*Proc. Natl. Acad. Sci. USA*
Vol. 96, pp. 8825–8826, August 1999
From the Academy

***This paper is a summary of a session presented at the first Japanese-American Frontiers of Science symposium, held August 21–23, 1998, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.***

# Functional genomics

STANLEY FIELDS*†, YUJI KOHARA‡, AND DAVID J. LOCKHART§

*Howard Hughes Medical Institute, Departments of Genetics and Medicine, University of Washington, Box 357360, Seattle, WA 98195; ‡Genome Biology Lab, National Institute of Genetics, Mishima 411, Japan; and §Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051

**ABSTRACT Complete genome sequences are providing a framework to allow the investigation of biological processes by the use of comprehensive approaches. Genome analysis also is having a dramatic impact on medicine through its identification of genes and mutations involved in disease and the elucidation of entire microbial gene sets. Studies of the sequences of model organisms, such as that of the nematode worm *Caenorhabditis elegans,* are providing extraordinary insights into development and differentiation that aid the study of these processes in humans. The field of functional genomics seeks to devise and apply technologies that take advantage of the growing body of sequence information to analyze the full complement of genes and proteins encoded by an organism.**

Biology and medicine are undergoing a revolution based on the accelerating determination of DNA sequences, including the complete DNA sequences (or genomes) of a growing number of organisms. Genomics refers to the analysis of these genomes and functional genomics to the component of this field that uses global approaches to understand the functions of genes and proteins. Organisms are complex and genomes can be immense, and thus new and powerful technologies are being developed to analyze large numbers of genes and proteins as a complement to traditional methodologies that study a small number at a time.

Our genes are encoded in our DNA. The DNA is copied into an intermediate, RNA, and the RNA molecules are used to make proteins that carry out the basic processes of life. In general, a gene is the stretch of DNA, including regulatory elements, that specifies a protein. Functional genomics seeks to contribute to the elucidation of some fundamental questions: How does the exact sequence of human DNA differ between individuals? What are the differences that result in disease or predisposition to disease? What is the specific role of each protein synthesized by a bacterial pathogen, by a model organism, e.g., *Escherichia coli,* yeast, the fruitfly, and the nematode, or by a human? How do proteins collaborate to perform the tasks required for life? Because not all genes are active in a given cell at a given time, which genes are used under which circumstances? How does this differential gene expression result in different types of cells and tissues in a multicellular organism?

Increasingly, a first step toward answering these answers is a determination of a genome's complete sequence. Genome sizes vary: thousands of base pairs of DNA in viruses, a few million base pairs (Mb) in bacteria, 13 Mb in the budding yeast *Saccharomyces cerevisiae*, 100 Mb in the nematode *Caenorhabditis elegans*, and 3,000 Mb in the human. More than 50 bacterial genomes are complete or in progress; the yeast genome was completed in 1996, and the nematode sequence is nearly completed. The human genome is projected to be completely sequenced within the next 5 years.

Genome sequences are used first to predict the set of possible proteins, which are compared with all known sequences in central databases. In many instances, for approximately half of the proteins, there is a reasonable match to a previously sequenced protein from another organism. Because the function of many proteins already has been determined in some organism, and similarity in sequence generally reflects similarity in function, these database comparisons can at least partially interpret a substantial fraction of the genome. But how can the function of the many uncharacterized genes be addressed on a genomewide basis?

## *C. elegans* as a Model Organism

The function of a protein often can be determined through the use of simple experimental organisms. These organisms have rapid doubling times and are amenable to uncomplicated genetic analyses yet they undergo complex cellular processes common to mammals. *C. elegans* consists of only about 1,000 somatic cells, but it has the basic body plan of an animal, including muscle, digestive organs, nervous system, epidermis, etc. Moreover, the body is transparent, which allowed the entire pattern of cell divisions from the fertilized egg to the adult, termed the cell lineage, to be described (1). Additionally, this transparency permits the analysis of gene expression at the level of individual cells throughout development. The targeted knockout or inhibition of gene function can be readily accomplished by injection of double-stranded RNA (2). Notably, about half of the genes so far implicated in human disease have a homologue in the worm. A striking example of the significance of this conservation was the ability of a human gene implicated in Alzheimer's disease to complement a worm mutation and the subsequent analysis of mutant human genes through assay of their activities in the worm (3). Finally, the availability of the complete genome sequence of the nematode adds exceptional power to the range of molecular and genetic approaches that can be applied to the study of this organism.

*C. elegans* has been used to address the issue of how a fertilized egg goes on to generate the varied cell types of the mature animal. Of ≈15,000 genes in this organism, 9,000 mRNAs have been identified and 3,000 of these have been analyzed with respect to when and in which cells they are expressed (for example, ref. 4). Various patterns of expression have been observed: some genes are expressed immediately after fertilization and some only after hatching. Genes whose products exist in eggs to very early embryos are of particular interest, because differential localization of these gene products leads to an unequal first division. Thus, the two cells that result when a fertilized egg divides are already different from each other, and these differences can at least partly be attributed to distinct localization patterns of known molecules.

†To whom reprint requests should be addressed. E-mail: fields@u.washington.edu.

