as EnsEMBL (Hubbard et al., 2002). While this might seem trivial, understanding which splice forms exist can inform the results of a microarray experiment and provide valuable information for their interpretation (Larkin et al., 2005). Thus a more comprehensive list of transcripts, particularly in light of what we are learning about the importance of non-protein-coding transcripts (see papers by Mattick and Carninci, elsewhere in this issue) (Mattick, 2007; Carninci, 2007) is an important addition to genome annotation. Beyond this, mapping TCs to the genome provides the opportunity to integrate them with other information such as genetic linkage and quantitative trait loci (QTL) maps.

The Eukaryotic Gene Orthologue database (EGO; http://biocomp.dfci.harvard.edu/tgi/ego) builds on The Gene Index databases to create a consistent framework for comparative genomics. EGO attempts to identify orthologous genes across species and kingdoms using a parsimony-based approach that searches for the best sequence matches across three or more species (Lee et al., 2002). Building on the TGI and EGO databases, RESOURCERER takes widely used microarray platforms and genome sequence datasets such as National Center for Biotechnology Information (NCBI)'s RefSeq and provides extensive annotation and cross-referencing capabilities (Tsai et al., 2001). RESOURCERER annotates microarray resources with TC numbers, potential orthologues, genomic locations, GO terms, EC numbers, links to PubMed references, and a wide range of other information that can be used in data analysis. RESOURCERER also allows users to associate resources across platforms and across species, essentially providing a linking table that identifies probes on one array corresponding to probes on another, facilitating comparison between experiments. RESOURCERER also links microarray probes to genetically defined regions, associates microarray probes with QTL maps in mouse and rat, and allows other analyses, such as extraction of upstream promoter regions for probes that can be mapped to genome sequences. Versions of RESOURCERER exist for both animals (http://biocomp.dfci.harvard.edu/cgi-bin/magic/r1.pl) and plants (http://biocomp.dfci.harvard.edu/cgi-bin/magic/r1.pl).

Tools for data analysis

We have also invested significant effort in the development of tools to facilitate analysis of microarray data. This has resulted in a collection of sophisticated open-source tools known collectively as TM4 (Saeed et al., 2003) and freely available for download (http://www.tm4.org). TM4 consists of four primary software tools. MADAM is a comprehensive DNA microarray database and data entry interface that allows collection of the information relevant for any particular assay, including that required by the MIAME standard (Ball et al., 2002; Brazma et al., 2001). MADAM also exports MAGE-ML (Spellman et al., 2002) output that can be submitted to public databases such as NCBI's *GEO* and the European Bioinformatic Institute's *ArrayExpress*, as is often mandatory for publication. Spotfinder is an image processing tool written in machine-independent C/C++ for use with spotted two-color microarrays. Spotfinder includes a range of quality control tools to help users identify high and low quality assays and eliminate uninformative hybridizations. MIDAS provides a variety of normalization methods, including lowess (Yang, I. et al., 2002; Yang, Y. et al., 2002) and variance regularization (Huber et al., 2002; Yang, Y. et al., 2002), as well as a number of data filtering options. To document the process, MIDAS produces a PDF log file containing a complete record of all the analyses and parameters along with diagnostic plots and summary statistics.

The most widely used tool in the TM4 suite is MeV, a data-mining tool that combines a number of clustering and statistical algorithms in an easy to use menu-driven format. Users of MeV can load data, use a *t*-test or Significance Analysis of Microarrays (SAM) (Tusher et al., 2001) to identify genes that correlate with the phenotypes under study, and explore relationships between gene expression profiles using hierarchical clustering. Gene sets identified during analysis can be subjected to a meta-analysis using EASE (Hosack et al., 2003) which, as described below, looks for over-represented Gene Ontology terms (GO terms; http://www.geneontology.org) and KEGG pathways (http:// www.genome.jp/kegg) in the set relative to their representation on the array. MeV is undergoing continuous improvement to include a wide range of new algorithms, and new releases happen at least twice per year.

Examples of integrative analysis

Assembling these databases and creating analysis tools is, however, only the first step. For these to be of value, they must be both useful and used. Consequently, in developing these tools, we have focused on addressing real biological problems. Here we describe two examples requiring very different analytical strategies that illustrate how these tools have been useful in the analysis of gene expression data and its

interpretation through integration with other sources of information.

Integration of expression with genetic mapping: innate immunity in the mouse

The innate immune response represents an organism's first line of defense against bacterial infection. Key elements in the innate immune system are the cell surface receptors known as the Toll-like receptors; in the context of this study, the key receptor is Toll-4 (TLR4) which, when it detects the presence of LPS, the lipopolysaccharide on the coats of Gram-negative bacteria, triggers a well-characterized signaling cascade resulting in recruitment of inflammatory cells, triggering of the adaptive immune response, and generation of a number of patho-physiological states, including asthma. Our collaborators, Donald Cook and David Schwartz, had been studying a model of innate immunity in the mouse and had identified two strains, C57/BL6 and DBA/2J which, despite possessing wild-type TLR4 receptors, had very different phenotypic responses to inhaled LPS. Their interest was in understanding the mechanisms that drive these differential responses, and in examining the responses of the BXD recombinant-inbred strains, derived from C57/BL6 and DBA/2J, they saw a spectrum of responses spanning the range between the parentals and extending beyond it. This suggested that the response to LPS was mediated by interaction of multiple genes, and so a genome-wide microarray-based approach promised the opportunity to discover genes whose expression might ultimately lead to the observed phenotypes.

