Annu. Rev. Immunol. 2000. 18:829–859

GENOMIC VIEWS OF THE IMMUNE SYSTEM*

Louis M. Staudt[1] and Patrick O. Brown[2]
[1]*Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland; e-mail: lstaudt@box-l.nih.gov*
[2]*Dept. of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine; e-mail: pbrown@cmgm.stanford.edu*

Key Words genomics, microarray, gene expression

■ **Abstract** Genomic-scale experimentation aims to view biological processes as a whole, yet with molecular precision. Using massively parallel DNA microarray technology, the mRNA expression of tens of thousands of genes can be measured simultaneously. Mathematical distillation of this flood of gene expression data reveals a deep molecular and biological logic underlying gene expression programs during cellular differentiation and activation. Genes that encode components of the same multi-subunit protein complex are often coordinately regulated. Coordinate regulation is also observed among genes whose products function in a common differentiation program or in the same physiological response pathway. Recent application of gene expression profiling to the immune system has shown that lymphocyte differentiation and activation are accompanied by changes of hundreds of genes in parallel. The databases of gene expression emerging from these studies of normal immune responses will be used to interpret the pathological changes in gene expression that accompany autoimmunity, immune deficiencies, and cancers of immune cells.

INTRODUCTION

The established, model-driven field of immunology is about to collide with the upstart, discovery-driven field of genomics. Traditional research in molecular biology and molecular immunology can be likened to trying to understand a movie by successively examining a few pixels (genes) at a time from each frame. Genomic approaches allow the scientist to view the entire movie in one sitting and discover complex interrelationships among the plot, characters, and recurring themes. The tension between genomic approaches and the more traditional single gene orientation of molecular biology often leads to criticism of genomic approaches as non-hypothesis-driven. Those who favor a genomic approach embrace this characterization, noting that genomic approaches are deliberately not hypothesis-limited and are instead discovery-driven. When the powerful

molecular tools of genomics are applied to a new biological question, discoveries will almost certainly be made that will generate new hypotheses and necessitate a reworking of existing models.

The field of immunology is especially primed to receive the new insights that genomics can provide. Numerous immune cell types have been defined with high precision, and methods to culture and manipulate these cells are well developed. Such experimental systems are ideal settings in which to study genome-wide phenomena under very well controlled circumstances. Powerful techniques for the analysis of single gene mutations in lymphocytes have been developed in the mouse, yielding a plethora of precise genetic models that are ideal substrates for genomic approaches. Finally, malfunctions of the immune system give rise to a host of autoimmune diseases, immune deficiency diseases, and malignancies in need of fresh insights that may be supplied by genomic views of the pathological processes.

The young field of genomics has already been somewhat arbitrarily subdivided into two separate disciplines. One branch of genomics, structural genomics, has as its immediate goal to determine complete genomic DNA sequences of the major model organisms. To date, the complete genomes of the yeast *Saccharomyces cerevisiae* (1), the worm *Caenorhabditis elegans* (2), and numerous prokaryotes have been sequenced (3). The complete genomes of these simple organisms have yielded a plethora of orthologues of human and mouse genes. New insights into the function of these evolutionarily conserved gene families are thus made possible using the more tractable genetics of these model organisms.

Much of this review focuses, however, on the newly coined field of functional genomics. Broadly construed, functional genomics encompasses any experimental approach that uses genomic structural information to view and understand biological processes in a systematic and comprehensive fashion. This vast frontier, opened up by the genome sequencing projects, is just beginning to be explored. Even at this early stage, a diversity of approaches have been developed for exploring the living genome. In this review, however, we focus primarily on one of them: the genome-wide analysis of mRNA expression using DNA microarrays. Because of the central role played by regulation of mRNA levels in development and physiology and because of the deep, logical connection between the function of a gene's product and its pattern of expression, this specific area of functional genomics research has been the richest source of new biological insights. One of the defining characteristics of functional genomic approaches is that they generate data streams that overwhelm the traditional analytical methods of biology and indeed make possible entirely new ways of looking at living systems. Throughout this review, we discuss how the field of bioinformatics has faced the challenge of organizing, distilling, and visualizing the information provided by genomic data in ways that allow biological insights to be found.

The field of genomics naturally intersects with classical genetics in the study of complex genetic diseases. In polygenic disorders, the contribution of any one

locus to the disease phenotype is small and may be apparent only in the context of specific alleles in other genes. The current race to define allelic variants of genes in human populations is largely fueled by the desire to understand their contribution to differential disease susceptibility. Millions of single nucleotide polymorphisms exist in the human population, and recognizing the linkage or association of a single polymorphism with a disease state is a considerable challenge (4). Techniques in functional genomics provide information that can complement linkage and association methods in making the connection between genes and disease risks. For virtually every gene, variation in its expression, as a function of cell specialization, physiology, or disease, is much richer than allelic variation in that gene. Because the pattern in which each gene is expressed is so closely connected to the biological role and effects of its product, systematic studies of variation in gene expression can provide an alternative approach to linking specific genes with specific diseases and to recognizing heritable variation in genes important for immune function. For example, allelic differences in the regulatory regions of cytokine genes may influence the expression levels of cytokines during particular immune responses. An appreciation for such quantitative traits in the immune system may help unravel the genetics of autoimmune diseases and lymphoproliferative disorders.

STRUCTURAL GENOMICS AND THE IMMUNE SYSTEM

Systematic studies of genomic expression programs are best pursued in two independent steps. The first step is to obtain as complete a catalog as possible of all the expressed genes in the genome. The second step is to use parallel methods, such as DNA microarray hybridization, to measure the expression of each gene in the genome over the range of conditions and cell types under investigation. Our still incomplete knowledge of the human and mouse genomic sequence and the incomplete catalog of genes in these genomes present an important challenge in functional exploration of mammalian genomes. Even when a full mammalian genomic sequence is known, it will not immediately be possible to identify all of the segments that are expressed as mRNA. Computer algorithms such as GRAIL (5) use machine learning techniques to identify putative coding regions in genomic sequences. In practice, however, these algorithms need to be supplemented by cDNA sequence data to completely annotate the exon-intron structure of a mammalian genome. Indeed, even in microbial genomes with few or no introns and much higher densities of transcribed and protein-coding sequences than are found in mammalian genomes, current algorithms for identifying genes in genomic sequences have significant false positive and false negative rates. Therefore, an indispensable component of any mammalian genome project is high-throughput, single-pass sequencing of cDNA libraries to generate expressed sequence tags (ESTs) (6), which provides a systematic set of unique labels for identifying the mRNAs that can be expressed from a genome. The current release

of the EST database dbEST [release 070999 (7)] contains 1,476,380 human ESTs and 658,511 mouse ESTs. These numbers are much larger than the numbers of distinct transcripts represented in each set because a very large fraction of the ESTs in each set is composed of multiple representations of mRNAs that are widely or highly expressed in the cells from which the source libraries were obtained. Indeed, despite these large numbers, it is clear that that not all of the human genes are represented by an EST in this public database. To illustrate this deficiency, consider the representation of interleukin sequences in the dbEST. ESTs representing about half of the known human interleukins can be found in this database, but no ESTs representing interleukins 2, 3, 5, 9, 11, 12 beta, 14, and 17 have yet been encountered. By contrast, of the 8963 known human genes with full-length cDNA sequences, 89% are represented by an EST in the dbEST database. This discrepancy reflects the bias in the dbEST database toward genes that are widely or highly expressed and the fact that very few of the EST sequences in the public domain have come from cDNA libraries made from activated cells of the immune system. Given this example from the immune system, one wonders how many inducible genes in other specialized or rare cell types have yet to be identified.

For the present, filling the gaps in our catalog of human expressed genes is a practical problem for which simple, though technically challenging, incremental solutions can often be found. Several years ago, the public EST database was strikingly deficient in sequences from B lymphocytes. This void was a serious impediment not only to the study of normal B cell development and physiology but also to the study of human lymphoid malignancies, the majority of which are derived from B cells. In order to fill this void, several libraries were created from normal and malignant human B cells and sequenced under the auspices of the Cancer Genome Anatomy Project (8, 9). As shown in Table 1, each of these cDNA libraries yielded a large number of novel ESTs, ranging from 12% to 22% of the total ESTs sequenced, most presumably representing genes never previously identified or studied. In part, this apparent high rate of gene discovery can be attributed to the paucity of previous EST sequences from B cell libraries and to the normalization process used in creating the NCI_CGAP_GCB1 library (10). Among the non-unique ESTs, some represented genes that were observed only in B cell libraries or only in other lymphoid libraries (Table 1). This example illustrates the challenge that will be faced in trying to discover the complete set of expressed human genes, including all the genes expressed at low levels or in highly specialized cells or conditions.

The Unigene project at the National Center for Biotechnology Information has attempted to provide a systematic classification of EST sequences (11). Unigene uses sequence alignment methods to group overlapping cDNA sequences into clusters, each of which provisionally corresponds to a unique gene. The Unigene analysis of the B cell library ESTs also reveals a high rate of gene discovery: 1652 of the 83,240 Unigene clusters at the time of this writing are defined only by ESTs derived from B cells. Viewed in another way, approximately 10% of the

TABLE 1 High-throughout sequencing of human B cell cDNA libraries

		BLAST Analysis				Unigene Analysis	
CDNA Library Name	**mRNA Source**	**3′ ESTs Total**	**3′ ESTs Unique to Library**	**3′ ESTs Only in B Cell Libraries**	**3′ ESTs Only in Lymphoid Libraries**	**Unigene Clusters Containing Library Clone**	**Unigene Clusters Uniquely Defined by Library Clone**
NCI_CGAP_GCB1	Tonsillar germinal center/memory B cells	40428	7388	233	443	13078	1058
NCI_CGAP_GCB0	Tonsillar germinal center/memory B cells	907	139	1	4	495	4
NCI_CGAP_Lym12	Follicular mixed small and large cell	4038	480	21	18	2381	200
NCI_CGAP_Lym5	Follicular lymphoma	1293	182	17	10	859	40
NCI_CGAP_Lym6	Mantle cell lymphoma	621	135	3	3	316	16
NCI_CGAP_CLL1	Chronic lymphocytic leukemia	8628	1085	127	54	4612	277
All B cell libraries		55915	9409	402	532	15992	1652

genes (Unigene clusters) that were sampled during the sequencing of B cell libraries were B cell–restricted. This result dramatically demonstrates our relative ignorance of the molecular biology of B lymphocytes and the need for systematic, genomic approaches to determine the expression patterns and functions of these novel genes in immune responses and other physiological processes.

