Genomic Perspective and Cancer

D. BOTSTEIN
Lewis-Sigler Institute, Princeton University, Princeton, New Jersey 08544

The discovery of the double-helical structure of DNA, and the concomitant realization that DNA molecules carry genetic information digitally encoded in their nucleotide sequences (Watson and Crick 1953a,b,c), neatly divide the history of biology in the 20th century. The first half featured the rise of classical genetics: analysis of the inheritance of traits obtained at first from natural variation and later by induced or selected mutations. Much was learned from breeding studies in plants and simple "model" organisms such as *Drosophila* and *Neurospora,* including quite detailed genetic linkage maps. The second half saw the rise of molecular biology: the elucidation, in considerable detail, of the information pathway that begins with the nucleotide sequence in DNA and ends with the specification of the phenotypes of cells and organisms.

The determination of the complete nucleotide sequences of the human, mouse, and many bacterial and eukaryotic model organism genomes is the logical culmination of both molecular biology and classical genetics. Just as the DNA structure transformed classical genetics into a molecular science, the genomic sequences are transforming molecular biology into an information science, marking the beginning of a third era in the history of biology. In each case, the newer science is of necessity firmly based in its predecessor, but thinking and research become transformed both in style and substance; each era adds a new perspective to our understanding of biology.

INTELLECTUAL ORIGINS OF GENOMICS

The ideas underlying the science of genomics can be traced back to the early years of molecular biology. The first truly "genomic" paper was presented at the Cold Spring Harbor Symposium in 1963; it summarized the results of a deliberate program to identify all the genes of bacteriophage T4 (Epstein et al. 1964). The paper described, in a general way, what each of the T4 genes does for the organism. This program was based on the idea that one might be able to obtain mutations in all the essential T4 genes by isolating conditional-lethal mutations. Two kinds of conditional-lethal mutations (chain-terminating and temperature-sensitive) had recently been described; strong arguments were made for the idea that either or both of these kinds of mutations could be found in any essential phage gene if one looked hard enough for them. The T4 genes themselves were then defined and enumerated genetically: The mutations were classified into genes by complementation and recombination mapping using their common conditional-lethal phenotypes. This was clearly a lot of work, even for an organism expected to have no more than about 100 genes. It thus seems worth noting that, like modern genomics papers, Epstein et al. (1964) was the result of an international collaboration among several laboratories and had 10 authors (remarkably many, in 1963).

Figure 1 is a composite, consisting of an electron micrograph of phage T4 on the left, and a diagram from Epstein et al. (1964). The genes are shown in the order they appeared on the circular T4 genetic linkage map, along with an abbreviation or ideogram that describes the outcomes of infections with mutant phages under nonpermissive circumstances: D0 for no DNA synthesis, DA for DNA synthesis arrest, MD for maturation defective, and ideograms for the presence of heads, tails, unassembled heads and tails, and various kinds of assembled, but defective, phage particles. Updates of this figure served, for many years, as the genome database for bacteriophage T4. A similar genomics program was successfully carried out with several other bacterial viruses as well, notably bacteriophages λ, P22, and ϕX174, and a few animal viruses (e.g., polyoma, adenovirus, and herpesvirus). In each of these cases, most (if not quite all) essential viral genes were identified in advance of any sequencing, and their biological roles were defined in at least a general way.

Shortly after the appearance of Epstein et al. (1964), two substantial efforts were undertaken to identify all the genes of two free-living organisms (the yeast *Saccharomyces cerevisiae* and the nematode worm *Caenorhabditis elegans*), despite the expectation that their genes would number in the many thousands. Once again, gene enumeration was to be via conditional-lethal mutations and classical genetic methods (complementation and recombination mapping). These efforts, led by Leland Hartwell and Sydney Brenner, respectively, also had substantial success well before the DNA sequence era (Hartwell 1970, 1974, 1978; Brenner 1974). It was this success that attracted a large and productive research community of molecular biologists to the study of these organisms. These active research communities and the progress they made in biology made it logical, even inevitable, that these would be the model organisms whose genomes would be sequenced first.

As molecular sequences accumulated, it became clear that the sequences and functions of most genes and proteins are strongly conserved in evolution. Today, the findings of the bacterial, yeast, worm, and other model organism research communities about individual genes and

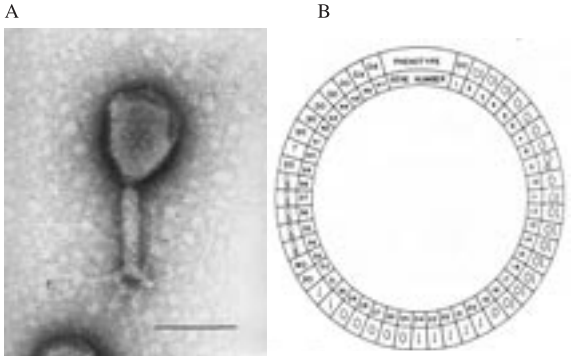

Figure 1. (*A*) Blectron micrograph of the bacteriophage T4. (*B*) Summary of the T4 genes in linkage map order with ideograms and abbreviations indicating mutant phenotypes of conditional-lethal alleles. (*A*, Reprinted, with permission, from Büchen-Osmond 2003.)

proteins are the basis for most of what is known about the roles they play in the biology of all organisms, including the human. This "grand unification" of biology is part of the genomic perspective.