The parental strains, the three highest-responding, and three lowest-responding strains, were exposed to LPS and compared to matched control animals using a spotted cDNA microarray containing nearly 27 000 gene-specific probes (Cook et al., 2004). This identified approximately 425 genes that were significantly differentially expressed between the high- and low-responding strains. At this stage, we faced a problem common in the analysis of microarray data: placing a long list of responsive genes into a biological context. We chose to develop QTL maps, identifying genetic markers whose inheritance correlates with the severity of the phenotype. Using a series of F2 crosses and building on the exquisite genetic resources available in mouse, we identified two regions in the mouse genome containing approximately 525 genes that could be linked to differences in either of our phenotypic measures: TNF-α levels or the recruitment of peripheral mononucleocytes to the lung (as measured by bronchial lavage). We then used RESOURCERER to compare the two lists, looking for genes that were both genetically linked to the response and differentially expressed. We expected, by chance, 13 genes in the overlapping set but, in fact, we found 46 ($P<10^{-20}$), clearly suggesting that these genes were significantly linked to expression of the phenotype.

We then used quantitative RT-PCR to compare expression for these 46 genes between LPS-exposed and control animals in the parental and BXD strains (Fig. 2). It is worth noting that if one orders the strains by increasing the number of upregulated genes, this closely resembles the ordering based on magnitude of the phenotypic response, consistent with them being related. We also realized that the gene expression profiles themselves could be used as quantitative traits for constructing QTL maps, something beautifully demonstrated by Eric Schadt and his group at Rosetta in a paper where they defined precisely this concept, defining the 'expressed QTL' (eQTL) and showing that one could use it to find interactions between genes in producing a phenotype (Schadt et al., 2005).

One might ask whether gene expression and genetic analysis together give us enough information to completely understand

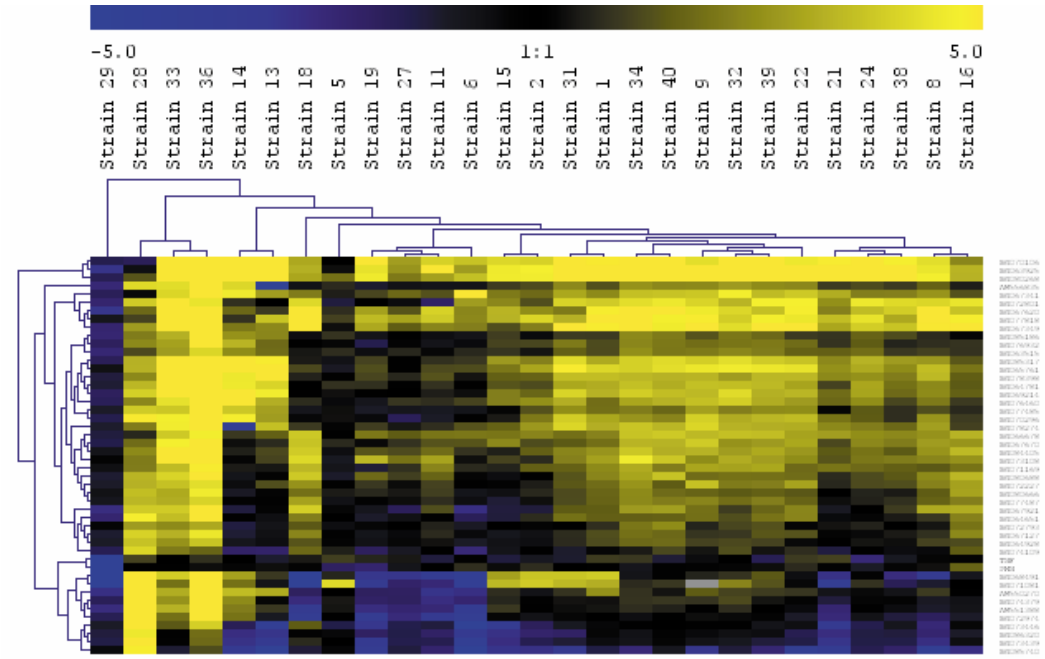

Fig. 2. Genes identified as both differentially expressed and also genetically linked to the differential response to inhaled LPS in a mouse model of environmentally induced asthma. Responses were measured by qPCR, comparing exposed mice to matched controls from the same strain. The ordering of the strains by expression levels for these genes closely mimics that produced when ordered by phenotype.

the observed phenotypes. The answer, not surprisingly, is that they do not. An example of the limitations of this approach can be seen in the BXD 29 strain. This strain shows the lowest phenotypic response using any measure and in Fig. 2 one can see that it hardly produces any transcriptional response following LPS exposure. This was puzzling because we expected to see at least some response as both parental strains, and therefore the BXD strains, should have wild-type TLR4 receptors. When we sequenced the TLR4 receptor gene in BXD29, however, we discovered that it had developed a spontaneous mutation rendering it insensitive to LPS exposure. As one might expect, use of QTL and gene expression analysis may miss key genes linked directly to the phenotype through mutation but which are not themselves differentially expressed.

Extracting meaning using GO-term meta-analysis

Another approach to placing a list of genes into a relevant biological context is to use the annotation for those genes in sophisticated ways. In collaboration with Haralambos Gavras of Boston University, we analyzed gene expression in a mouse model of hypertension and used Gene Ontology assignments to facilitate the interpretation of the data (Larkin et al., 2004). Hypertension is a significant disease affecting nearly one in four Americans and is strongly associated with heart disease, the leading cause of death in the United States. Angiotensin II (Ang II) is a significant contributor to the development of hypertension and target organ damage and is known to produce vasoconstriction and increased blood pressure following acute exposure, and cardiac necrosis, fibrosis and hypertrophy following chronic exposure. Ang II is produced by cleavage of Ang I by Ang I Cleavage Enzyme (ACE); ACE inhibitors are one of the primary therapeutic classes used in the treatment of hypertension.