GENOMIC-SCALE ANALYSIS OF GENE EXPRESSION

Although posttranscriptional mechanisms are important in regulating the expression of many genes, most cellular regulation is achieved by changes in mRNA levels. Consequently, systematic studies of gene expression patterns have proven to be remarkably powerful sources of insight into gene function and biological processes. Four aspects of genome-wide gene expression analysis are particularly appealing. First is its feasibility: DNA microarrays make it easy to measure, in a single hybridization, the mRNA abundance of every gene for which either a clone or sufficient DNA sequence information exists. Second, there is a biologically rational connection between the function of a gene product and its expression pattern. Natural selection has acted to optimize simultaneously the functional properties of the product encoded by a gene and the program that dictates where, when, and in what amounts the product is made. As a rule, each gene is expressed in the specific cells and under the specific conditions in which its product makes a contribution to fitness. The richness and precision with which mRNA levels can be controlled is such that virtually every gene in the yeast genome can be distinguished from every other gene based on its pattern of expression. Therefore, even subtle variations in the expression patterns of genes can be related to corresponding differences in the functions of the products they encode. Third, promoters and the regulatory systems that act upon them function as transducers, integrating diverse kinds of information about the identity, environment, and internal state of a cell. Thus, a diversity of information that is difficult or impossible to measure is transformed into a signal that can readily be measured systematically using DNA microarrays. Learning to decode this transduced information is one of the immediate priorities of functional genomics. Finally, the set of genes expressed in a cell determines how the cell is built, what biochemical and regulatory systems are operative, and what it can and cannot do. Thus, as we learn to infer the biological consequences of gene expression patterns, using our growing knowledge of the functions of individual genes, we can use microarrays as microscopes to see a comprehensive, dynamic molecular picture of the living cell.

Several methods have been developed over the last several years to quantitate the mRNA expression of thousands of genes in parallel. One method, termed serial analysis of gene expression (SAGE), relies on high-throughput sequencing of 14-bp, gene-specific cDNA tags to enumerate the expression of individual genes in a cell (12). Because of its reliance on DNA sequencing, SAGE can identify novel transcripts that have not been observed in other high-throughput

sequencing projects. On the other hand, it is difficult to analyze large numbers of samples, or to measure changes in the abundance of rare transcripts, using SAGE, and thus this method is most suited to binary questions in which the transcriptional response to a particular cellular stimulus or to a single transcription factor is assessed. Within the immune system, SAGE has recently been used to analyze gene expression in mast cells before and after stimulation through the high-affinity IgE receptor (13). An interesting and unanticipated finding was the expression in resting mast cells of macrophage inhibition factor (MIF), a cytokine that was previously known to be constituitively expressed only in macrophages and anterior pituitary cells. MIF is an important mediator of delayed-type hypersensitivity (DTH) reactions, and this observation suggests an important role for mast cells in some forms of DTH. Despite extensive prior study of cytokine production by mast cells, the expression of MIF had not been reported, pointing to the value of unbiased, genome-wide gene expression surveys.

In the other common methods of genomic expression analysis, DNA fragments derived from individual genes are placed in an ordered array on a solid support. These arrays are then hybridized with radioactive or fluorescent cDNA probes prepared from total cellular mRNA by reverse transcription. Following washing, the hybridization of the cDNA probes to each array element is quantitated using either a phosphorimager for radioactive probes or a scanning confocal microscope for fluorescent probes. Three styles of arrays are used most commonly. Nitrocellulose filter arrays are prepared by robotic spotting of purified DNA fragments or lysates of bacteria containing cDNA clones, and the filter arrays are hybridized with radioactive cDNA probes (14–17). Oligonucleotide arrays can be produced by in situ oligonucleotide synthesis in conjunction with photolithographic masking techniques and are hybridized with fluorescent cDNA probes (18–22).These two array formats are typically hybridized with a single cDNA probe at a time. In order to compare the mRNA expression profiles of two samples, therefore, two probes are generated and hybridized to separate arrays. The relative hybridization of the two probes to each array element is determined indirectly by mathematical normalization of the two data sets. A third type of microarray is fabricated by robotic spotting of PCR fragments from cDNA clones onto glass microscope slides (23–29). These cDNA microarrays are simultaneously hybridized with two fluorescent cDNA probes, each labeled with a different fluorescent dye (typically Cy3 or Cy5). In this format, therefore, the relative mRNA expression in two samples is directly compared for each gene on the microarray (Figure 1A, see color insert). For a given gene, the fluorescence ratio corresponds well with more conventional measures of relative gene expression including Northern blot hybridization and quantitative RT-PCR (23, 29, 30). Scanning and interpreting large bodies of relative gene expression data is a formidable task, which is greatly facilitated by algorithms designed to organize the results in ways that highlight systematic features and by visualization tools that represent the differential expression of each gene as varying intensities and hues of color (Figure 1B, see color insert) (31).

Mathematical Analysis of Gene Expression Data

The ability to produce large systematic sets of measurements of gene expression on a genomic scale using DNA microarrays is becoming commonplace. A single group, in a year, can print several thousand microarrays with a single microarraying robot and can produce tens of millions of individual measurements of gene expression. The mathematical analysis of the resulting data is a rapidly evolving science that is nevertheless based on a rich mathematics of pattern recognition developed in other contexts (32). Typical goals of these analyses are to identify groups of genes that are coregulated within a biological system, to recognize and interpret similarities between biological samples on the basis of similarities in gene expression patterns, and to recognize features of gene expression patterns that can be related to distinct biological processes or phenotypes. In other words, the biologist wishes to identify systematic features in the data that can be understood as a molecular picture of a biological system.

The expression pattern for each gene on an array across *n* experimental samples can be represented by a point in *n*-dimensional space, with each coordinate specified by an expression measurement in one of the *n* samples. In order to determine the proximity of points in this gene expression space (a measure of the similarity in the expression patterns of the corresponding genes), one must first define a metric that quantitates the distance between any two of these points. In the clustering algorithms that have been implemented thus far, the most commonly used metric is essentially the standard correlation coefficient of the two data vectors (31). Although there are other possible ways of measuring distance in gene expression space, this metric is well suited to gene expression data because it corresponds well to the intuitive idea of coordinate regulation of two genes (31).

The second step in the mathematical treatment of array data is to apply one of many clustering algorithms that use the distance metrics to find clusters of genes in this *n*-dimensional space, corresponding to genes with similar patterns of variation in expression over a series of experiments. The clustering methods that have been applied to array data thus far are hierarchical clustering (31), self-organizing maps (SOMs) (33), *k*-means (34), and deterministic annealing (35). Each of these algorithms easily captures the main biological features within data sets. For example, hierarchical clustering, SOMs, and *k*-means algorithms have all been applied to cell cycle data in yeast and have each revealed several broad classes of cell cycle–regulated genes (33, 34, 36). Nonetheless, the differences in the various algorithms produce views of the data that differ in detail with respect to the assignment of genes or samples to particular clusters. There is no ideal approach to the problem that these clustering methods address, namely the projection of a very high dimensional body of data to a lower-dimensional space (often just a one-dimensional ordered list). A reasonable approach, therefore, is to use a variety of different algorithms, each emphasizing distinct orderly features of the data, in order to glean the maximal biological insight.

Figure 2 (see color insert) presents a simple example of hierarchical clustering applied to data from T cell and fibroblast activation experiments (30, 37). Hierarchical clustering begins by determining the gene expression correlation coefficients for each pair of the n genes studied. The two genes with the most correlated expression across all of the samples are fused into a node that is subsequently represented by the average expression of the two genes. This clustering process is then repeated on the $n-1$ genes/nodes that remain. After $n-1$ iterations, all genes are incorporated into a dendrogram that connects each of the nodes generated during the clustering (Figure 2, see color insert). The length of each fork in the dendrogram is inversely proportional to the similarity of the two nodes or genes that it connects. The data in Figure 2 are taken from one experiment with human peripheral blood T cells activated by phytohemagglutinin (PHA) and phorbol-myristoyl-acetate (PMA) and another experiment with human serum-starved fibroblasts activated by readdition of serum. In both experiments, the cells were initially in the G0 stage of the cell cycle and synchronously entered G1 and S phase following stimulation. Each experiment used microarrays containing the same set of 9000 human cDNAs to monitor changes in gene expression over time, comparing mRNA from each stimulated culture with mRNA from resting cells. Figure 2 (see color insert) shows data from a subset of the induced and repressed genes, presented at the left in an unclustered form and, at the right, arranged by hierarchical clustering to reveal coordinately expressed genes.

In this example, the clustering algorithm identified three broad clusters that contain genes activated (*a*) in T cells only, (*b*) in both T cells and fibroblasts, or (*c*) in fibroblasts only. The genes upregulated in both T cells and fibroblasts include *c-myc*, a gene known to be important for progression from G0 to S phase, and genes involved in energy metabolism, presumably reflecting the increased energy requirements of activated cells. Within the T cell–specific cluster are the chemokines MIP-1-alpha and MIP-1-beta, which are known to be coordinately regulated during T cell activation and are important for recruitment of monocytes to regions of immune activation. Interestingly, the aryl-hydrocarbon receptor, the molecular target of dioxin, is specifically induced during T cell activation, perhaps accounting for the ability of dioxin to induce apoptosis in activated, but not resting, mouse T cells (38). The SH2- and SH3-containing protein SLAP (src-like adapter protein) was preferentially induced in T cells. This is noteworthy because SLAP has recently been shown to inhibit cell cycle progression in fibroblasts (39). These microarray data may thus have revealed an unsuspected differential function of SLAP in T cell and fibroblast mitogenesis. Notable among the genes induced preferentially in fibroblasts are basic fibroblast growth factor (basic FGF) and vascular endothelial growth factor (VEGF), both of which are involved in a wound healing response (see below) (30). In addition to these three broad gene expression clusters, there is biologically important fine structure. For example, c-fos, jun B, and MAP kinase phosphatase were all downregulated in late T cell activation, whereas they were induced during the serum response of fibroblasts. The above example highlights several general principles that can

emerge from clustering of gene expression data. As described in the following section, studies of global gene expression patterns in yeast have shown that genes with related biological roles are often tightly coregulated (28, 31, 36, 40, 41). A corollary is that novel genes of unknown function that are clustered with a large group of functionally related genes are likely to participate in the same biological process. In this light, it is interesting to note that several novel genes were selectively induced in T cells rather than fibroblasts (Figure 2, see color insert). Cluster analysis provides a systematic way to focus attention on subsets of the novel genes represented in a survey of gene expression patterns that warrant further investigation in relation to specific biological processes. Finally, Figure 2 demonstrates the usefulness of systematic databases of gene expression measurements that allow fresh biological insights to be made by juxtaposing and comparing data sets from disparate biological systems.

Genomic-Scale Gene Expression Analysis in Model Systems

Whole Genome Gene Expression Analysis in Yeast The most extensive and systematic studies of global gene expression patterns to date have been carried out in *Saccharomyces cerevisiae.* The yeast genome was the first genome of a free-living organism to be completely sequenced, and it has thus been the first model used for development of many functional genomic approaches that can now be applied to mammalian genomes. Over the past three years, several groups have reported studies of genome-wide patterns of gene expression in response to physiological stimuli, drugs, developmental programs or specific mutations in yeast (28, 36, 40, 42–45).

Each such study has provided a wealth of new information and insight into a specific process: the switch from glycolysis to respiration, progression through the cell division cycle, the program of gametogenesis and spore formation, and the targets of specific and global transcriptional regulators. Trivially, these studies provide comprehensive catalogs of the genes whose expression varies in each specific process or in response to each specific perturbation, and the studies define the temporal pattern of each gene's response. But the systematic nature of these observations, involving comprehensive, quantitative measurements of variation in each transcript from the yeast genome, makes it possible to view the entire set of data as one large and expanding survey of the expression program of the yeast genome. A new and remarkably useful kind of gene expression map emerges from this approach. In contrast to conventional genetic maps based on the physical order of genes in the genome, gene expression maps derive their order from the logic underlying the expression program of a genome.