EXTRACTING BIOLOGICAL INFORMATION FROM GENOME SEQUENCE

Even before the Human Genome Project had been organized, the need for suitable archives for the onrushing flood of genomic data became obvious, as did the need for ways to compare and display data in a manner useful to biologists. Computation quickly became indispensable; thanks to the rapid pace of advance in the productivity of computers, computation per se has rarely if ever been limiting in genomics. It was the provision of suitable biological context for sequences and computed results about sequences that became the challenge. At first, most effort went into "primary annotation," which includes finding the open reading frames, splice junctions, homology and synteny with other organisms, etc. Most of this annotation today is being done essentially automatically by an increasingly sophisticated and powerful set of computer programs.

It soon became clear that if biologists were to have useful access to the fruits of genomic sequencing, another level of "biological" annotation would be necessary. Databases were organized to meet this larger challenge, ranging from the very basic and important archival sequence databases (NCBI, EBI, SwissProt, etc.) to the more specialized organism-specific databases (SGD, MGD, FlyBase, WormBase, etc.). Today, these have grown into a veritable armamentarium, including many more focused databases that catalog such things as sequence motifs or mutations in particular gene families. These databases have already become indispensable to working biologists of every kind.

A considerable part of the challenge facing biological annotators concerns nomenclature and language. The classical methods for naming and describing the functions of genes, proteins, protein assemblies, and even biological processes themselves remain different for each species and for each sub-field of biology, producing something of a Tower of Babel. The genomic database organizations recognized the need for a common language describing the biology associated with genes and proteins, and banded together to produce what is now called the "Gene Ontology" (GO; this is not, in a strict sense, an ontology, but the name has caught on nevertheless; Ashburner et al. 2000; Harris et al. 2004). GO, which is described elsewhere in this volume (Ashburner et al.), emerged as a limited vocabulary organized in a set of directed acyclic graphs that represent the "biological processes," "molecular activities," and "subcellular locations" associated with genes and proteins of an organism. GO has rapidly become popular with genome biologists, as it facilitates biological annotation in a way that allows, among other things, computational connections among the functional annotations of orthologs and makes it possible to begin to assess quantitatively the significance, in the context of biological function, of the coexpression of two genes (see, e.g., Raychaudhuri et al. 2003; Troyanskaya et al. 2003).

ASSESSING GENE EXPRESSION GENOME-WIDE

Complete genomic sequences have provided biologists with a finite universe of genes and proteins for each organism. For the first time, it has become possible to design experiments that interrogate every gene for its activity in a biological process. It is this kind of comprehensive experiment that provides a global perspective and that we think of as "genomic." The genomic technology that has advanced the most rapidly in recent years is DNA microarray hybridization. Many variants of this technology have come into use. All have in common the intent to measure, by hybridization, the relative amounts of nucleic acid in a sample corresponding to each gene. As with any method, there are limitations in practice, some of which apply to all the technologies, and others that affect some methods more than others; we do not discuss further the technology per se; instead, the reader is directed to a collection of recent reviews (Brown and Botstein 1999; *Nature Genetics* [supplement] 2002). Despite these limitations, DNA microarray technology has provided a wide-ranging and comprehensive view of gene expression patterns both in experimental model systems and in normal and diseased human tissues. DNA microarrays have also been used to study, genome-wide, changes in DNA copy number, once again in both model systems and human tissues.

As with DNA sequences themselves, the value of DNA microarray analysis depends on the ability to connect results with biology. The large numbers of measurements represented in a single array (typically tens of thousands) require considerable computation not only to recover and organize the data, but also to present them in a form that is simultaneously comprehensive and intuitive. In 1998, Eisen et al. described a system for analysis and display of microarray data, many features of which have come into common use. The most important and general feature is

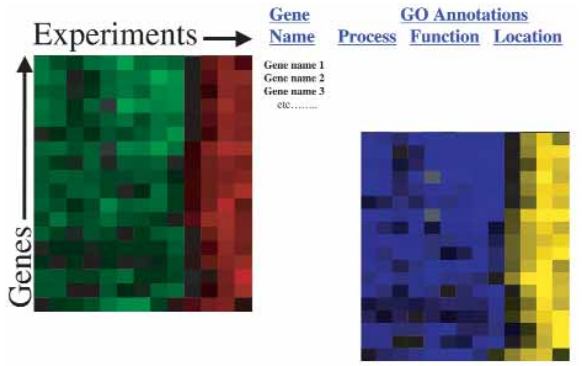

Figure 2. Display of relative degrees of gene expression in a set of DNA microarray experiments. A table of $\log_2$ of ratios of gene expression between an experimental sample and (usually) a common reference is colored according to the relationship of each cell in a row to the median (or mean) for that row. Increasing intensity of red indicates higher ratios, and increasing intensity of green indicates lower ratios; often yellow and blue are substituted for red and green, respectively (*inset*). The display is connected to biology by the text annotations of gene name and abbreviations of appropriate GO terms. For more detail, see Eisen et al. (1998) and Ashburner et al. (2000).

the method of display (adapted from Weinstein et al. 1997): Tables of suitably analyzed gene expression values are presented with cells colored according to the magnitude of the difference between the value in the cell and the mean or median for that gene in the group of arrays being compared. Generally, each row of the table represents a single gene, and each column a single array. Much useful analysis can be performed, without removal of any data, just by manipulating the order of the rows and columns according to an analysis scheme (usually some form of clustering), after which patterns of gene expression become manifest as patches of color. Viewing the entire colored table, one can see an overview of patterns consisting of literally millions of individual gene expression values. One can also zoom in on portions of the pattern. As shown in Figure 2, the gene names are listed next to each row along with summary descriptions of what is known about the genes (e.g., GO annotations). At this level, biologists can often not only see relationships among the genes in their experiments, but also begin to make inferences based on what is held in common by the annotations for the genes clustered together by the analysis. A Windows implementation (TreeView) and an enhanced platform-independent version (JavaTreeView) of this display system are freely available from genome-www.stanford.edu.