Using a mouse model of acute and chronic exposure to Ang II developed by Gavaras and his coworkers, we used DNA microarrays to compare patterns of gene expression that were induced following both acute (24 h) and chronic (2 week) exposure to that of matched controls (Larkin et al., 2004). DNA microarray expression profiles were analyzed using SAM (Tusher et al., 2001) to identify genes significantly up- and downregulated in cardiac tissue for both acute and chronic treatments, as well as those that were generally responsive to Ang II. To make sense of the numerous gene lists, we used functional class assignments based on assigned Gene Ontology terms (GO) (Ashburner et al., 2000). GO attempts to describe each encoded protein by the molecular function it carries out, the biological process in which it participates, and the cellular component to which it is localized. We also examined assignments of gene products to metabolic and other pathways in both the KEGG (Ogata et al., 1999) and GenMapp databases (Dahlquist et al., 2002).

Rather than simply producing lists of associated terms and pathways, we instead looked for classes that were over-represented relative to the population of probes on the array. We used EASE (Hosack et al., 2003), a method developed by Doug Hosack and Glynn Dennis at the National Institute of Allergy and Infectious Disease (NIAID) of the US National Institutes of Health, which was integrated into MeV (Saeed et al., 2003). EASE uses a Fisher Exact test to compare the fractional representation of any one class within a set of 'significant' genes to its representation on the array, estimating the probability that any group is over-represented by chance. For example, if 30% of the genes on an array were energy metabolism genes, then in any randomly selected set of genes, we would expect about 30% of them to be related to energy metabolism. However, if we found 70% or 75% of the genes in our set to be classified as energy metabolism, this would be very suggestive that the phenotypes we are analyzing are related to changes in expression of energy metabolism genes.

Fig. 3 shows a heat map and hierarchical clustering dendrogram for the GO term assignments found to be significant with the heat map showing $-\log_{10}(P\text{-values})$ based on EASE analysis. This analysis associated a wide range of biological responses to Ang II treatment and suggested a potential mechanism. One surprising result was the identification of group of genes significantly downregulated in acute Ang II exposure but upregulated following chronic exposure and which mapped to significantly over-represented KEGG pathways annotated as Alzheimer disease and neurodegenerative disorders pathways. The Alzheimer disease pathway is particularly interesting, as Alzheimer disease plaques and heart disease have been shown in clinical studies to co-occur (Sparks et al., 2000). Heart disease and hypertension may be a forerunner to Alzheimer disease, and Alzheimer disease-like β-amyloid plaques have been found in the brains of non-demented individuals with heart disease (Sparks et al., 2000). Furthermore, cardiac amyloidosis is known to cause restrictive cardiomyopathy (Artz and Wynne, 2000), a particularly lethal form that does not respond to standard treatments. This was the first analysis to report the upregulation of a brain-specific pathway in cardiac tissue, but a pathway has been clinically identified as a pre-determining risk factor for a neurological disorder. This suggests that the investigation of therapeutics developed for the treatment of plaque formation in Alzheimer disease may act as a viable alternative therapy for hypertension or cardiac amyloidosis. It also argues for the proactive treatment of hypertension as a way of reducing risk for Alzheimer disease. The interrelationship between heart disease and Alzheimer disease and their treatment has already been demonstrated by the effects of statins, used to lower cholesterol and treat ischemic heart disease, on Alzheimer disease. Although some statins appear to be protective against subsequent development of Alzheimer disease, there are also indications that patients with Alzheimer disease may be more susceptible to adverse effects of statins than are age-matched controls (Algotsson and Winblad, 2004).

Although the methods employed in this analysis did provide important insight, their application was somewhat limited. While broad classes of genes can provide insight into the general response of an organism to a particular stimulus, they are not precise enough to provide direct inferences of

mechanism. Rather, they provide hypotheses that can be further refined and tested in the laboratory. Based on our experience with EASE, we also developed MeSHer (http://biocomp.dfci.harvard.edu/mesher.html) (Djebbari et al., 2005), which uses Medical Subject Headings assigned to PubMed references associated with particular genes, but this approach too suffers from similar limitations. Further, both EASE and MeSHer are limited by the scope of functional assignments, as many genes still lack functional classes, and for most genes the classification is not comprehensive. What are truly needed are methods that will allow more comprehensive analysis of the mechanisms associated with the phenotypes we observe based on the pathways and networks that are underlie cellular processes. Nevertheless, while suggesting potential new areas of investigation, associations discovered through such analyses remain hypotheses that must be validated experimentally.

Modeling pathways and networks

Organisms use a combination of *cis*- and *trans*-acting elements to respond to intra- and extracellular environmental stimuli by regulating gene and protein expression. The biological process of transcription begins with the binding of transcription factors to specific sequence motifs lying upstream to a gene's transcription initiation site. This induces conformational changes in the DNA and initiates the assembly of the RNA polymerase complex. This process is rather complex, with promoters, inhibitors and enhancers interacting in complex ways in the regulation of gene expression. One consequence of transcriptional activation is that the levels of transcription factors themselves can be affected through the same promotion and repression mechanism. What emerges is not a single set of interactions, or even a single pathway, but a complex network of interacting genes and gene products. In principle, it is this network and the interactions between its components that we would like to understand, since this underlies the way in which organisms respond to environmental and other cues.