Gene expression maps are constructed by first organizing the gene expression data using any of the various clustering algorithms outlined above. The ordered tables of data are then displayed graphically in a way that allows biologists to assimilate the choreography of gene expression on a broad scale as well as the fine distinctions in expression of individual genes. The large panel on the left of

Figure 3 (see color insert) shows one example of such a map: Each row represents the expression pattern of one of the 6220 known or predicted genes of yeast, and each column represents the results of one of 204 genome-wide microarray gene expression experiments. The expression measurements in this figure were derived from yeast that were placed under 28 distinct physiological or nutritional conditions and assayed multiply over time. For this map, the hierarchical clustering strategy was used to group genes on the basis of similarity in gene expression patterns (31). It is worth noting that this set of over one million measurements of gene expression represents considerably less than 10% of the genome-wide expression data that has been collected over the past two years for this one organism.

A simple evolutionary logic emerges from an analysis of yeast gene expression maps: Genes with similar expression patterns under a particular set of conditions encode protein products that play related roles in the physiological adaptation to those conditions. The extent and precision with which this simple organizing principle determines the geography represented in this map of the genome is unexpected and remarkable (31, 41). Genes encoding products that invariably function together in a stoichiometric complex are virtually always among the most highly coregulated groups in the genome. For example, the vertical bar at the upper right of this map (Figure 3) marks the position of a cluster comprising about 2% of the genes in the yeast genome, including, almost exclusively, all the genes that encode ribosomal proteins. Similar coregulated clusters identify the histones, the subunits of the proteosome, and subunits of numerous other multimeric enzymes (31, 41). Genes whose products work together in a metabolic pathway or a discrete physiological or developmental program are typically less tightly coregulated than components of stoichiometric complexes, but they are sufficiently similar in their expression patterns to cluster together in this genomic expression map. Expanded views of two such clusters are shown at the right of Figure 3 (see color insert): One cluster is composed of genes encoding components of the mitochondrial electron transport and ATP synthase complexes (labeled ‘‘respiration’’), and the other is composed of genes that play key roles in chromosome synapsis and meiotic recombination (labeled ‘‘meiosis’’). Each cluster includes genes without a presently identified function. The consistent relationship between a gene’s expression pattern and its function, reflected in this map, provides the basis for imputing functions to these previously uncharacterized genes. Indeed, the essential role in sporulation for one of the previously uncharacterized genes in the sporulation cluster, YPR007C, which is predicted to encode a putative chromosome cohesion protein, was established following its identification by this cluster analysis method (40). Conversely, each of the conditions represented in this gene expression map (the vertical columns) is characterized by a unique and recognizable signature in its gene expression pattern. Each cell transduces variation in its environment, internal state, and developmental state into readily measured and recognizable variation in gene expression patterns.

Thus the global pattern of gene expression provides a distinctive and accessible molecular picture of the state and identity of biological samples.

The prospects for mapping the regulatory networks that control gene expression programs and connecting them to the corresponding environmental stimuli and the physiological processes that they mediate are already apparent from studies in yeast. These studies have revealed unsuspected complexity in the relationships among regulatory proteins and the genes they control and, at the same time, have provided compelling evidence for the experimental tractability of this problem to systematic dissection (28, 36, 40, 43, 45).

Unconventional Pictures of Biological Responses from Genome-Scale Gene Expression Profiles One of the most useful qualities of the systematic characterization of gene expression programs is that the results are much less constrained by preconceived models than traditional, ''hypothesis-limited,'' experimental approaches. A vivid example of this feature was provided by a genome-scale survey of gene expression changes during the response of serum-deprived cultured human fibroblasts to serum (30). A cDNA microarray representing approximately 9000 different human genes was used to measure gene expression changes at 14 time points following the readdition of serum, beginning 15 min after stimulation and continuing for 24 h. The experiment was intended to provide new insights into the transition from the G0 cell cycle state to a proliferating state since, historically, the serum response of fibroblasts had been viewed as a simple model for this transition. However, the proliferation-related changes in gene expression accounted for only a small fraction of the program of gene expression that was observed in this experiment.

The gene expression program of serum-stimulated fibroblasts was far richer than anticipated and pointed to an important physiological role of fibroblasts in the wound healing response. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that was observed in response to serum suggested that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal signifying a wound. Numerous genes with known roles in processes relevant to wound healing were induced by the serum stimulus. These included genes involved in the direct role of fibroblasts in remodeling the clot and the extracellular matrix as well as genes encoding intercellular signaling proteins that promote inflammation, angiogenesis, and re-epithelialization.

Although this study focused exclusively on the fibroblast and was not intended or expected to address any aspect of immunity, the observed expression program pointed to an important role for fibroblasts in orchestrating the immune response to a wound. The serum-induced genes encoded proteins that promote chemotaxis and activation of neutrophils, monocytes and macrophages, T lymphocytes and B lymphocytes, thus providing innate and antigen-specific defenses against

wound infection. In addition, the recruitment of phagocytic cells is required to clear out the debris during wound remodeling.

The results, unexpectedly, remind us of the importance of viewing an immune response as a concerted physiological program, involving not only cells normally regarded as components of the immune system per se but also virtually any cell that finds itself in a setting where an immune response is called for. The picture painted by the transcriptional response to serum suggests that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals that control inflammation, angiogenesis, and epithelial regrowth during the response to an injury. Another implication of this experiment is that fibroblasts, and very likely many other cells, are programmed to recognize exposure to serum as a signal representing a serious injury. Inclusion of serum in mammalian cell culture medium has become a common, almost ubiquitous, practice. Yet, this experiment suggests that trying to study the normal behavior of cells in the presence of serum may be analogous to trying to study normal human behavior in a burning building.

Signal Transduction One of the natural arenas for genomic-scale gene expression analysis in mammalian systems is signal transduction. It is clear from studies of protein-protein interactions and inducible phosphorylation events that proximal signaling pathways are considerably interwoven. However, not yet known is the extent to which the downstream transcriptional targets of different signaling pathways are overlapping or distinct. For one class of target genes, the immediate early genes, the answer appears to be that disparate signaling pathways converge on virtually identical immediate early target genes (46). Oligonucleotide microarrays were used to compare the immediate early gene response (i.e. genes induced within 4 h of stimulation) of fibroblasts to platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), and epidermal growth factor (EGF), all of which signal through distinct tyrosine kinase receptors. Out of 5938 genes on the array, 66 genes displayed an immediate early response to PDGF. Almost all of these genes were also induced by FGF to the same degree as by PDGF. Correspondingly, these two growth factors cause a quantitatively similar mitogenic response in fibroblasts. Although EGF induced many, but not all, of the same immediate early genes, the magnitude of the induction was quantitatively lower than observed with PDGF and FGF. In this experimental system, therefore, the immediate early genes behave as a transcriptional ''module'' that is invoked to a greater or lesser degree by different cellular stimuli. A second important conclusion from this study was that none of the tyrosines in the cytoplasmic tail of the PDGF receptor was absolutely required for any discrete feature of the immediate early response. This was a surprise because previous work had shown that each tyrosine serves as the docking site for a different signal transduction protein. The results therefore suggest that signal transduction networks must be extensively ramified proximal to the membrane tyrosine kinase receptors, converging on a common set of nuclear immediate early responses.

This is, of course, only one snapshot of the genomic response to membrane signaling events. Since the serum response of fibroblasts induced a stereotypical set of genes beyond the immediate early time frame, it is quite plausible that different receptor kinases will cause distinct delayed transcriptional responses (30). The signaling events through other cell surface receptors certainly will lead to receptor-specific transcriptional responses in some cases. Microarray analysis of cytokine responses, for example, reveals both cytokine-specific and generic transcriptional responses (37). This is not surprising given the direct docking of distinct STAT family transcription factors to the various cytokine receptors (47). Thus, when membrane signaling events lead more directly to the activation and/or nuclear translocation of transcription factors without invoking extensively interconnected proximal signaling networks, signature transcriptional responses may be elicited. Finally, the cell type chosen for signaling experiments will inevitably influence the genomic transcriptional response. For example, a microarray analysis of PMA-responsive genes in myeloid and lymphoid cell lines revealed sets of induced genes that were cell line–specific as well as genes that were PMA-responsive in all myeloid cell lines but not in Jurkat T cells (33). The developmental history of a cell, preserved within heritable chromatin structure or by DNA methylation, will shape the outcome of signaling, as will the different repertoires of transcription factors that are available to various cell types.

The direct target genes of transcription factors can be revealed by genomic-scale gene expression analysis, as illustrated by studies of p53 and BRCA1 (48–51). Inducible overexpression of transcription factors is the experimental design that is currently adopted in most cases. Although valuable, this approach is somewhat risky in that artificial overexpression can lead to nonphysiological titration of protein-protein interactions and binding of transcription factors to inappropriate sites within the genome. Genomic studies of loss-of-function mutants will be an important goal in this field. Analysis of cells taken from knockout animals will be helpful, particularly in cases in which the developmental program has not been overtly altered by the engineered mutation. Large-scale loss-of-function studies in somatic cells in culture await the development of robust methods of gene disruption or interference.

Genomic-Scale Gene Expression Analysis in the Immune System

Ultimately, studies of gene expression in the immune system will examine the entire genomic repertoire of genes in each sample investigated. Although this complete repertoire is not yet available, many insights into the gene expression programs evoked during immune responses can be made using large DNA microarrays that deliberately include many genes known to be expressed in immune cells. An example of such a specialized subgenomic microarray is the Lymphochip, a specialized human cDNA microarray that is enriched for genes related to immune function (8). The Lymphochip microarray is composed of

17,853 cDNA clones derived from three sources. The majority of clones (~80%) were derived from the lymphoid cDNA libraries that were subjected to high-throughput EST sequencing (Table 1). The selection of these clones was based on bioinformatics algorithms that identified ESTs that were either unique or enriched in lymphoid cDNA libraries (8). A second set of Lymphochip clones was identified during the course of previous microarray analyses of immune responses using first-generation microarrays of ~10,000 human genes (37). Last, a curated collection of 3183 ''named'' genes that are of known or suspected importance to immune function, cell proliferation, apoptosis, or oncogenesis and 57 open reading frames from the pathogenic human viruses HIV-1, HTLV-I, EBV, and HHV-6, 7, and 8 were incorporated into the Lymphochip. One of the virtues of mechanically printed microarrays like the Lymphochip, in this era of continuing gene discovery, is that they can be readily upgraded: New genes that are discovered during further high-throughput sequencing or as a result of directed molecular biology experiments can be added to new editions of the Lymphochip in days.