Several points are worth emphasizing about the practical advantages of this style of analysis and display. First, and probably most important, the analyis preserves the comprehensive nature of experiments intended to interrogate the entire genome. In this way, it provides and maintains a genomic perspective. The experimenter gets an overview, through the patterns of color, of all the data as analysis proceeds. Second, because data are not removed, it facilitates the unsupervised discovery of relationships of the patterns of gene expression between uncharacterized genes and those that have been well-studied. Inclusion of an uncharacterized gene in a cluster of coexpressed genes has become one of the most common leads to characterization of such a gene's role in the cell. Third, it allows the comparison (and, under the right circumstances, even the amalgamation) of data from many different kinds of experiments. For example, the Eisen et al. analysis and display system facilitated the discovery that the so-called "proliferation cluster" observed in a number of studies of tumors consists of genes periodically expressed in synchronized HeLa cells (Whitfield et al. 2002). The ability to usefully compare diverse data, collected by different groups under different conditions, is an important property that microarray data share with molecular sequence data. The data have cumulative value, which makes it important that all data, not just the subsets used to make a point in a paper, should be made freely available at the time of publication.

MOLECULAR PORTRAITS OF CELLS, TISSUES, AND TUMORS

DNA microarrays that contain many thousands of different human cDNA sequences can be used to assess patterns of gene expression, producing a highly detailed and nuanced map of gene expression across the genome. Each individual microarray shows the relative abundance of transcripts of each of the genes represented on the array, and thereby gives a characteristic and nuanced picture of the biological state of the cells or tissues from which the mRNA was extracted. After application of clustering algorithms, the patterns of a number of microarrays can be assessed together, not only visually, but also quantitatively, using a variety of statistical methods and computer algorithms that relate gene expression patterns to each other and to external information, including the identities of the cells or tissues, their environment, their response to previously applied stimuli, or disease state. An example of such a map is shown in Figure 3, in which the patterns of gene expression of about 6000 different human genes in 440 different cell lines and tissues are shown together, after the data had been clustered in both the gene and array dimension. It is easy to discern visually that similar cell types and tissues, collected under similar conditions, display similar patterns of gene expression. Likewise, despite the extreme diversity of cell and tissue types and environmental conditions, it is easy to discern groups of genes that appear regularly to be expressed similarly over the entire gamut of cell type and condition.

What can be learned from clustering of gene expression patterns of large numbers of cell and tissue samples? First, as pointed out above, one can obtain, for relatively uncharacterized genes, quite specific suggestions regarding their role in the biology of the tissue or organism. Second, clustering of arrays according to the patterns of gene expression allows inferences to be made about the biology of the cells from which the RNA was drawn. A good example of this was the demonstration of substantial and reproducible biological differences among apparently similar cell types (e.g., fibroblasts or endothelial cells), depending on their anatomical site of origin (Chang et al. 2002; Chi

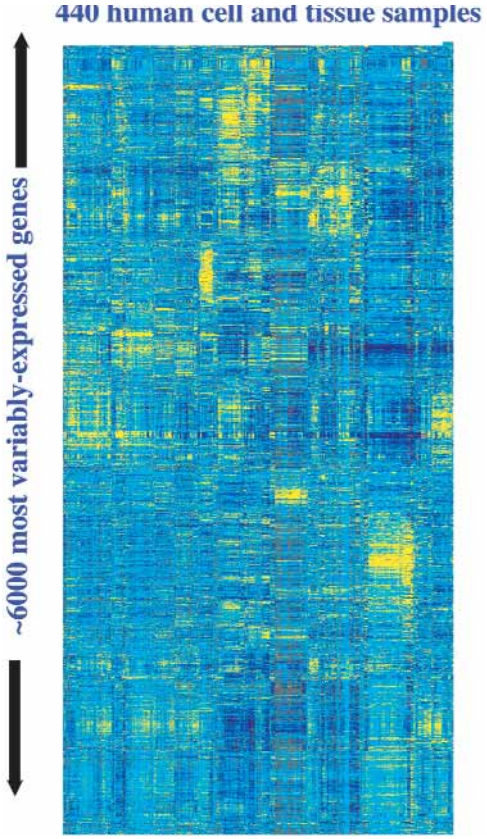

Figure 3. Cluster diagram as described in Fig. 2 that includes gene expression data from more than 400 cell and tissue samples whose gene expression was measured relative to a common reference; about 6000 of the most variably expressed genes are represented. (This figure was made by Pat Brown and Mike Eisen and includes data collected by Max Diehn, Xin Chen, Jon Pollack, Chuck Perou, Therese Sorlie, Mitch Garber, Marci Schaner, Matt van de Rijn, Gavin Sherlock, and Mike Fero.)