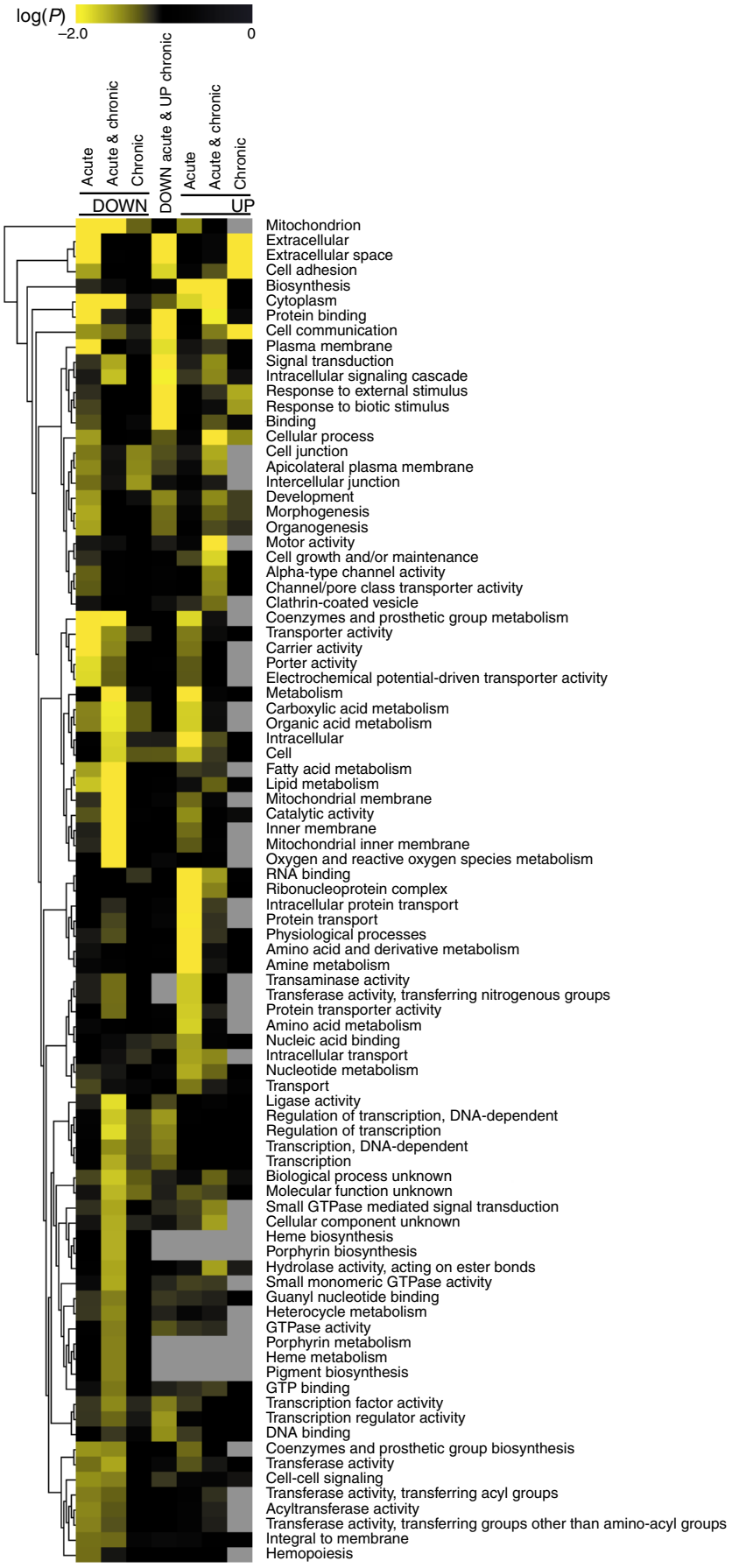

Fig. 3. Heat map and hierarchical clustering dendrogram, in which the elements being clustered are the GO term assignments and the values represented in each cell are the $-\log_{10}(P\text{-values})$ of this being significantly different from the null hypothesis, based on EASE analysis. From Larkin et al. (Larkin et al., 2004).

A natural approach to modeling this process is to assume that there is some logical combination of elements that must be present in the cell in order to initiate transcription. Fundamentally, this process is governed by a network of interacting components connected by pair-wise interactions. One way to conceptualize these interactions is to represent the components as 'nodes' connected to each other by 'links' or 'edges' (most often directed edges), such that the edges represent the interactions between any two components. In this way, we can combine elements such as genes, proteins, metabolites and other factors, and represent them as a network, or graph, that can be modeled in a variety of ways.

Such a graph can represent the relationships between elements within a network on two levels. First, the nature of the interactions between the elements can be reflected in the architecture of the graph – the patterns of nodes and edges – and the conditional relationships between the graph elements, so that we know which respond to which others. Second, a set of network parameters can describe the strength of the dependencies between elements. For example, if two nodes, A and B, are connected by an edge, then A and B are dependent in some way. Such an interaction might model a transcription factor, B, that activates the expression of a particular kinase, A. But if A and B are separated by a third node C, then A and B are conditionally independent, given C. Here one might imagine that a transcription factor B activates expression of a second transcription factor C, which in turn induces expression of the kinase A. To model biological systems, we have to define the rules by which B activates C and C activates A, taking into account important physical parameters such as expression levels. We can, of course, build up even more complex interactions, such that any node can have multiple incoming edges and, consequently, its activation can depend on complex interactions between those input edges. To model biological systems, the challenge is to derive both the network architecture and the set of rules by which the inputs to any node interact to produce the output.