The Genomic Expression Program in Lymphocyte Differentiation Systematic exploration of gene expression programs during human lymphocyte development and activation is under way. Early work has focused on late-stage B cell differentiation, following mature, naive B cells from the resting state through the germinal center reaction and into terminal differentiation. The germinal center is an inducible microenvironment formed during an immune response by the concerted action of antigen-specific B and T cells together with follicular dendritic antigen-presenting cells (FDCs) (52, 53). The germinal center reaction is initiated when the surface immunoglobulin receptor on a B cell encounters its cognate antigen, and activated T cells signal the B cell through CD40. FDCs secrete a gradient of the chemokine BLC, which signals the activated B cell through the BLR1/CXCR5 receptor to migrate toward the FDC (54). Activated T cells also migrate to the nascent germinal center where they continue to interact with germinal center B cells. The germinal center becomes polarized, with highly proliferative centroblast B cells in the ''dark'' zone and less proliferative centrocytes in the ''light'' zone. The process of somatic hypermutation of immunoglobulin genes is initiated in centroblasts, which then migrate to the light zone to become centrocytes. If the hypermutation process has improved, or at least preserved, the ability of the B cell to bind antigen on the surface of the FDC, the B cell is rescued from programmed cell death. The B cell may then migrate back to the dark zone and continue somatic hypermutation or may terminally differentiate into a memory B cell or plasma cell.

B cells at each of these stages of differentiation were purified from human tonsils or peripheral blood, and their transcript patterns were characterized using the Lymphochip microarray (8). As important controls, B cells were activated polyclonally in vitro by ligation of the antigen receptor and activation with CD40 ligand, with and without IL-4. Additionally, T cells were mitogenically activated

with phorbol ester and ionomycin. The gene expression profiles shown in Figure 4 (see color insert) reveal that germinal center B cells represent a distinct stage of B cell differentiation that activates a broad gene expression program that is not observed in mitogenically activated peripheral blood B cells. Germinal center B cells not only express scores of genes that are missing in activated peripheral blood B cells but also lack expression of many genes that are induced during in vitro B cell activation. Thus, coligation of undefined B cell surface receptors, together with stimulation through the antigen receptor and CD40, may be needed to generate the germinal center gene expression profile. Indeed, no convincing in vitro culture system has yet been developed that is able to induce resting peripheral B cells to adopt a full germinal center phenotype. The large set of germinal center B cell–specific genes discovered by microarray analysis can therefore serve as a yardstick to measure the success of in vitro cultures in mimicking the germinal center state.

Mitogenically activated B and T cells shared a common set of activation genes (Figure 4, see color insert), which may reflect the convergence of multiple signaling pathways on common nuclear targets (46) and the fact that the cell cycle gene expression program was activated in both cell types. However, mitogenically activated T cells expressed a distinct set of genes not observed in resting T cells or in activated B cells (not shown). This set of genes includes, of course, various cytokines such as IL-2 and TNF alpha but also a number of novel genes. Based on the coordinate expression of these novel genes with cytokines and the lineage specificity of their expression, they are attractive candidates for functional analysis in the future.

The Relationship of Lymphoid Malignancies to Normal Lymphocyte Differentiation Genomic-scale gene expression profiling is certain to illuminate many aspects of cancer pathogenesis, cancer diagnosis, and the mechanisms underlying treatment resistance and susceptibility. Traditionally, studies of mutations, amplifications, and deletions in the genomic DNA of cancer cells have revealed many of the key genetic events that occur during the progression to cancer. Many of these genetic alterations may have acted for many years prior to diagnosis to bypass key checkpoints and allow cell cycle progression. On the other hand, gene expression profiling of cancer cells reflects the molecular phenotype of the cancer cell at diagnosis. As a consequence, the detailed picture provided by the genomic expression pattern may provide the basis for a new systematic classification of cancers and more accurate predictions of the responses of a cancer to treatment.

A major determinant of the biological potential of a cancer cell is likely to be the normal cell from which it was derived. About 90% of human lymphoid malignancies are derived from B cells, and each of these malignancies has been provisionally assigned to a particular stage of B cell differentiation based on analysis of immunoglobulin gene rearrangement and mutation together with cell surface phenotyping. However, the extent to which the gene expression program of nor-

mal B cells is retained in the cancer cell is best addressed by genomic-scale gene expression analysis.

A particular breeding ground for human lymphomas is thought to be the germinal center reaction. This notion is based on analysis of rearranged immunoglobulin genes in these malignancies, which often show extensive somatic hypermutation (55). Indeed, in two categories of non-Hodgkin's lymphoma, follicular lymphoma and MALT lymphoma, the immunoglobulin sequences from a single biopsy specimen show evidence of ongoing mutation (56–59). In other malignancies in which the immunoglobulin sequences are mutated but invariant, the cell of origin could as well be a postgerminal center B cell. Even the presence of immunoglobulin mutations in a B cell malignancy is not conclusive evidence that the cell of origin passed through the germinal center microenvironment, since in some mutant mouse models, somatic hypermutation of immunoglobulin genes can occur in the absence of detectable germinal centers (60).

The most common form of non-Hodgkin's lymphoma is diffuse large cell lymphoma (DLCL), comprising ~40% of all cases. The immunoglobulin genes in DLCL are invariably mutated. Furthermore, a recurrent translocation in this malignancy involves the BCL-6 gene, a gene also required for normal germinal center development (61–63). However, this translocation occurs in only ~32% of DLCLs, thus revealing potential heterogeneity in this diagnostic category. Patterns of gene expression in a large number of DLCLs were therefore analyzed, using the Lymphochip microarray, to determine the relationship of this malignancy to normal germinal center cells and to investigate the possibility that this diagnostic category may harbor more than one disease entity. Figure 5 (see color insert) shows the expression of a subset of 60 genes from the Lymphochip in 25 different lymph node biopsies of DLCL and in a variety of normal B cell preparations. It is evident that the gene expression patterns in DLCLs are strikingly heterogeneous and that a subset of DLCLs shows a pattern with a strong resemblance to the pattern seen in normal germinal center B cells. Distinct patterns of gene expression identify at least two different subtypes in what has previously been considered a single disease. The similarities in gene expression patterns strongly imply that the cell of origin of one DLCL subtype is the germinal center B cell, but the origin of the other cases is enigmatic. These cases could be derived from a postgerminal center B cell that had extinguished the germinal center gene expression program. Alternatively, the oncogenic transforming event(s) may have disrupted signaling pathways that are critical to maintain the germinal center phenotype.

Preliminary surveys of other B cell malignancies demonstrate that each diagnostic category has its own gene expression signature. Gene expression patterns observed in follicular lymphomas share significant features with the patterns seen in germinal center B cells, whereas the expression patterns in chronic lymphocytic leukemia cells do not resemble those in germinal center cells but instead are reminiscent of resting peripheral blood lymphocytes. Within each of these diagnostic categories, however, the molecular heterogeneity reflected in the gene

expression profiles suggests the existence of disease subtypes, as were revealed in DLCL. The stratification of patients according to gene expression signatures could ultimately contribute to clinical decisions directing the patient to the most appropriate therapy.

Gene Expression Changes During Immune Responses Oligonucleotide arrays have been used to discover gene expression correlates of antigen-induced anergy and activation in B lymphocytes (R Glynne, C Goodnow, personal communication). Transgenic animals expressing heavy and light chains for anti-HEL (hen egg lysozyme) antibody provide B cells of near monoclonality that can be either anergized or activated depending on the method and form of antigen administration (64). Anergic/tolerant B cells are profoundly resistant to subsequent exposure to antigen under activation conditions. B cell anergy involves activation of some but not all of the signaling pathways that are engaged during lymphocyte activation: NF-AT and erk MAP kinase pathways are activated in tolerant cells, whereas NF-κB and jnk pathways are not (65).

Microarray analysis of gene expression in antigen-stimulated naïve B cells demonstrated that 59 genes were significantly induced or repressed after 1 h of stimulation, whereas more than 300 genes were altered in expression after 6 h (R Glynne, C Goodnow, personal communication). By contrast, only 8 of these genes were regulated in tolerant B cells. Instead, tolerant B cells displayed a distinct gene expression signature consisting of 20 upregulated genes and 8 downregulated genes that were not altered during activation of naïve B cells. Interestingly, pharmacological inhibition of NF-AT by the immunosuppresive drug FK506 was less efficient than tolerance in blocking B cell activation responses: One third of the antigen-induced gene expression changes in naïve B cells were unaffected by FK506.

These findings could have important implications for the discovery of novel immunosuppresive drugs. An ideal immunosuppressive drug would have all of the functional effects of natural tolerance without eliciting the side effects that limit the utility of FK506 and cyclosporin in some patients. The gene expression signature of tolerant B cells could by used as a surrogate marker in drug screens for compounds that might mimic the anergic state (R Glynne, C Goodnow, personal communication). Furthermore, if a novel compound in a drug screen induces gene expression changes not found in tolerant cells, this might signal an unwanted ''off-target'' effect of that compound (42).

T cell responses to antigenic and mitogenic stimulation have also been analyzed by cDNA and oligonucleotide microarray analysis (37) (P Marrack, personal communication). Since T cell activation is a well-trodden path, many of the induced genes are well known, including some depicted in Figure 2 (see color insert). Interestingly, an equal number of genes were repressed as were induced during T cell activation, leaving the total diversity of mRNAs roughly equivalent between resting and activated cells. Immunologists have evidently spent less

effort investigating the genes that are downregulated during lymphocyte activation because this class contained more novel genes than the upregulated class.

FUNCTIONAL GENOMICS AND THE GENETICS OF COMPLEX IMMUNOLOGICAL DISEASES

Positional cloning of susceptibility genes for diseases with simple Mendelian inheritance is now routine. However, the majority of medically important genetic diseases show familial clustering with an indeterminant inheritance pattern. The heritable risk for these diseases may be determined by many genes, each of which can affect relative risk for the disease phenotype. In human autoimmune diseases, a sibling of an affected individual has a relative risk of developing the same autoimmune disease of 6–100-fold, compared with the prevalence of the disease within the general population (66). The genetics of autoimmune mouse models has been particularly illustrative of the complexity of some immune-mediated diseases. For example, autoimmune diabetes in the NOD mouse may be controlled by 15 genes on 11 chromosomes (66). The genetic complexity of these diseases is most likely a reflection of their complex pathophysiology. In most autoimmune diseases, the major histocompatibility complex (MHC) plays a dominant role, presumably by dictating which autoantigens can be presented to the immune system. Nevertheless, MHC alleles confer a relative disease risk of only 1.3–8.3-fold (66). Other genetic loci may control the breaking of immunological tolerance, the repertoire of autoimmune T and B cells, the expansion of pathogenic CD4, CD8, and/or B lymphocyte subsets, and the skewing of immune responses by cytokines. One approach to such complex diseases is to artificially simplify the genetics. In the NOD mouse diabetes model, transgenic expression of a single pathogenic T cell receptor has been used to short-circuit some of the disease pathogenesis and to reduce the number of disease susceptibility loci to five genomic intervals (67). Further breeding of this simplified mouse model to knockout animals has revealed a role for interferon gamma in the development of diabetes (68).

Recent reviews have focused on the ways in which genome-wide application of polymorphic markers can identify which genomic intervals may harbor the disease susceptibility genes (66), and therefore we do not review this genomic arena extensively here. A large number of genome-wide screens in immune-mediated diseases of humans and animals have been conducted (67, 69–94). In most cases, however, the genomic interval containing the susceptibility gene has not been narrowed to < 1 centiMorgan ($\sim 1 \times 10^6$ base pairs) by these methods. Interestingly, the susceptibility loci in these various diseases often coincide, suggesting that some common genes may influence many autoimmune and inflammatory diseases (66, 95).