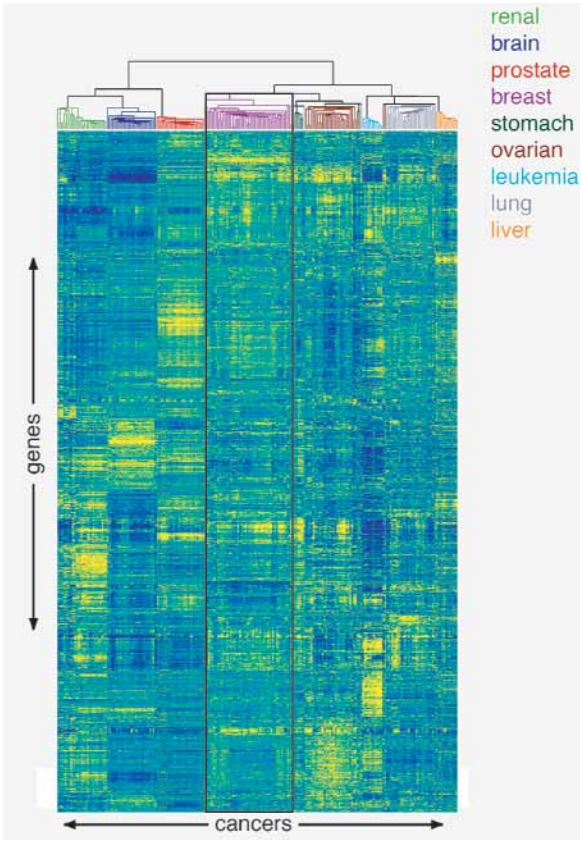

Figure 4. Cluster diagram as described in Fig. 2 that includes samples of about 500 diverse tumors relative to a common reference; about 6000 of the most variably expressed genes are represented. Source of the data is the same as in Fig. 3.

et al. 2003). This conclusion emerged when gene expression patterns of fibroblast or endothelial cell cultures from several individuals were compared by cluster analysis. The patterns for cells derived from similar anatomical sites but from different individuals clustered tightly together, indicating very little interindividual variation compared to the variation found between similar cell cultures derived from different parts of the human body.

Similarly strong biological inferences could be drawn from the analysis of gene expression profiles of human tumors. A large set of arrays representing the patterns of expression of about 6000 genes in a variety of diverse, crudely dissected tumors is shown in Figure 4. Once again, the clustering was done in both the gene and array dimension. Inspection of the figure shows that tumors of similar tissue of origin, but from many different patients, cluster together, indicating that tumors of each type (e.g., breast) are more similar in pattern of gene expression to each other than any of them is to another tumor type (e.g., ovarian or liver). This is despite the fact that these tumors consist of many different cell types, each of which contributes characteristic patterns of gene expression to the overall portrait of the tumor (for a fuller discussion of this point, see Perou et al. 1999, 2000).

Gene expression patterns thus appear to reflect accurately, as might have been expected, the biological differences among cell types and tumor tissues. Considering the many thousands of genes whose expression varies among the various cell and tissue tumor types, and the relatively small variation in interindividual gene expression for each different cell type and tissue, these molecular portraits may represent the best and most nuanced distinctions that can today be made among human cells and tissues. Molecular profiles thus provide a genomic perspective, faithfully representing, in considerable detail, the genomic contribution to cell and tissue phenotype, identity, and developmental history.

TUMOR SUBTYPES BASED ON MOLECULAR PORTRAITS

Gene expression patterns also appear to reflect the genomic contribution to the development of tumors, consistent with everything that is known about the genetic events that underlie tumor initiation, progression, and metastasis. It therefore seemed particularly significant that among the portraits of tumors of similar origin and pathological diagnosis (e.g., breast tumors), there appeared, based on the clustering patterns, clear indications of distinguishable subtypes.

Figure 5 (Sørlie et al. 2003) shows the molecular portraits of breast tumors derived from 115 different patients. From the dendrogram (Fig. 5B), one can see that breast tumors are very diverse, especially when compared with the patterns of three typical normal breast samples (shown in black). Similarly wide diversity in tumor profiles and relatively minimal variation in normal tissue profiles have been found not only for breast cancers (Perou et al. 2000; Sørlie et al. 2001, 2003), but also for lung (Garber et al. 2001), liver (Chen et al. 2002), and gastric (Leung et al. 2002) cancers. Another common feature (not shown) is that tumor samples from the same breast cancer patient, either by repeated surgical sampling or from lymph node metastases, tend to have profiles very similar to each other (Perou et al. 2000; Sørlie et al. 2001, 2003); similar results were obtained in our studies of lung and liver tumors (Garber et al. 2001; Chen et al. 2002). This property is useful for defining subsets of genes whose expression patterns contain the most information for distinguishing subtypes, as was done for Figure 5.

The simplest interpretation of the dendrogram in Figure 5 is to suppose that there are five subtypes corresponding to the top-level nodes of the dendrogram. The samples whose patterns are best correlated in each sub-