These rules can take on a variety of complex forms, and several gene network modeling techniques have been applied to the analysis of microarray data, including weight matrices (Weaver et al., 1999), Boolean networks (Akutsu et al., 1999), and differential equations (Chen et al., 1999), but Bayesian networks (BNs), in particular, have shown the greatest promise in the analysis of expression data (Friedman et al., 2000). Formally, Bayesian networks are directed acyclic graphs (DAGs) in which nodes represent random variables (typically genes or gene products, referred to hereafter as genes) and directed edges represent dependencies between variables (interactions between the genes); conditional probability distributions associated with each node imply conditional independence statements that describe how the state of one gene influences the state of another. Bayesian networks are particularly appropriate to the study of biological systems as the underlying variables are probabilistic, can be discrete (on/off) or continuous, and describe variation across conditions. Example variables include mRNA concentrations, protein concentrations, protein modifications or complexes, metabolites, experimental conditions, genotypic information, or phenotypes such as prognoses or drug susceptibility.

Gene 1	Gene 2=1|Gene 1
–1	0.1
0	0.2
1	0.7

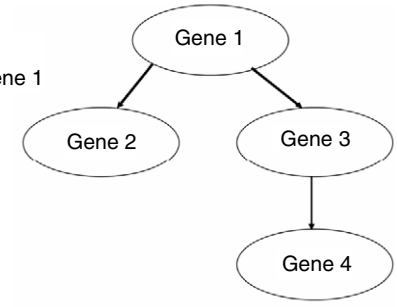

Fig. 4. An example of a Bayesian Network model for four genes. If we assume that Gene 1 activates Gene 2, then we can construct a conditional probability table (shown at the right) that captures our observations of the state of Gene 1 when we observe Gene 2 to be upregulated. Here the values for Gene 1 of –1, 0, and +1 represent states where Gene 1 is downregulated, unchanging or upregulated, respectively.

Consider a simple model of a Bayesian Network (Fig. 4), in which we assume that Gene 1 controls other genes. If we focus on Gene 2, its 'parent' is Gene 1. We can describe the observed relationship between these two genes by constructing a conditional probability table showing the likelihood that we observe Gene 2 to be upregulated (Gene 2=1) given that we also observe the state of Gene 1. In the example in the figure, we find that the probability that Gene 2 is upregulated given Gene 1 is upregulated is 0.7. This value may be interpreted as Gene 1 activating Gene 2. Bayesian Networks encode dependencies in the data such that the only dependencies are between a gene and its direct parents.

Applying a BN approach to microarray data analysis is challenging for a number of reasons. Most notably, the Bayesian network analysis requires that we deduce the structure of the network graph from the available data and then use the structure to define conditional probability tables that describe the interactions between genes. In learning the structure, one must consider all possible network topologies and this is a computationally intractable problem as the number of possible graphs is super-exponential in the number of nodes; formally, learning Bayesian networks is NP-hard, meaning that the number of possible network structures to be tested is so large that it is not amenable to exact computational solution. Consequently, most approaches use heuristic search algorithms that start with a random graph and look at perturbations on the network structures in order to find the best network given the data. The problem with this approach is that it is susceptible to getting 'trapped' at a local maxima and so not finding the global optimal network. As a result, BN analysis of most biological datasets produce networks that have little resemblance to real biological networks and, despite their initial promise, this has severely limited their applicability.

Based on our experience with the analysis of complex data, we realized that one could use a more intelligent approach to

'seeding' the search for network structure based on integration of prior knowledge. Recently, we demonstrated that one can use prior best guesses as to the network structure, derived from the biomedical literature or protein–protein interaction (PPI) datasets, or combinations thereof, to learn biologically realistic networks that reproduce known biological interactions and pathways relevant to the disease state under analysis (A. Djebbari and J. Quackenbush, manuscript submitted), an approach that is also described by Lehner in this issue (Lehner, 2007). In our method, gene networks were constructed from the literature using the co-occurrences method (Jenssen et al., 2001) by defining genes as nodes and connecting them with an edge if they are mentioned in the same article, assigning edge weights based on the number of articles mentioning those genes. Given a set of genes of interest, one can construct a literature network by taking the union of edges in networks constructed from databases including Entrez Gene and PubMed.

PPI networks can also be used as constraints on Bayesian Networks topology. The recently published CCSB-HI1 dataset (http://vidal.dfci.harvard.edu/HIPaperSup) detected interacting proteins using a high-throughput yeast two-hybrid assay (Rual et al., 2005). While this is one of the most comprehensive surveys of the human interactome, its relatively low coverage (2754 edges representing binary interactions) represents approximately 1% of the interactome; in order to explore the potential benefit of using PPI as prior information for learning Bayesian Networks, we allowed our starting 'significant' gene lists to expand by including all genes in the interactome dataset with k or fewer links using Floyd's all-pairs shortest paths algorithm.

We then apply this method to deducing Bayesian Networks from gene expression data. To avoid the over-fitting problem that arises from learning Bayesian Networks from too few samples, one can generate many networks and perform model averaging to find important features that are supported by the data. To this end, we used a bootstrapping approach to estimate the confidence in features learned. The bootstrapping method consists of resampling the data with replacement (a non-parametric bootstrap) (Friedman et al., 2000) to estimate the confidence in features learned. The features considered are: directed edges, undirected edges, order relations (one variable is the ancestor of the other variable) and Markov relations (if two variables are connected either way or if they are both parents of another variable). If the feature is strongly induced from the data, the confidence of this feature is expected to be closer to 1 or otherwise closer to 0.