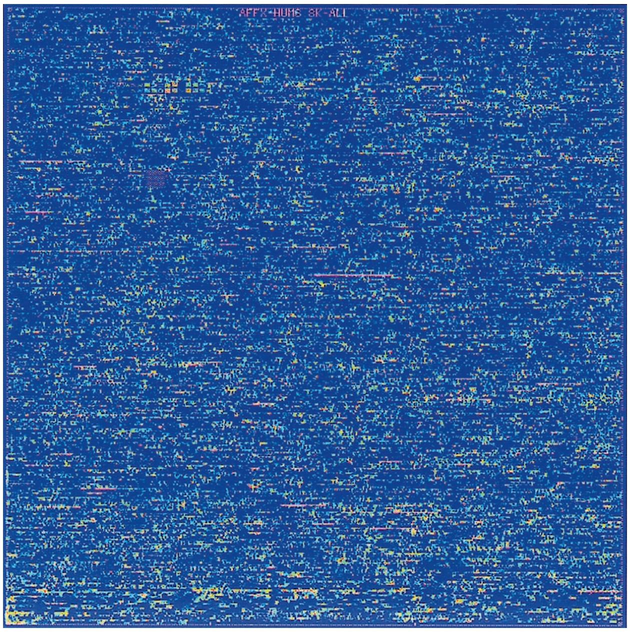

FIG. 1. Fluorescence image of an array that contains more than 280,000 different 25-mer oligonucleotide probes in an area of 1.28 × 1.28 cm. The array contains probe sets for more than 6,800 human genes, and the image was obtained after overnight hybridization of an amplified and labeled human mRNA sample.

By the next division, specific cells have the ability to give and receive instructive signals from neighboring cells.

For the future, efforts will focus on determining the expression pattern for all genes and the distribution of all proteins. Each gene will be knocked out and the phenotype examined of the resulting worms or embryos. Interactions among proteins will be identified. Computer representations of embryogenesis may be generated to simulate the process of development. The goal is to understand the complex processes of differentiation and development: how is the body plan of a multicellular organism specified in the genes of a single cell?

## DNA Chips

Efficient methods for determining the sequence of DNA and measuring the abundance of RNA molecules in cells have been developed that allow the direct use of sequence information to construct DNA arrays. One approach uses high-density two-dimensional arrays of chemically synthesized molecules on a glass surface (DNA chips, or GeneChip arrays) that are designed based on the sequences of the genes to be monitored (5). These oligonucleotides are synthesized *in situ* on a solid surface in a predefined spatial pattern by using a combination of chemistry and photolithographic methods borrowed from the semiconductor industry. Highly parallel chip-based analysis methods are used to scan large regions of DNA from many individuals for mutations or to assay the cellular abundance of thousands of different RNA molecules.

Chip-based quantitative RNA monitoring experiments are being applied to the study of humans, mice, rats, yeast, and bacteria to help understand the genes and pathways involved in biological processes. For example, DNA chips designed for monitoring the expression levels of >6,000 genes in *S. cerevisiae* identified more than 400 RNAs whose levels are significantly modulated as the cell progresses through the cell cycle (6). In combination with the sequence of yeast, this information made possible the finding of potential regulatory elements important for the use of different genes at different stages and allowed new insight into the relationship between the local organization of genes in the genome and their temporal regulation. Also, yeast DNA arrays were used in studies to elucidate the mode of action of new drug candidates on whole cells and to identify effects of drugs that could not be seen by more conventional *in vitro* screening assays (7). Similar methods are being applied to the study of mammals in efforts to understand the genes involved in cancer, other diseases, and addiction to alcohol or cocaine. Chips recently have been made to monitor the expression levels of ≈40,000 human genes and 30,000 mouse genes designed based on information in public sequence databases (Fig. 1).

DNA arrays also are being used to characterize human genetic variation. Extensive stretches of DNA sequence can be screened at once, and more than 4,000 common genetic variations (single-nucleotide polymorphisms, or SNPs) have been found across the human genome (8). These small differences provide markers that can be used in subsequent studies to identify the genes responsible for particular traits and to analyze the sequences of important genes to uncover associations between specific genetic variations and either predisposition to common diseases or the efficacy and safety of therapies.

In the coming years highly parallel methods (oligonucleotide arrays as well as arrays and high-throughput approaches of other types) that allow the collection and analysis of unprecedented amounts of cellular and genetic information may help revolutionize our understanding of biology, the way drugs are developed, and the way diseases are diagnosed, prevented, and treated.

1. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. (1983) *Dev. Biol.* **100,** 64–119.
2. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998) *Nature (London)* **391,** 806–811.
3. Levitan, D., Doyle, T. G., Brousseau, D., Lee, M. K., Thinakaran, G., Slunt, H. H., Sisodia, S. S. & Greenwald, I. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14940–14944.
4. Tabara, H., Motohashi, T. & Kohara, Y. (1996) *Nucleic Acids Res.* **24,** 2119–2124.
5. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. P. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5022–5026.
6. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. & Davis, R. W. (1998) *Mol. Cell* **2,** 65–73.
7. Gray, N. S., Wodicka, L., Thunnissen, A. M., Norman, T. C., Kwon, S., Espinoza, F. H., Morgan, D. O., Barnes, G., LeClerc, S., Meijer, L., *et al.* (1998) *Science* **281,** 533–538.
8. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al.* (1998) *Science* **280,** 1077–1082.