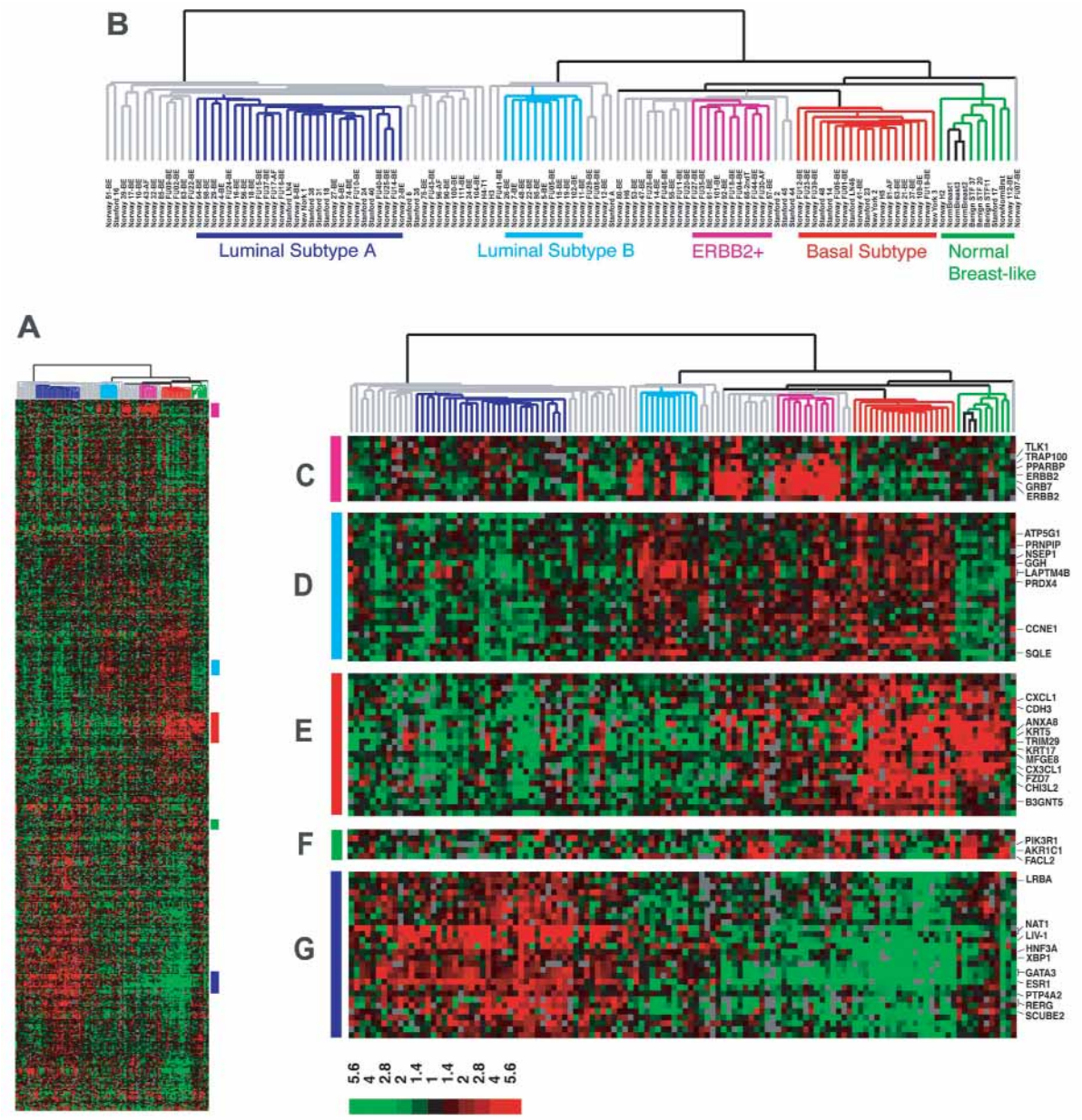

Figure 5. Cluster analysis of 115 breast tumors. (*A*) Representation of the entire data set clustered according to the expression of the 534 genes that vary least in repeated samples from the same individual and most across all the samples. (*B*) Dendrogram showing the clustering of the tumor samples into five groups, color coded as indicated. Black indicates the three normal breast samples. *C,D,E,F*, and *G* show the clusters of genes whose expression is characteristic of the ERBB2+, luminal B, basal, normal-like, and luminal A subtypes, respectively. Scale bar shows the fold difference relative to the median for each gene. (Reprinted, with permission, from Sørlie et al. 2003.)

group are color-coded. Each subtype has been named according to previous practice (Perou et al. 2000, Sørlie et al. 2001). The "luminal" tumor subtypes express genes (e.g., cytokeratins 5 and 17) normally expressed in the epithelial cells that normally line the lumen of breast, whereas the "basal" tumor subtypes express genes (e.g., cytokeratins 8 and 18) normally expressed by the basal epithelial cells that normally are located one or more cell diameters away from the lumen (Perou et al. 2000; van de Rijn et al. 2002). This suggests that the origins of luminal and basal tumors are somehow related to the differences between the development of normal basal and luminal epithelial cell types in the breast.

Three additional lines of evidence support the biological significance of at least some of the distinctions among the subtypes. First, the several subtypes are associated with different disease severity. Second, some of the subtype distinctions, and their different clinical consequences, are reproducible in completely separate cohorts of patients (Sørlie et al. 2003). Third, the tumors of patients genetically predisposed to breast cancer appear always to be of the basal subtype, suggesting this subtype is biologically distinct from the others.

Figure 6 shows disease outcomes for women with different breast tumor subtypes. Data for two separate patient cohorts, differing in age at onset and methods of treatment, are shown. In both cohorts the relationship of subtype to clinical course is similar. The most prominent features of the Kaplan-Meier curves are that women with tumors of the luminal A subtype have markedly less severe outcomes than do those with the basal subtype. In both cohorts, women with luminal B subtype tumors appear to have disease with intermediate severity. These results are in considerable agreement with previous studies relating expression of particular individual genes or proteins (e.g., estrogen receptor or Her2/neu) to disease outcome (cf. Henson et al. 1995; Allred et al. 1998).