We considered the dataset presented by Golub et al. as a test set (Golub et al., 1999), where the authors analyzed two forms of leukemia, acute lympboblastic leukemia (ALL) and acute Myeloblastic leukemia (AML). Using a simple between-groups t-test, we identified a set of 40 genes that best distinguished the

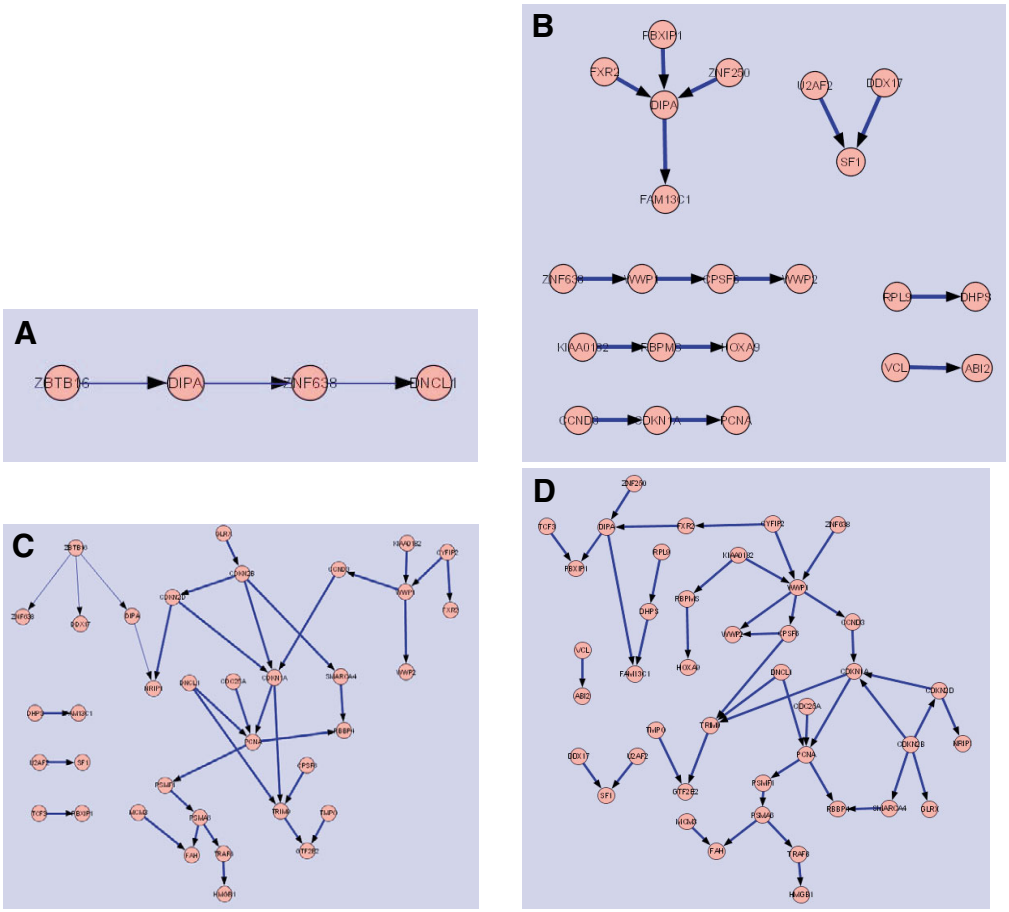

Fig. 5. Representations of the networks produced by a Bayesian Network analysis of the top 40 genes selected as distinguishing ALL and AML in the microarray dataset of Golub et al. (Golub et al., 1999) for links with confidence greater than 0.7; links with confidence greater than 0.9 are shown in bold. Networks represent the consensus of 200 iterations for (A) the microarray data alone, (B) the microarray data with constraints from protein-protein interaction (PPI) data, (C) microarray data with constraints from literature networks, and (D) microarray data with constraints from a combination of microarray and PPI data.

two disease states and applied our Bayesian Network formalism. For each of 200 bootstrap iterations, we ran a Bayesian Network heuristic search algorithm using the Bayesian Network package from within the WEKA toolkit (Witten and Frank, 2005), which has been integrated into the MeV software package developed by our group. We used a variety of combinations of prior network structures arising from the literature, PPI data, and a combination of the two. For each graph feature, we counted the number of times it occurs for each pair of genes over the total number of iterations. Taking the null hypothesis that the no prior and prior cases provide the same confidence estimates for each gene pair (as computed by taking the average for all the combinations of algorithms and scoring schemes), we perform a two-tailed paired *t*-test to find the corresponding *P*-value for each comparison. The results of this analysis are shown schematically for the four cases in Fig. 5.

In the final network combining both the literature and PPI data, many of the genes play important biological roles associated with the cell cycle. MCM3 is critical in S-phase cell cycle progression, RBBP4 in chromatin remodeling and GTF2E2 in transcription. Many of these genes are either directly or indirectly involved in the Rb/E2F pathway. Rb and p53 are tumor-suppressor genes than can check cell-cycle progression and prevent cells from becoming cancerous. E2F is a transcription factor required for the expression of proteins involved in dNTP and DNA synthesis. In normal cells, Rb is hypophosphorylated and complexes with E2F, blocking it from activating transcription. In this way, inhibition of E2F activity by Rb can block entry into the S phase. Rb is inactive when hyperphosphorylated, however, thereby releasing E2F and the cell cycle progresses. This over-representation of cell-cycle genes suggests that, indeed, differences in expression of cell-cycle-related genes are responsible for some of the observed differences in ALL and AML phenotypes.