The molecular definition of specific susceptibility alleles in complex immune-mediated diseases will clearly require new strategies that complement genetic linkage analysis. One functional genomics strategy that holds promise is the ''positional candidate gene'' approach (reviewed in 96). Having narrowed the susceptibility interval by linkage analysis, the known genes that map within the interval can be identified. Since mutation detection is still technically cumbersome and a 1-centiMorgan susceptibility region could contain 30 or more genes, it may be helpful to first focus attention on candidate genes with functions that can be plausibly connected to the disease phenotype. One potential identifying characteristic of a candidate susceptibility gene would be expression in the cells or tissues presumed to be at fault in the disease. Soon, public databases of gene expression measurements will make this analysis routine. Even when genetic linkage has not been performed, the candidate gene approach may quickly reveal potential disease susceptibility loci. For example, the candidate gene approach was used to examine the genetic differences between chronic granulomatous disease (CGD) patients who differed in susceptibility to immune-mediated complications (97). CGD results from a primary defect in genes for NADPH oxidase that control superoxide production in phagocytes. CGD patients differ dramatically in the frequency with which they develop a variety of chronic complications, including granulomatous diseases of the gastrointestinal and urinary systems as well as autoimmune and rheumatological disorders. A priori, such differences may be due to differences in the mutations present in the 4-NAPDH oxidase subunit genes found in different patients. Alternatively, polymorphisms in other disease-modifying genes could contribute to the risk of immune complications. Foster et al examined polymorphisms in seven candidate genes encoding myeloperoxidase, mannose binding lectin, TNF alpha, IL-1 receptor antagonist, and the Fc gamma receptors IIa, IIIa, and IIIb (97). Alleles of myeloperoxidase and Fc gamma receptor IIIb were significantly associated with an enhanced risk for gastrointestinal complications. Alleles of myeloperoxidase were associated with increased risk of autoimmune and rheumatological disorders. Combinations of specific alleles of different genes conferred an even greater relative risk for chronic immune-mediated complications. In this relatively rare genetic disease it would be difficult, if not impossible, to enroll enough patients to conduct a standard linkage analysis of immune complications; thus the candidate gene approach provides a tractable alternative.

cDNA microarray analysis of gene expression promises to aid significantly in the search for disease susceptibility loci. One example of this approach comes from analysis of the spontaneously hypertensive SHR rat, which is a model for human diabetes, hyperlipidemia, obesity, and hypertension (98). Genetic linkage analysis of this disorder had focused attention on an interval from rat chromosome 4, but the causative gene had not been identified. cDNA microarrays were used to compare gene expression in adipose tissue from the SHR strain and control, nonhypertensive rat strains. SHR cDNA probes hybridized poorly to a microarray spot representing CD36, a gene that maps to regions of mouse and human chro-

mosomes that are syntenic to rat chromosome 4. This microarray finding prompted a further analysis of the CD36 gene, which revealed multiple coding region mutations in the CD36 gene of SHR rats. CD36 is a fatty acid receptor and transporter whose overexpression in transgenic mice or deletion by gene disruption in mice leads to alterations in blood lipid levels (98, 99); thus it is a strong candidate gene for the hyperlipidemic quantitative trait in SHR rats. The apparently diminished expression of CD36 in SHR rats, detected by cDNA microarray hybridization, was traced to a genomic deletion in this strain within the CD36 3′ untranslated region, the only region of the CD36 gene represented on the array.

Although this may appear to be an exceptional case, in that the genetic lesion in this example directly affected the ability to measure the expression of the gene, it is likely that many disease-causing mutations will be found to affect transcript levels. Nonsense mutations and mutations that disrupt normal splicing can lead to reduced mRNA levels via nonsense-mediated decay mechanisms. Many mutations, including many classical genetic disease-causing mutations (e.g. many thalassemias), directly alter transcription of the affected gene. Indeed, the evolutionary constraints on mutation in perigenic noncoding regions, reflected in limited sequence polymorphism observed in these regions as compared to degenerate positions in coding sequences, argue that the potential for deleterious consequences from mutations in regulatory sequences of genes rivals that of mutations in protein-coding sequences (4, 100).

Perhaps a more common use of cDNA microarrays in the investigation of genetic diseases will be to detect quantitative differences in the expression of genes between different animal strains or different human individuals. Quantitative traits that distinguish individuals of the same species undoubtedly arise as a result of both coding region polymorphisms that alter the function of a gene product and regulatory region polymorphisms that affect the expression level of the mRNA or protein. The relative contribution of these two types of allelic differences to genetic diversity is unclear at present, but cDNA microarray analysis may soon reveal a broad range of quantitative gene expression traits within the immune system. Polymorphisms in the mouse TNF alpha gene have been described that affect TNF alpha levels and can modulate the development of nephritis in animals predisposed to systemic lupus erythematosis (101). Similarly, regulatory mutations in the human TNF alpha gene have been associated with a wide variety of diseases, but the interpretations of such studies in humans is complicated by the location of the TNF alpha gene in the MHC and the difficulty in teasing apart the contributions of allelic differences in the TNF alpha gene and the MHC genes.

An interesting quantitative gene expression trait was recently described involving *FRIP*, a gene that encodes an adapter protein involved in IL-4 signaling (102). The *FRIP* gene was mapped to a region of mouse chromosome 14 very close to the gene for the hairless mutation. The hairless mutation results from the insertion of an endogenous mouse retrovirus into the mouse hairless locus (103). The hair-

less gene is also mutated in human alopecia universalis and encodes a pioneer protein of unknown function (104). The hairless mouse also has immune abnormalities, including lymphadenopathy and augmented response of anti-CD3 stimulated T cells to IL-4 (102). The *FRIP* gene is expressed at significantly lower levels in hairless mice compared with wild-type mice, possibly as a result of the same retroviral insertion event. Given the proximity of *FRIP* to the hairless gene and the IL-4-related abnormalities of hairless mice, the *FRIP* quantitative gene expression trait may well account for the lymphoproliferative disorder in these mice.

Genomic-scale gene expression analysis may help to unravel complex genetic diseases by defining more precisely the disease ''phenotype.'' As a hypothetical example, suppose one of the disease susceptibility genes involved in a complex immunological disease regulates responsiveness of T lymphocytes to IL-2. cDNA microarray analysis of IL-2-stimulated peripheral blood T cells might therefore reveal a gene expression profile that correlates with the presence of this susceptibility allele. This gene expression correlate of the susceptibility gene might be observed in family members of affected individuals who are clinically ''normal'' due to segregation of other susceptibility alleles. Linkage analysis using polymorphic markers could be applied to this gene expression phenotype rather than to the whole clinical syndrome, thereby isolating one component of the complex disease phenotype.

GENE EXPRESSION PROFILES OF PERIPHERAL BLOOD AS SENTINELS OF DISEASE

Immune cells circulate throughout the body responding to internal and external threats to homeostasis. Circulating white blood cells are charged with the task of seeking out, recognizing, and mounting a suitable response to the earliest signs of an infection or injury. The sensitive and diverse repertoire of receptors and signal transduction systems that cells use to monitor and respond to trouble at any site in the body may well give rise to signature patterns of altered gene expression in peripheral blood cells reflecting the nature and site of an infection or injury. It is plausible that gene expression patterns in specific subsets of peripheral blood cells might be altered in characteristic ways in response to the presence of specific occult infectious agents. As a consequence, peripheral blood mononuclear cells might display a pathognomonic gene expression signature that could be used to diagnose occult disease. Gene expression changes induced by cytomegalovirus (CMV) infection of fibroblasts and HTLV-I infection of T lymphocytes were studied using microarrays and revealed, not surprisingly, largely distinct gene expression profiles (105, 106). Infection with CMV invoked a strong interferon response that was not observed with HTLV-I, whereas HTLV-1 induced a number of NF-κB target genes as a consequence of the nuclear translocation of NF-κB induced by the HTLV-I tax protein. Recently, the response of monocytic

cells to bacterial exposure in vitro was monitored using Lymphochip cDNA microarrays (D Relman, in preparation). *B. pertussis, H. pylori,* and *S. typhimurium* each induced a distinctive gene expression profile in the monocytes. Further, mutant strains of *B. pertussis* that lacked individual toxin genes elicited gene expression changes that differed from the response to the wild-type strain. Thus, gene expression profiles could be used not only to recognize exposure to an infectious agent but perhaps to identify the agent or category of agent, based on specific characteristics of the response. This ability would clearly be especially useful in cases in which the agent cannot be readily cultured from the host. Since gene expression responses to infectious agents take place within the first few hours after exposure, gene expression profiling might be useful in diagnosing infectious exposure in advance of clinical symptoms, allowing exposed patients to be rapidly triaged for treatment. Finally, the course of infection and the ensuing host response could potentially be monitored by changes in peripheral blood gene expression. This approach could aid in the management of sepsis, which is a disease characterized by an orderly progression of pathophysiological events (107). Gene expression profiles could thus be used to stratify patients into distinct pathophysiological groups and, ultimately, treat each group with a therapy tailored to the disease stage.

It is not difficult to imagine a wider range of clinical settings in which peripheral blood gene expression profiles might aid in patient management. Exposure to toxic xenobiotic compounds such as dioxin should be readily detectable by virtue of the expression of the aryl hydrocarbon receptor in activated T cells (Figure 2, see color insert). T cells from cancer patients display an anergic phenotype, partly due to loss of the zeta chain of the T cell receptor (108, 109), which should result in a gene expression signature in peripheral blood cells. In cases of occult malignancy, such as often occurs in ovarian cancer, this gene expression signature might be detectable in advance of clinical symptoms. In autoimmune diseases such as multiple sclerosis, changes in peripheral blood gene expression may precede a clinical exacerbation, allowing clinicians to time immunosuppressive treatment optimally. Recognition of characteristic patterns of gene expression in circulating peripheral blood cells may thus prove broadly useful as an approach to noninvasive diagnostics, in effect recruiting these readily accessible cells as ''spies'' to report the presence of occult infection or injury they have encountered during their surveillance of the integrity of the body. Of course, as with any clinical test, gene expression profiling of blood mononuclear cells must be both sensitive and specific to be useful. If only a small fraction of peripheral blood cells responds to a pathological event, microarray analysis of gene expression may not be a sufficiently sensitive test. Furthermore, we do not yet know the extent of normal variation in gene expression patterns in peripheral blood cells, nor the extent to which they are altered by everyday, non–life threatening events. For example, it will be necessary (and very interesting in its own right) to catalog the effects of upper respiratory viral infections, stress hormones, age, sex, and even circadian rhythms on gene expression in peripheral blood cells. As already mentioned, genetic variation in immune regulatory genes may be associated with

quantitative gene expression traits that will need to be considered in interpreting gene expression profiles of blood cells. Given the rich insights that genomic-scale gene expression analysis has already provided, we can be optimistic that this new mode of biological discovery will illuminate many issues in clinical pathophysiology.

ACKNOWLEDGMENTS

The microarray work summarized in this paper results from a collaborative effort of the authors' laboratories with a variety of researchers at the National Cancer Institute, NIH; Stanford University; Center for Information Technology, NIH; CBER, FDA, University of Nebraska Medical Center, and Research Genetics. Key individuals who contributed to the Lymphochip project are Ash Alizadeh, Mike Eisen, R Eric Davis, Chi Ma, Hajeer Sabet, Truc Tran, John Powell, Liming Yang, Gerry Marti, Troy Moore, Jim Hudson, John Chan, Tim Greiner, Denny Weisenburger, Jim Armitage, Izadore Lossos, Ron Levy, and David Botstein. The authors thank Richard Glynne, Chris Goodnow, Philippa Marrack, and David Relman for communicating results prior to publication.