Our ability to discern a relatively small number of biologically coherent breast tumor subtypes suggests a systematic explanation of such results: The different subtypes have many correlated differences in gene expression, and the differences in outcome are related to the difference in subtype, and not generally the expression of individual genes or the presence or absence of particular proteins. The success in correlating, on a large scale (more than 600 patients), the presence of cytokeratin 17 (by immunohistochemistry) with outcome underscores this point (van de Rijn et al. 2002).

As indicated above, many different tumor types have been studied using genome-wide gene expression profiling. In our experience, subtypes are more often discernible in such studies than not: We have found evidence for subtypes in diffuse large-cell lymphomas (Alizadeh et al. 2000); lung (Garber et al. (2001), liver (Chen et al. 2002), gastric (Chen et al. 2003), soft tissue (Nielsen et al. 2002), ovarian (Schaner et al. 2003), and follicular lymphoma (Bohen et al. 2003). Interestingly, in the case of follicular lymphoma, the subtypes appeared to have dichotomous responses to treatment with rituximab, providing a somewhat different line of evidence for the biological and clinical significance of the distinction between the subtypes of follicular lymphoma.

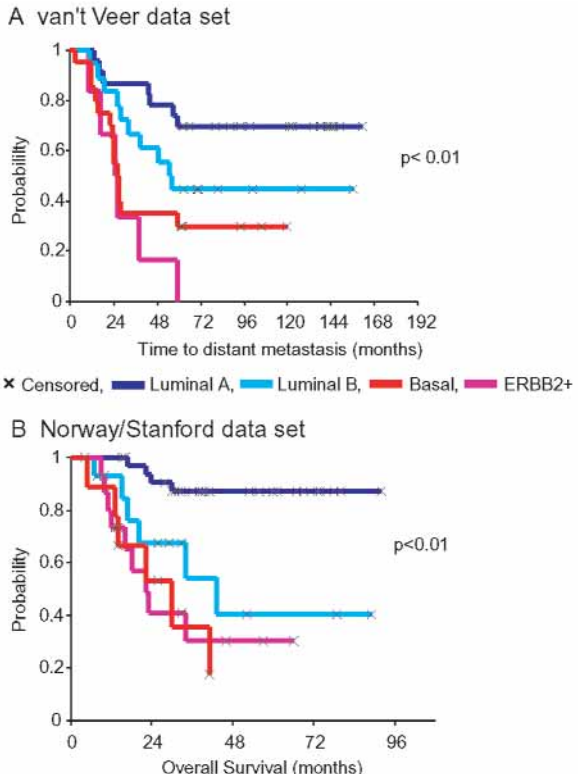

Figure 6. Kaplan-Meier analysis of disease outcome in two patient cohorts (from Sørlie et al. 2003). Color codes are the same as in Fig. 5. (*A*) Data from van't Veer et al. (2002) showing time to metastasis for 97 sporadic cases. (*B*) Data from Sørlie et al. (2003) showing overall survival for 72 patients with locally advanced breast cancer. The normal-like class was omitted from this analysis.

GENOMICS AND BIOLOGICAL PERSPECTIVE

The availability of genomic sequences has changed the way in which we think about biology, even as it has changed the way in which we do research. Where once we were limited to studies applicable directly only to a limited set of organisms, we now can make inferences that apply, with high likelihood, to most organisms; where once we were limited to a view, in an experiment, of only a few genes and/or gene products, we are beginning to be able to see our experimental results in the context of all the genes and gene products.

The genome sequences have resulted in a "grand unification" of biology based on molecular sequence conservation. Molecular sequence comparisons have all but ended the intellectual fragmentation along taxonomic lines that has been a feature of the biological sciences for centuries. It is now routine to make detailed studies of common ancestry at the whole-genome level, at the level of individual gene and protein sequences, and even at the level of oligonucleotide-length sequence motifs. Results

from such studies have facilitated and stimulated the development of limited vocabularies and a common language about biological functions and relationships (e.g., GO) that allows information in experimentally tractable systems to be used effectively in understanding, and devising experimental tests of that understanding, under less tractable circumstances, including humans and human disease.

The genome sequences themselves have made possible comprehensive, genome-wide experimentation where previously only a few genes and proteins could be studied simultaneously. The most advanced of these technologies are the genome-wide gene expression techniques, but others, such as the production of comprehensive sets of deletion (or "knockout") mutations in all genes (Winzeler et al. 1999; Giaever et al. 2002), comprehensive sets of fluorescently labeled proteins in vivo (Ghaemmaghami et al. 2003), comprehensive two-hybrid protein interaction screens (Uetz et al. 2000), and genome-wide synthetic lethality tests (Tong et al. 2001) are coming into use. Such methods are qualitatively different because they provide relatively complete information in context. This context is causing a new appreciation of the global consequences of phenomena where only the behavior of a few genes had been examined before. Taking a few very basic examples just from our own experience, this approach expanded severalfold the number of known cell-cycle-regulated genes in yeast (Spellman et al. 1998) and animal cells (Whitfield et al. 2002), as well as the stress response genes in yeast (Gasch et al. 2000, 2001).

The context provided by genomics has stimulated great interest in understanding globally interactions among genes and proteins. It is, for example, routine in gene expression studies to find genes and proteins that interact or participate in a process or pathway simply because they show characteristic coexpression with the other genes and proteins involved under many different conditions. New fields (called integrative genomics or system biology) are emerging whose explicit goal is to understand biological function and regulation in context, capitalizing on the new perspective and technology provided by the genome sequences.