It must be noted that these networks are not biological networks in the sense of metabolic pathways, signal transduction networks or biochemical pathways. Rather, the networks produced from a Bayesian network analysis provide us with a framework for understanding the interactions between genes or their encoded proteins and for building predictive models that can be used to evaluate how an organism might respond to perturbations in its environment. Our work demonstrates that the use of prior information, derived from the published literature, PPI data, or both, can improve our ability to learn realistic networks from gene expression data. What is most encouraging about this work is that the starting data, a comparison of AML and ALL, was not designed to probe the cell-cycle network, but this network emerged from the data nonetheless. This suggests that directed experiments where a particular network or pathway is perturbed and followed over time may further improve the overall performance of a BN approach. Using such an approach in an iterative manner, in which a network is first learned, then perturbed and the resulting data used to refine the predicted network structure, may allow us to discover novel players in many known networks and to learn previously unknown networks from DNA microarray expression profiles.

Stochastics in gene expression: the foundation of systems biology

Recently, much of the focus in genomics has moved to developing models of cellular systems that extend beyond the ‘parts list’ provided by the genome and the types of relationship-based models represented in Bayesian Networks. Systems biology has focused on developing quantitative and predictive models describing cellular systems. However, the network models developed to date are largely deterministic, meaning that if the right initial conditions are met and the right interactions are represented, then the model will predict a specific outcome. What these models ignore is the fact that biological systems have inherently stochastic components. In 1995, McAdams and Shapiro attempted to model one of the simplest organisms, lambda phage, and realized that stochastic inputs to the system made reliable prediction of outcome nearly impossible (McAdams and Shapiro, 1995).

Evidence for stochastic processes in biology has been mounting for quite some time. There are a number of reports indicating that protein production has a stochastic component that gives rise to very different rates of protein synthesis in genetically identical cells in essentially identical environments (Blake et al., 2003; Elowitz et al., 2002; Ozbudak et al., 2002). However, until recently there has only been a single published report of the variability of gene expression in single cells, which did not provide an underlying statistical model for mRNA representation within the cell (Levsky et al., 2002). While this may seem to be minor, it represents a significant gap in our knowledge if we are to construct the sort of predictive models that are the aim of systems biology.

To address this problem, we turned our attention to understanding the stochastic nature of steady-state gene expression. We tend to think of a tissue sample as being homogeneous and to discuss levels of gene expression in terms of absolute numbers of copies per cell. However, every transcription factor has a dissociation constant (K_D) that can be measured, which implies that the transcription factor binds and unbinds with some fixed probability per unit time. There is a long history of modeling such stochastic processes using Poisson statistics, so it was a natural assumption to consider transcription as a Poisson process. While measuring transcription in individual cells is a challenge, we recognized that if we looked at small numbers of cells, we might be able to see echoes of the underlying stochastic processes in individual cells. Sampling statistics applied to Poisson processes tell us that the variance in gene expression levels should decay as $1/N$, where N is the number of cells sampled.

We developed an approach we refer to as ‘mesoscopic biology’, which looks between the macroscopic and microscopic (single cell) realms. Using quantitative RT-PCR, and sampling variable numbers of cells, we were able to demonstrate that steady state gene expression does, in fact,

obey Poisson statistics (Mar et al., 2006). The beauty of this approach is that it can provide experimental measurements even for genes expressed at very low levels. It further suggests that other stochastic events occurring in single cells, even complex interactions in pathways, may reveal themselves through the analysis of samples of mesoscopic size. In many ways, this situation is analogous to one in statistical mechanics and thermodynamics – the relationship between the Maxwell–Boltzman distribution and the Ideal Gas Law. While we understand that the Ideal Gas Law describes gas dynamics for macroscopic samples, we know that, on a microscopic scale, the behavior of the gas molecules themselves is described by the Maxwell–Boltzman distribution. In a biological system, the compromise between looking at large tissue samples, where the stochastic events are averaged out, and single cells, where experimental analysis is difficult, is to look at small numbers of cells – mesoscopic samples – where one can begin to see deviations from the average behavior. And although a first step, we believe that this is a crucial one for developing an understanding of the way in which biological systems operate.

Conclusions

The biological sciences are in a state of rapid development, driven largely by the tools that have developed as byproducts of the Human Genome Project. These new technologies are profoundly changing the manner in which we approach a wide range of problems and questions, and demanding that we develop new methods that will allow us effectively to manage the data they produce. In many ways, the true revolution inspired by genomics has been one that is changing what was exclusively a laboratory science into an information science. What we have presented here is a cursory overview of some of the ways in which we have attempted to deal with this transition and is in no way meant to be a comprehensive review of the field. There are, for example, many freely available software systems for the analysis of gene expression and other genomic data, the most notable being the BioConductor package developed in R (Gentleman et al., 2004). While not being exhaustive, we hope that the examples we have provided illustrate the challenges presented by the analysis of genomic data and some possible ways of addressing them.

We also hope that we have provided some insight into the complexities of modeling biological systems and the challenges inherent in the growing discipline of systems biology. While there are many problems remaining to be solved, the focus on developing predictive models promises to move us toward a more secure analytical framework in which to study processes relevant to a wide range of enquiries, including analysis of the mechanisms underlying human disease.

In all of this, it is important to remember that every experiment is an attempt to understand a real biological system. The new approaches and new technologies that have become so widespread in the past years cannot in and of themselves provide us with biological insight. In fact, in any large-scale analysis, the best result we can generally hope for is the development of new, testable hypotheses that can lead us back to directed experiments in the laboratory. The good news is that genomic technologies have provided us with a ‘macroscope’ that allows us to consider biological systems holistically, examining their entire gene content in a single assay, and in so doing has opened up new areas of investigation and provided us with many new and exciting, testable, hypotheses.