LITERATURE CITED

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science* 274:546:563–67
2. Consortium TC e S. 1998. Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science* 282:2012–18
3. Clayton RA, White O, Fraser CM. 1998. Findings emerging from complete microbial genome sequences. *Curr. Opin. Microbiol.* 1:562–66
4. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalayanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22:231–38
5. Uberbacher EC, Xu Y, Mural RJ. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. *Meth. Enzymol.* 266:259–81
6. Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC. 1992. Sequence identification of 2,375 human brain genes. *Nature* 355:632–34
7. Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST—database for ''expressed sequence tags.'' *Nat. Genet.* 4:332–33
8. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell J, Yang L, Marti G, Moore T, Hudson J, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM. 1999. The Lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lympho-

cytes. *Cold Spring Harbor Symp. Quant. Biol.* In press
9. Strausberg RL, Dahl CA, Klausner RD. 1997. New opportunities for uncovering the molecular basis of cancer. *Nat. Genet.* 15(Spec. No.):415–16
10. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* 91:9228–32
11. Schuler GD. 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* 75:694–98
12. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* 270:484–87
13. Chen H, Centola M, Altschul SF, Metzger H. 1998. Characterization of gene expression in resting and activated mast cells. *J. Exp. Med.* 188:1657–68
14. Southern EM, Maskos U, Elder JK. 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* 13: 1008–17
15. Southern EM, Case-Green SC, Elder JK, Johnson M, Mir KU, Wang L, Williams JC. 1994. Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res.* 22:1368–73
16. Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Sampson R, Houlgatte R, Soularue P, Auffray C. 1996. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* 6:492–503
17. Gress TM, Muller-Pillasch F, Geng M, Zimmerhackl F, Zehetner G, Friess H, Buchler M, Adler G, Lehrach H. 1996. A pancreatic cancer-specific expression profile. *Oncogene* 13:1819–30
18. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* 91:5022–26
19. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274:610–14
20. Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP. 1995. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 19:442–47
21. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675–80
22. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15:1359–67
23. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70
24. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93:10614–19
25. Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6:639–45
26. Lashkari DA, DeRisi JL, McCusker H, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94:13057–62
27. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE,

Davis RW. 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94:2150–55
28. DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–86
29. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14:457–60
30. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83–87
31. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68
32. Kohonen T. 1997. *Self-Organizing Maps.* Berlin: Springer
33. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–12
34. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22:281–85
35. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745–50
36. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* 9:3273–97
37. Alizadeh A, Eisen M, Botstein D, Brown PO, Staudt LM. 1998. Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.* 18:373–79
38. Pryputniewicz SJ, Nagarkatti M, Nagarkatti PS. 1998. Differential induction of apoptosis in activated and resting T cells by 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and its repercussion on T cell responsiveness. *Toxicology* 129:211–26
39. Roche S, Alonso G, Kazlauskas A, Dixit VM, Courtneidge SA, Pandey A. 1998. Src-like adaptor protein (SLAP) is a negative regulator of mitogenesis. *Curr. Biol.* 8:975–78
40. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705
41. Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21:33–37
42. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, Bassett DE Jr, Hartwell LH, Brown PO, Friend SH. 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* 4:1293–1301
43. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2:65–73
44. Gray NS, Wodicka L, Thunnissen AM, Norman TC, Kwon S, Espinoza FH, Morgan DO, Barnes G, LeClerc S, Meijer L, Kim SH, Lockhart DJ, Schultz PG. 1998. Exploiting chemical libraries,

structure, and genomics in the search for kinase inhibitors. *Science* 281:533–38
45. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–28
46. Fambrough D, McClure K, Kazlauskas A, Lander ES. 1999. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* 97:727–41
47. Darnell JE, Kerr IM, Stark GR. 1994. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* 264:1415–21
48. Harkin DP, Bean JM, Miklos D, Song YH, Truong VB, Englert C, Christians FC, Ellisen LW, Maheswaran S, Oliner JD, Haber DA. 1999. Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* 97:575–86
49. Amundson SA, Bittner M, Chen Y, Trent J, Meltzer P, Fornace AJ Jr. 1999. Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. *Oncogene* 18:3666–72
50. Hermeking H, Lengauer C, Polyak K, He TC, Zhang L, Thiagalingam S, Kinzler KW, Vogelstein B. 1997. 14–3-3 sigma is a p53-regulated inhibitor of G2/M progression. *Mol. Cell.* 1:3–11
51. Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B. 1997. A model for p53-induced apoptosis. *Nature* 389:300–5
52. Kelsoe G. 1996. The germinal center: a crucible for lymphocyte selection. *Semin. Immunol.* 8:179–84
53. MacLennan ICM. 1994. Germinal centers. *Annu. Rev. Immunol.* 12:117–39
54. Gunn MD, Ngo VN, Ansel KM, Ekland EH, Cyster JG, Williams LT. 1998. A B-cell-homing chemokine made in lymphoid follicles activates Burkitt's lymphoma receptor-1. *Nature* 391:799–803
55. Klein U, Goossens T, Fischer M, Kanzler H, Braeuninger A, Rajewsky K, Kuppers R. 1998. Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* 162:261–80
56. Du M, Diss TC, Xu C, Peng H, Isaacson PG, Pan L. 1996. Ongoing mutation in MALT lymphoma immunoglobulin gene suggests that antigen stimulation plays a role in the clonal expansion. *Leukemia* 10:1190–97
57. Bahler DW, Levy R. 1992. Clonal evolution of a follicular lymphoma: evidence for antigen selection. *Proc. Natl. Acad. Sci. USA* 89:6770–74
58. Bahler DW, Miklos JA, Swerdlow SH. 1997. Ongoing Ig gene hypermutation in salivary gland mucosa-associated lymphoid tissue-type lymphomas. *Blood* 89:3335–44
59. Qin Y, Greiner A, Hallas C, Haedicke W, Muller-Hermelink HK. 1997. Intraclonal offspring expansion of gastric low-grade MALT-type lymphoma: evidence for the role of antigen-driven high-affinity mutation in lymphomagenesis. *Lab. Invest.* 76:477–85
60. Matsumoto M, Lo SF, Carruthers CJ, Min J, Mariathasan S, Huang G, Plas DR, Martin SM, Geha RS, Nahm MH, Chaplin DD. 1996. Affinity maturation without germinal centres in lymphotoxin-alpha-deficient mice. *Nature* 382: 462–66
61. Fukuda T, Yoshida T, Okada S, Hatano M, Miki T, Ishibashi K, Okabe S, Koseki H, Hirosawa S, Taniguchi M, Miyasaka N, Tokuhisa T. 1997. Disruption of the Bcl6 gene results in an impaired germinal center formation. *J. Exp. Med.* 186:439–48
62. Ye BH, Cattoretti G, Shen Q, Zhang J, Hawe N, de Waard R, Leung C, Nouri-Shirazi M, Orazi A, Chaganti RS, Rothman P, Stall AM, Pandolfi PP, Dalla-Favera R. 1997. The BCL-6 proto-

oncogene controls germinal-centre formation and Th2- type inflammation. *Nat. Genet.* 16:161–70

63. Dent AL, Shaffer AL, Yu X, Allman D, Staudt LM. 1997. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* 276:589–92
64. Goodnow CC, Cyster JG, Hartley SB, Bell SE, Cooke MP, Healy JI, Akkaraju S, Rathmell JC, Pogue SL, Shokat KP. 1995. Self-tolerance checkpoints in B lymphocyte development. *Adv. Immunol.* 59:279–368
65. Healy JI, Goodnow CC. 1998. Positive versus negative signaling by lymphocyte antigen receptors. *Annu. Rev. Immunol.* 16:645–70
66. Vyse TJ, Todd JA. 1996. Genetic analysis of autoimmune disease. *Cell* 85:311–18
67. Gonzalez A, Katz JD, Mattei MG, Kikutani H, Benoist C, Mathis D. 1997. Genetic control of diabetes progression. *Immunity* 7:873–83
68. Wang B, Andre I, Gonzalez A, Katz JD, Aguet M, Benoist C, Mathis D. 1997. Interferon-gamma impacts at multiple points during the progression of autoimmune diabetes. *Proc. Natl. Acad. Sci. USA* 94:13844–49
69. Moser KL, Neas BR, Salmon JE, Yu H, Gray-McGuire C, Asundi N, Bruner GR, Fox J, Kelly J, Henshall S, Bacino D, Dietz M, Hogue R, Koelsch G, Nightingale L, Shaver T, Abdou NI, Albert DA, Carson C, Petri M, Treadwell EL, James JA, Harley JB. 1998. Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. *Proc. Natl. Acad. Sci. USA* 95:14869–74
70. Todd JA. 1995. Genetic analysis of type 1 diabetes using whole genome approaches. *Proc. Natl. Acad. Sci. USA* 92:8560–65
71. Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, Farmer J, Williams SR, Cox NJ, Bell GI, Risch N, Spielman RS. 1998. A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* 19:292–96
72. Satsangi J, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K, Terwilliger JD, Lathrop GM, Bell JI, Jewell DP. 1996. Two-stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat. Genet.* 14:199–202
73. Sundvall M, Jirholt J, Yang HT, Jansson L, Engstrom A, Pettersson U, Holmdahl R. 1995. Identification of murine loci associated with susceptibility to chronic experimental autoimmune encephalomyelitis. *Nat. Genet.* 10:313–17
74. Ebers GC, Kukay K, Bulman DE, Sadovnick AD, Rice G, Anderson C, Armstrong H, Cousin K, Bell RB, Hader W, Paty DW, Hashimoto S, Oger J, Duquette P, Warren S, Gray T, O'Connor P, Nath A, Auty A, Metz L, Francis G, Paulseth JE, Murray TJ, Pryse-Phillips W, Risch R, et al. 1996. A full genome search in multiple sclerosis. *Nat. Genet.* 13:472–76
75. Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, Robertson N, Clayton D, Goodfellow PN, Compston A. 1996. A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. *Nat. Genet.* 13:464–68
76. Haines JL, Ter-Minassian M, Bazyk A, Gusella JF, Kim DJ, Terwedow H, Pericak-Vance MA, JB Rimmler, Haynes CS, Roses AD, Lee A, Shaner B, Menold M, Seboun E, Fitoussi RP, Gartioux C, Reyes C, Ribierre F, Gyapay G, Weissenbach J, Hauser SL, Goodkin DE, Lincoln R, Usuku K, Oksenberg JR, et al. 1996. A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. *The Multi-*