The study of molecular portraits of tumors provides a good illustration of the change in perspective provided by the genomic view. It was difficult to distinguish what turn out to be reproducible and robust subtypes of tumors on the basis of expression of one or a few genes or proteins. Only when it became possible to study in parallel the expression of thousands of genes was it possible to see these subtypes. Instead of thinking of each new molecular marker as a central actor in tumorigenesis, progression, or metastasis, one can now see that there may be hundreds of genes with the same expression patterns. Similarly, only by following many genes at once could one distinguish differences in apparently normal fibroblasts or endothelial cells based on their anatomical origin. It is the perspective provided by the still novel ability to study and appreciate biological phenomena in a global context that will characterize biological thinking and research for years to come.

ACKNOWLEDGMENTS

I am indebted, first of all, to P.O. Brown for a decade-long collaboration at Stanford that produced many of the ideas and results summarized above. The illustrations contain the work of virtually all of the members of our joint laboratory and our many collaborators. I am also indebted to Mike Cherry and the staff of the Saccharomyces Genome Database. Research was supported by grants from the National Cancer Institute, the National Human Genome Research Institute, and the National Institute of General Medical Sciences.

REFERENCES

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., Hudson J., Jr., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O., and Staudt L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403:** 503.

Allred D.C., Harvey J.M., Berardo M., and Clark G.M. 1998. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Mod. Pathol.* **11:** 155.

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., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., and Sherlock G. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25.

Bohen S.P., Troyanskaya O.G., Alter O., Warnke R., Botstein D., Brown P.O., and Levy R. 2003. Variation in gene expression patterns in follicular lymphoma and the response to rituximab. *Proc. Natl. Acad. Sci.* **100:** 1926.

Brenner S. 1974. The genetics of *Caenorhabditis elegans. Genetics* **77:** 71.

Brown P.O. and Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* (suppl. 1) **21:** 33.

Büchen-Osmond C., Ed. 2003. Myoviridae. In *ICTVdB—The Universal Virus Database*, version 3. ICTVdB Management, The Earth Institute, Biosphere 2 Center, Columbia University, Oracle, Arizona.

Chang H.Y., Chi J.T., Dudoit S., Bondre C., van de Rijn M., Botstein D., and Brown P.O. 2002. Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proc. Natl. Acad. Sci.* **99:** 12877.

Chen X., Cheung S.T., So S., Fan S.T., Barry C., Higgins J., Lai K.M., Ji J., Dudoit S., Ng I.O., Van De Rijn M., Botstein D., and Brown P.O. 2002. Gene expression patterns in human liver cancers. *Mol. Biol. Cell* **13:** 1929.

Chen X., Leung S.Y., Yuen S.T., Chu K.M., Ji J., Li R., Chan A.S., Law S., Troyanskaya O.G., Wong J., So S., Botstein D., and Brown P.O. 2003. Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell* **14:** 3208.

Chi J.T., Chang H.Y., Haraldsen G., Jahnsen F.L., Troyanskaya O.G., Chang D.S., Wang Z., Rockson S.G., van de Rijn M., Botstein D., and Brown P.O. 2003. Endothelial cell diversity revealed by global expression profiling. *Proc. Natl. Acad. Sci.* **100:** 10623.

Eisen M.B., Spellman P.T., Brown P.O., and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863.

Epstein R.H., Bolle A., Steinberg C.M., Kellenberger E., Boy de la Tour E., Chevalley R., Edgar R.S., Susman M., Denhardt G.H., and Lielausis A. 1964. Physiological studies of conditional lethal mutations of bacteriophage T4D. *Cold Spring Harbor Symp. Quant. Biol.* **28:** 375.

Garber M.E., Troyanskaya O.G., Schluens K., Petersen S., Thaesler Z., Pacyna-Gengelbach M., van de Rijn M., Rosen G.D., Perou C.M., Whyte R.I., Altman R.B., Brown P.O., Botstein D., and Petersen I. 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci.* **98:** 13784.

Gasch A.P., Huang M., Metzner S., Botstein D., Elledge S.J., and Brown P.O. 2001. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell* **12:** 2987.

Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., and Brown P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11:** 4241.

Ghaemmaghami S., Huh W.K., Bower K., Howson R.W., Belle A., Dephoure N., O'Shea E.K., and Weissman J.S. 2003. Global analysis of protein expression in yeast. *Nature* **425:** 737.

Giaever G., Chu A.M., Ni L., Connelly C., Riles L., Veronneau S., Dow S., Lucau-Danila A., Anderson K., Andre B., Arkin A.P., Astromoff A., El-Bakkoury M., Bangham R., Benito R., Brachat S., Campanaro S., Curtiss M., Davis K., Deutschbauer A., Entian K.D., Flaherty P., Foury F., Garfinkel D.J., and Gerstein M., et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387.

Harris M.A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., Eilbeck K., Lewis S., Marshall B., Mungall C., Richter J., Rubin G.M., Blake J.A., Bult C., Dolan M., Drabkin H., Eppig J.T., Hill D.P., Ni L., Ringwald M., Balakrishnan R., Cherry J.M., Christie K.R., Costanzo M.C., and Dwight S.S., et al. (Gene Ontology Consortium). 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258.

Hartwell L.H. 1970. Biochemical genetics of yeast. *Annu. Rev. Genet.* **4:** 373.