This work was supported by grants from the National Library of Medicine and the National Science Foundation.

References

Akutsu, T., Miyano, S. and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* **1999**, 17-28.

Algotsson, A. and Winblad, B. (2004). Patients with Alzheimer’s Disease may be particularly susceptible to adverse effects of statins. *Dement. Geriatr. Cogn. Disord.* **17**, 109-116.

Artz, G. and Wynne, J. (2000). Restrictive cardiomyopathy. *Curr. Treat. Options Cardiovasc. Med.* **2**, 431-438.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29.

Ball, C. A., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H. C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F. et al. (2002). Standards for microarray data. *Science* **298**, 539.

Blake, W. J., Kærn, M., Cantor, C. R. and Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature* **422**, 633-637.

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., Quackenbush, J. and Yeatman, T. J. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am. J. Pathol.* **164**, 9-16.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C. et al. (2001). Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* **29**, 365-371.

Carninci, P. (2007). Constructing the landscape of the mammalian transcriptome. *J. Exp. Biol.* **210**, 1497-1506.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling gene expression with differential equations. *Pac. Symp. Biocomput.* **1999**, 29-40.

Cook, D. N., Wang, S., Wang, Y., Howles, G. P., Whitehead, G. S., Berman, K. G., Church, T. D., Frank, B. C., Gaspard, R. M., Yu, Y. et al. (2004). Genetic regulation of endotoxin-induced airway disease. *Genomics* **83**, 961-969.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31**, 19-20.

Djebbari, A., Karamycheva, S., Howe, E. and Quackenbush, J. (2005). MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. *Bioinformatics* **21**, 3324-3326.

Elowitz, M. B., Levine, A. J., Siggia, E. D. and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* **297**, 1183-1186.

Eschrich, S., Yang, I., Bloom, G., Kwong, K. Y., Boulware, D., Cantor, A., Coppola, D., Kruhoffer, M., Aaltonen, L., Orntoft, T. F. et al. (2005). Molecular staging for survival prediction of colorectal cancer patients. *J. Clin. Oncol.* **23**, 3526-3535.

Flores-Morales, A., Stahlberg, N., Tollet-Egnell, P., Lundeberg, J., Malek, R. L., Quackenbush, J., Lee, N. H. and Norstedt, G. (2001). Microarray analysis of the in vivo effects of hypophysectomy and growth hormone treatment on gene expression in the rat. *Endocrinology* **142**, 3163-3176.

Friedman, N., Linial, M., Nachman, I. and Pe’er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601-620.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

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.

et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Hosack, D. A., Dennis, G., Jr, Sherman, B. T., Lane, H. C. and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4**, R70.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. et al. (2002). The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38-41.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96-S104.

Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* **28**, 21-28.

Larkin, J. E., Frank, B. C., Gaspard, R. M., Duka, I., Gavras, H. and Quackenbush, J. (2004). Cardiac transcriptional response to acute and chronic angiotensin II treatments. *Physiol. Genomics* **18**, 152-166.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R. and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337-344.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J. et al. (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12**, 493-502.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., Antonescu, V., Chan, A., Cheung, F. and Quackenbush, J. (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* **33** Database Issue, D71-D74.

Lehner, B. (2007). Modelling genotype–phenotype relationships and human disease with genetic interaction networks. *J. Exp. Biol.* **210**, 1559-1566.

Levsky, J. M., Shenoy, S. M., Pezo, R. C. and Singer, R. H. (2002). Single-cell gene expression profiling. *Science* **297**, 836-840.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239-240.

Malek, R. L., Irby, R. B., Guo, Q. M., Lee, K., Wong, S., He, M., Tsai, J., Frank, B., Liu, E. T., Quackenbush, J. et al. (2002). Identification of Src transformation fingerprint in human colon cancer. *Oncogene* **21**, 7256-7265.

Mar, J. C., Rubio, R. and Quackenbush, J. (2006). Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples. *Genome Biol.* **7**, R119.

Mattick, J. S. (2007). A new paradigm for developmental biology. *J. Exp. Biol.* **210**, 1526-1547.

McAdams, H. H. and Shapiro, L. (1995). Circuit simulation of genetic networks. *Science* **269**, 650-656.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29-34.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nat. Genet.* **31**, 69-73.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B. et al. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N. et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M. et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-378.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C. et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710-717.

Shan, L., He, M., Yu, M., Qiu, C., Lee, N. H., Liu, E. T. and Snyderwine, E. G. (2002). cDNA microarray profiling of rat mammary gland carcinomas induced by 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine and 7,12-dimethylbenz[a]anthracene. *Carcinogenesis* **23**, 1561-1568.

Sparks, D. L., Martin, T. A., Gross, D. R. and Hunsaker, J. C., 3rd (2000). Link between heart disease, cholesterol, and Alzheimer's disease: a review. *Microsc. Res. Tech.* **50**, 287-290.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M. et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, research0046.

Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., Cho, J., Parvizi, B., Cheung, F. and Quackenbush, J. (2001). RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol.* **2**, SOFTWARE0002.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116-5121.

Weaver, D. C., Workman, C. T. and Stormo, G. D. (1999). Modelling regulatory networks with weight matrices. *Pac. Symp. Biocomput.* **1999**, 112-123.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kauffman.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J. et al. (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* **3**, research0062.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.