ple Sclerosis Genetics Group. *Nat. Genet.* 13:469–71
77. Kuokkanen S, Sundvall M, Terwilliger JD, Tienari PJ, Wikstrom J, Holmdahl R, Pettersson U, Peltonen L. 1996. A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2. *Nat. Genet.* 13:477–80
78. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM, et al. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 371:130–36
79. Jacob HJ, Pettersson A, Wilson D, Mao Y, Lernmark A, Lander ES. 1992. Genetic dissection of autoimmune type I diabetes in the BB rat. *Nat. Genet.* 2:56–60
80. Hashimoto L, Habita C, Beressi JP, Delepine M, Besse C, Cambon-Thomsen A, Deschamps I, Rotter JI, Djoulah S, James MR, et al. 1994. Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. *Nature* 371:161–64
81. Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, Bonaiti-Pellie C, Weissenbach J, Mathew CG, Lennard-Jones JE, Cortot A, Colombel JF, Thomas G. 1996. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 379:821–23
82. Matthews D, Fry L, Powles A, Weber J, McCarthy M, Fisher E, Davies K, Williamson R. 1996. Evidence that a locus for familial psoriasis maps to chromosome 4q. *Nat. Genet.* 14:231–33
83. Tomfohrde J, Silverman A, Barnes R, Fernandez-Vina MA, Young M, Lory D, Morris L, Wuepper KD, Stastny P, Menter A, et al. 1994. Gene for familial psoriasis susceptibility mapped to the distal end of human chromosome 17q. *Science* 264:1141–45
84. Daniels SE, Bhattacharrya S, James A, Leaves NI, Young A, Hill MR, Faux JA, Ryan GF, le Souef PN, Lathrop GM, Musk AW, Cookson WO. 1996. A genome-wide search for quantitative trait loci underlying asthma. *Nature* 383:247–50
85. Remmers EF, Longman RE, Du Y, O'Hare A, Cannon GW, Griffiths MM, Wilder RL. 1996. A genome scan localizes five non-MHC loci controlling collagen-induced arthritis in rats. *Nat. Genet.* 14:82–85
86. Kono DH, Burlingame RW, Owens DG, Kuramochi A, Balderas RS, Balomenos D, Theofilopoulos AN. 1994. Lupus susceptibility loci in New Zealand mice. *Proc. Natl. Acad. Sci. USA* 91:10168–72
87. Morel L, Rudofsky UH, Longmate JA, Schiffenbauer J, Wakeland EK. 1994. Polygenic control of susceptibility to murine systemic lupus erythematosus. *Immunity* 1:219–29
88. Baker D, Rosenwasser OA, O'Neill JK, Turk JL. 1995. Genetic analysis of experimental allergic encephalomyelitis in mice. *J. Immunol.* 155:4046–51
89. Jirholt J, Cook A, Emahazion T, Sundvall M, Jansson L, Nordquist N, Pettersson U, Holmdahl R. 1998. Genetic linkage analysis of collagen-induced arthritis in the mouse. *Eur. J. Immunol.* 28:3321–28
90. Ghosh S, Palmer SM, Rodrigues NR, Cordell HJ, Hearne CM, Cornall RJ, Prins JB, McShane P, Lathrop GM, Peterson LB, et al. 1993. Polygenic control of autoimmune diabetes in nonobese diabetic mice. *Nat. Genet.* 4:404–9
91. Dallas-Pedretti A, McDuffie M, Haskins K. 1995. A diabetes-associated T-cell autoantigen maps to a telomeric locus on mouse chromosome 6. *Proc. Natl. Acad. Sci. USA* 92:1386–90
92. Rowe RE, Wapelhorst B, Bell GI, Risch N, Spielman RS, Concannon P. 1995. Linkage and association between insulin-dependent diabetes mellitus (IDDM) sus-

ceptibility and markers near the glucokinase gene on chromosome 7. *Nat. Genet.* 10:240–42
93. Field LL, Tobias R, Magnus T. 1994. A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* 8:189–94
94. Gaffney PM, Kearns GM, Shark KB, Ortmann WA, Selby SA, Malmgren ML, Rohlf KE, Ockenden TC, Messner RP, King RA, Rich SS, Behrens TW. 1998. A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. *Proc. Natl. Acad. Sci. USA* 95:14875–79
95. Becker KG, Simon RM, Bailey-Wilson JE, Freidlin B, Biddison WE, McFarland HF, Trent JM. 1998. Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc. Natl. Acad. Sci. USA* 95:9979–84
96. Odunsi K, Kidd KK. 1999. A paradigm for finding genes for a complex human trait: polycystic ovary syndrome and follistatin. *Proc. Natl. Acad. Sci. USA* 96:8315–17
97. Foster CB, Lehrnbecher T, Mol F, Steinberg SM, Venzon DJ, Walsh TJ, Noack D, Rae J, Winkelstein JA, Curnutte JT, Chanock SJ. 1998. Host defense molecule polymorphisms influence the risk for immune-mediated complications in chronic granulomatous disease. *J. Clin. Invest.* 102:2146–55
98. Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PF, Wahid FN, Al-Majali KM, Trembling PM, Mann CJ, Shoulders CC, Graf D, St. Lezin E, Kurtz TW, Kren V, Pravenec M, Ibrahimi A, Abumrad NA, Stanton LW, Scott J. 1999. Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats [see comments]. *Nat. Genet.* 21:76–83
99. Febbraio M, Abumrad NA, Hajjar DP, Sharma K, Cheng W, Pearce SF, Silverstein RL. 1999. A null mutation in murine CD36 reveals an important role in fatty acid and lipoprotein metabolism. *J. Biol. Chem.* 274:19055–62
100. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22:239–47
101. Jacob CO, McDevitt HO. 1988. Tumour necrosis factor-alpha in murine autoimmune 'lupus' nephritis. *Nature* 331:356–58
102. Nelms K, Snow AL, Hu-Li J, Paul WE. 1998. FRIP, a hematopoietic cell-specific rasGAP-interacting protein phosphorylated in response to cytokine stimulation. *Immunity* 9:13–24
103. Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM. 1988. Role of endogenous retroviruses as mutagens: the hairless mutation of mice. *Cell* 54:383–91
104. Ahmad W, Faiyaz ul Haque M, Brancolini V, Tsou HC, ul Haque S, Lam H, Aita VM, Owen J, deBlaquiere M, Frank J, Cserhalmi-Friedman PB, Leask A, McGrath JA, Peacocke M, Ahmad M, Ott J, Christiano AM. 1998. Alopecia universalis associated with a mutation in the human hairless gene. *Science* 279:720–24
105. Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T. 1998. Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 95:14470–75
106. Harhaj EW, Good L, Xiao G, Sun SC. 1999. Gene expression profiles in HTLV-I-immortalized T cells: deregulated expression of genes involved in apoptosis regulation. *Oncogene* 18:1341–49
107. Natanson C, Hoffman WD, Suffredini AF, Eichacker PQ, Danner RL. 1994.

Selected treatment strategies for septic shock based on proposed mechanisms of pathogenesis. *Ann. Intern. Med.* 120: 771–83

108. Mizoguchi H, O'Shea JJ, Longo DL, Loeffler CM, McVicar DW, Ochoa AC. 1992. Alterations in signal transduction molecules in T lymphocytes from tumor-bearing mice. *Science* 258:1795–98
109. Whiteside TL. 1998. Immune cells in the tumor microenvironment. Mechanisms responsible for functional and signaling defects. *Adv. Exp. Med. Biol.* 451:167–71

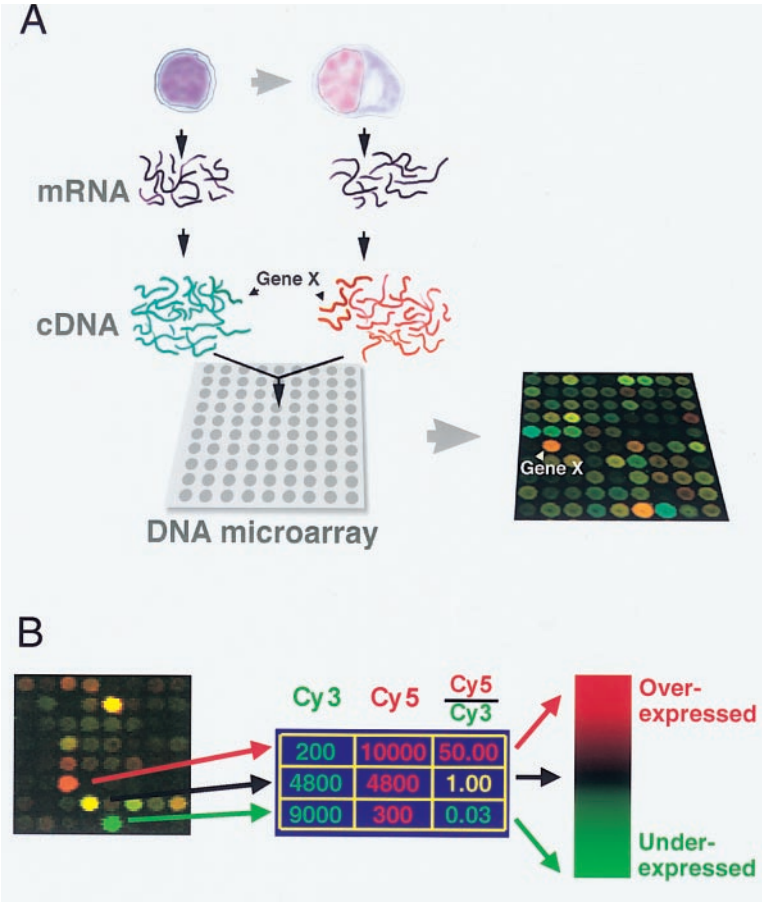

Figure 1 **A** Schematic of cDNA microarray gene expression analysis. In this illustration, the relative gene expression in a mature B cell and a plasma cell is compared. Gene X represents a gene more highly expressed in the plasma cell. See text for details. **B** Quantitative analysis of relative gene expression using cDNA microarrays. For each spot on the microarray, the fluorescence intensities of hybridized Cy3- and Cy5-labelled cDNA probes are separately quantitated, as shown in the middle panel. The Cy5/Cy3 fluorescence intensity ratio is a measure of relative gene expression in the two starting mRNA samples. The fluorescence ratios are divided into numerical bins and depicted visually using the color scale shown at the right.