———. 1974. *Saccharomyces cerevisiae* cell cycle. *Bacteriol. Rev.* **38:** 164.

———. 1978. Cell division from a genetic perspective. *J. Cell Biol.* **77:** 627.

Henson D.E., Fielding L.P., Grignon D.J., Page D.L., Hammond M.E., Nash G., Pettigrew N.M., Gorstein F., and Hutter R.V. 1995. College of American Pathologists Conference XXVI on clinical relevance of prognostic markers in solid tumors. Summary. Members of the Cancer Committee. *Arch. Pathol. Lab. Med.* **119:** 1109.

Leung S.Y., Chen X., Chu K.M., Yuen S.T. Mathy J., Ji J., Chan A.S., Li R., Law S., Troyanskaya O.G., Tu I.P., Wong J., So S., Botstein D., and Brown P.O. 2002. Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. *Proc. Natl. Acad. Sci.* **99:** 16203.

Nature Genetics. 2002. Supplement, volume 32, pp. 461–552. Nature Publishing Group, Nature America, New York.

Nielsen T.O., West R.B., Linn S.C., Alter O., Knowling M.A., O'Connell J.X., Zhu S., Fero M., Sherlock G., Pollack J.R., Brown P.O., Botstein D., and van de Rijn M. 2002. Molecular characterisation of soft tissue tumours: A gene expression study. *Lancet* **359:** 1301.

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. *Proc. Natl. Acad. Sci.* **96:** 9212.

Perou C.M., Sørlie T., Eisen M.B., van de Rijn M., Jeffrey S.S., Rees C.A., Pollack J.R., Ross D.T., Johnsen H., Akslen L.A., Fluge O., Pergamenschikov A., Williams C., Zhu S.X., Lonning P.E., Børresen-Dale A.L., Brown P.O., and Botstein D. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747.

Raychaudhuri S., Chang J.T., Imam F., and Altman R.B. 2003. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.* **31:** 4553.

Schaner M.E., Ross D.T., Ciaravino G., Sørlie T., Troyanskaya O., Diehn M., Wang Y.C., Duran G.E., Sikic T.L., Caldeira S., Skomedal H., Tu I.P., Hernandez-Boussard T., Johnson S.W., O'Dwyer P.J., Fero M.J., Kristensen G.B., Børresen-Dale A.L., Hastie T., Tibshirani R., van de Rijn M., Teng N.N., Longacre T.A., Botstein D., Brown P.O., and Sikic B.I. 2003. Gene expression patterns in ovarian carcinomas. *Mol. Biol. Cell* **14:** 4376.

Sørlie T., Tibshirani R., Parker J., Hastie T., Marron J.S., Nobel A., Deng S., Johnsen H., Pesich R., Geisler S., Demeter J., Perou C.M., Lønning P.E., Brown P.O., Børresen-Dale A.L., and Botstein D. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* **100:** 8418.

Sørlie T., Perou C.M., Tibshirani R., Aas T., Geisler S., Johnsen H., Hastie T., Eisen M.B., van de Rijn M., Jeffrey S.S., Thorsen T., Quist H., Matese J.C., Brown P.O., Botstein D., Eystein-Lønning P., and Børresen-Dale A.L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98:** 10869.

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. *Mol. Biol. Cell* **9:** 3273.

Tong A.H., Evangelista M., Parsons A.B., Xu H., Bader G.D., Page N., Robinson M., Raghibizadeh S., Hogue C.W., Bussey H., Andrews B., Tyers M., and Boone C. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364.

Troyanskaya O.G., Dolinski K., Owen A.B., Altman R.B., and Botstein D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci.* **100:** 8348.

Uetz P., Giot L., Cagney G., Mansfield T.A., Judson R.S., Knight J.R., Lockshon D., Narayan V., Srinivasan M., Pochart P., Qureshi-Emili A., Li Y., Godwin B., Conover D., Kalbfleisch T., Vijayadamodar G., Yang M., Johnston M., Fields S., and Rothberg J.M. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403:** 623.

van de Rijn M., Perou C.M., Tibshirani R., Haas P., Kallioniemi O., Kononen J., Torhorst J., Sauter G., Zuber M., Kochli O.R., Mross F., Dieterich H., Seitz R., Ross D., Botstein D., and Brown P. 2002. Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am. J. Pathol.* **161:** 1991.

van'T Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R., and Friend S.H. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530.

Watson J.D. and Crick F.H.C. 1953a. A structure for deoxyribose nucleic acid. *Nature* **171:** 737.

———. 1953b. Genetic implications of the structure of deoxyribonucleic acid. *Nature* **171:** 964.

———. 1953c. The structure of DNA. *Cold Spring Harbor Symp. Quant. Biol.* **18:** 123.

Weinstein J.N., Myers T.G., O'Connor P.M., Friend S.H., Fornace A.J., Jr., Kohn K.W., Fojo T., Bates S.E., Rubinstein L.V., Anderson N.L., Buolamwini J.K., van Osdol W.W., Monks A.P., Scudiero D.A., Sausville E.A., Zaharevitz D.W., Bunow B., Viswanadhan V.N., Johnson G.S., Wittes R.E., and Paull K.D. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275:** 343.

Whitfield M.L., Sherlock G., Saldanha A.J., Murray J.I., Ball C.A., Alexander K.E., Matese J.C., Perou C.M., Hurt M.M., Brown P.O., and Botstein D. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13:** 1977.

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.