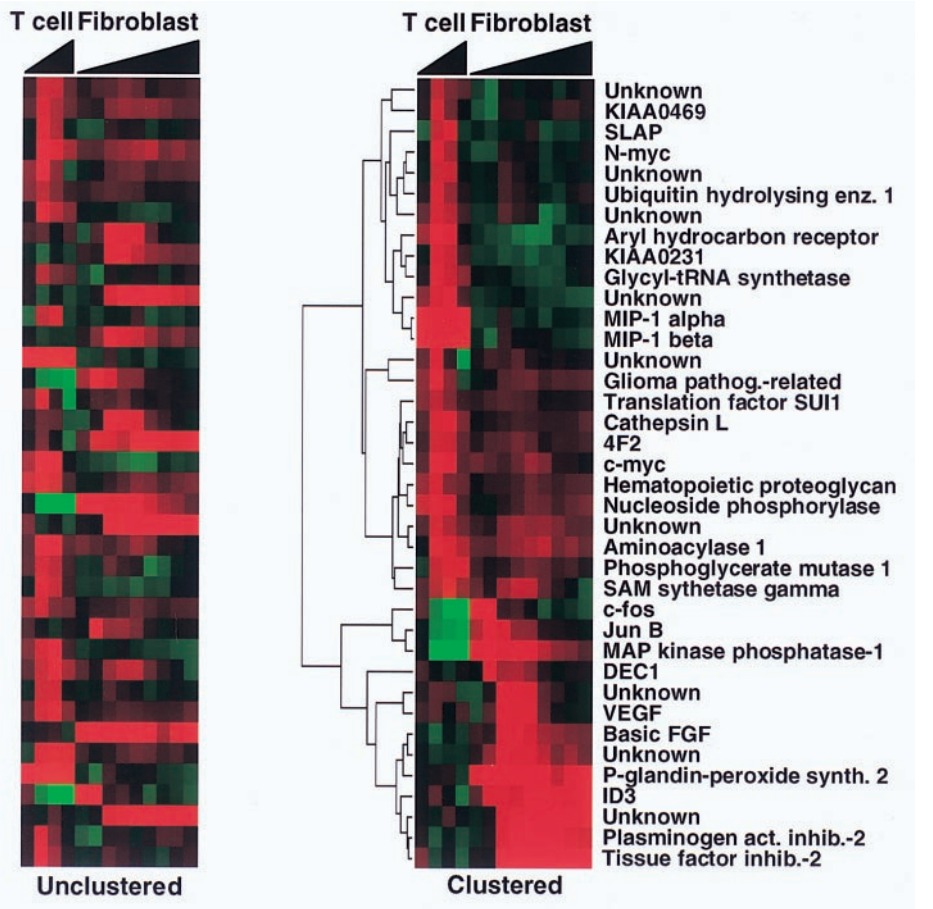

Figure 2 Gene expression profiles identified using hierarchical clustering algorithms. DNA microarray measurements of gene expression were taken from mitogenically activated T lymphocytes and serum-stimulated fibroblasts over a time course. Gene expression at each time point was measured relative to gene expression in unstimulated cells. The relative gene expression data is depicted using the color scheme of Figure 1B with the brightest red and green boxes representing eightfold induced or repressed genes, respectively. The left panel displays the genes in random order, while the right panel displays the genes in the order determined by hierarchical clustering.

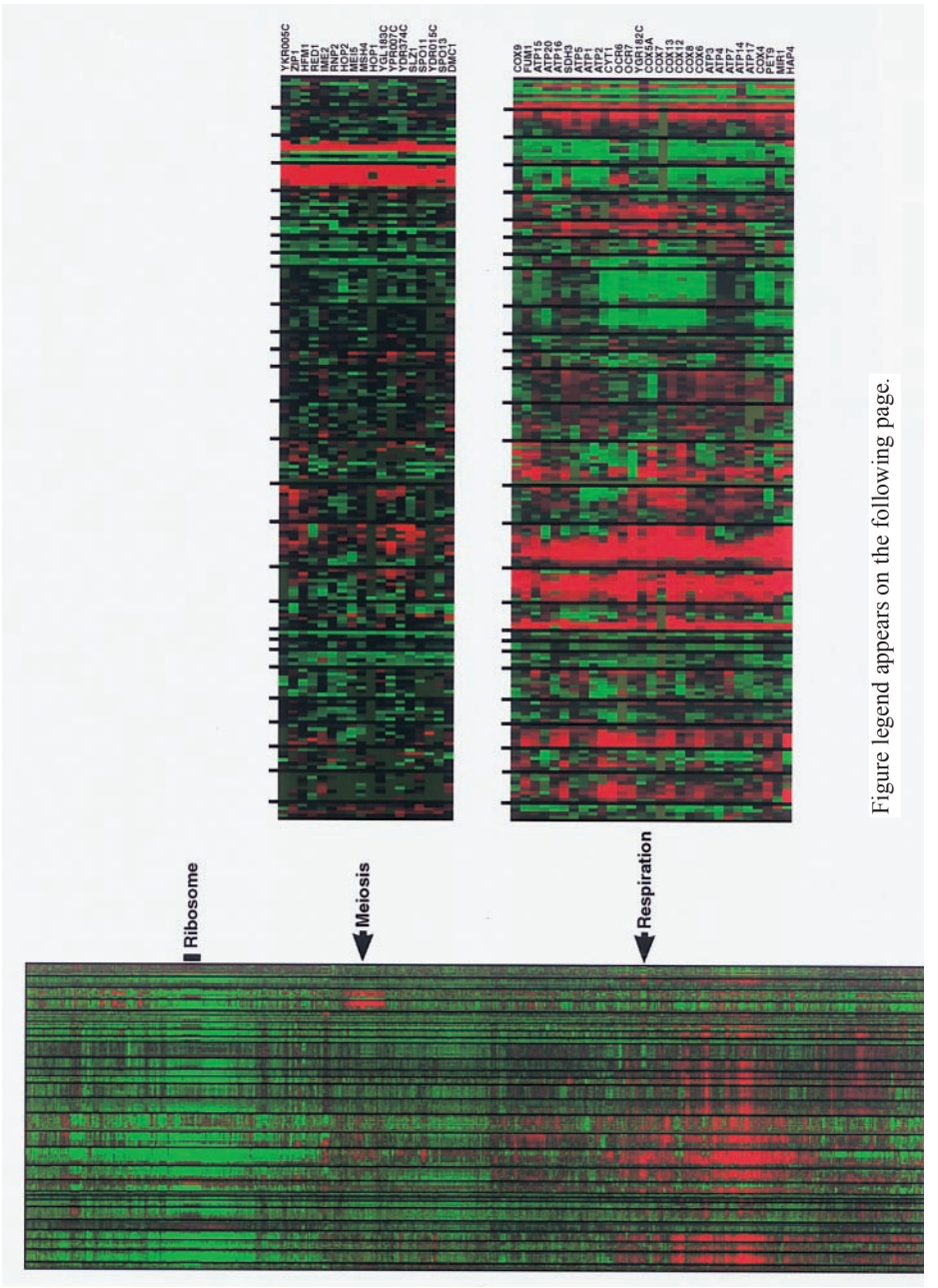

Figure legend appears on the following page.

Figure 3 (see preceding page) A gene expression map of *Saccharomyces cerevisiae*. The left panel depicts the results from 204 genome-wide microarray gene expression experiments in *Saccharomyces cerevisiae*. Yeast were placed under 28 distinct physiological or nutritional conditions (delineated by the vertical black stripes) and assayed multiply over time. Each column represents one microarray experiment, and each row represents one of the 6220 known or predicted genes of yeast. The coordinate regulation of genes encoding ribosomal subunits is indicated. Genes involved in meiosis and respiration form separate clusters which are expanded at the right.

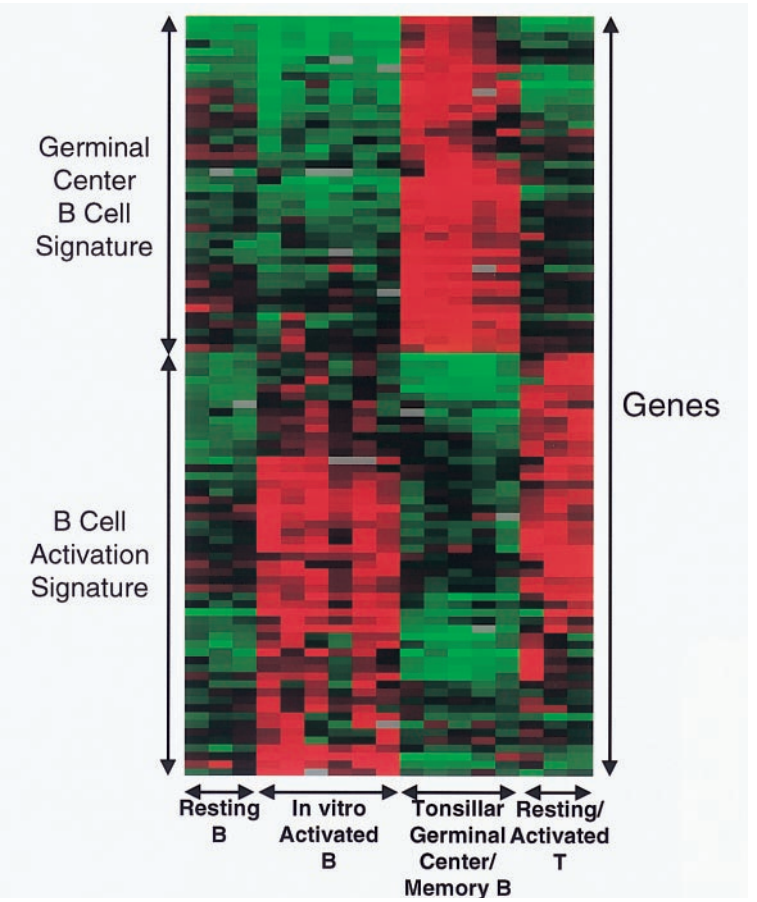

Figure 4 Gene expression signatures in lymphocyte differentiation. Gene expression measurements were taken from four categories of lymphocyte differentiation/activation: resting peripheral blood B cells (both naive and memory), in vitro activated peripheral blood B cells (anti-IgM +/– CD40 ligand +/– IL-4), tonsillar germinal center and memory B cells, and resting or activated (PMA + ionomycin) T cells. The genes were chosen to highlight the difference between germinal center B cells and in vitro activated peripheral blood B cells.

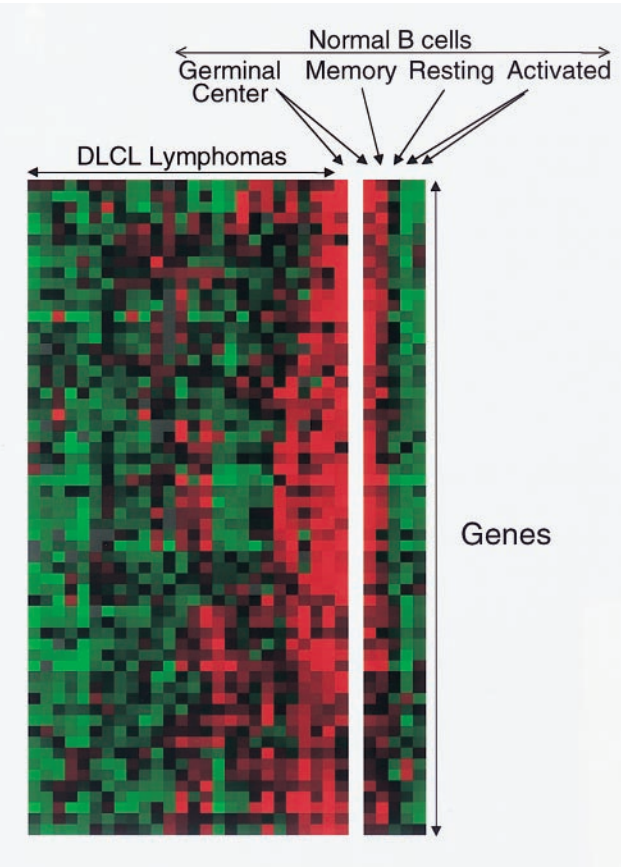

Figure 5 Gene expression in normal and malignant B cells. Gene expression in diffuse large B cell lymphoma (DLCL) lymph node biopsies was compared with gene expression in tonsillar germinal center B cells, tonsillar memory B cells, resting peripheral blood B cells, and peripheral blood B cells activated in vitro with anti-IgM + CD40 ligand + IL-4 for 6 and 24 h. A subset of diffuse large B cell lymphomas resembles normal germinal center B cells.