"the function of protein A is to interact with protein B" or "the expression of gene X is correlated (or anticorrelated) with the expression of gene Y." Although conclusions of this nature may have a hollow ring to those trained in a different mode of investigation, the methodologies leading to such conclusions may redeem themselves by providing powerful new perspectives on the holistic operation of biological systems—the "big picture" view. This outlook may even change the way in which we phrase our questions from "what is the function of this protein?" to "what roles does this sequence play in one or more biological processes that are operational under these conditions?"

According to Weiner (*10*), the *Drosophila* genetics of Thomas Hunt Morgan started a whole century of talk of "a gene for _____" (fill in the blank). Actually the question of "what does it mean to ascribe a function to something" has permeated biological thought for centuries. Following the discussion by Sober (*11*, p. 85), functional statements make claims about why an entity is there. What is the function of the heart? To serve as the seat of emotion? To occupy space in the chest and make noise? Or to pump blood? The true function of the heart only became apparent in 1616 when William Harvey considered it as part of a larger system, the circulatory system (although that "making noise" thing did prove to be useful clinically).

Philosophers of science have suggested that biology might benefit from a deemphasis on teleological (functional) concepts and explanations (*11*, p. 83). (Physics abandoned its teleology in favor of causal explanations during the scientific revolution of the 17th century, and it will be interesting to see if modern physicists now moving into biological research will maintain this tradition.) Despite the great heuristic value of teleological explanations and our desire to annotate individual genes with specific functions, this mind-set might actually hinder our ability to fully comprehend the outputs of "functional" genomics methodologies. An unanticipated benefit of these new technologies may indeed be an expansion of our biological epistemology in a new world of teleologically independent, discovery-driven research.

"What is true for *E[scherichia] coli* is true for the elephant," asserted Jacques Monod during the heroic age of molecular biology when it was first imagined that all of the complexities of living systems could be derived from a few basic principles and mechanisms (*12*, p. 592). However, organisms are multiform, intricate, and elaborate physical systems with their operational and regulatory parts assembled by a series of evolutionary contingencies. In the words of Erwin Chargaff, living things display an "immensely diversified phenomenology" that is subject to change in response to innumerable environmental conditions and developmental states. Gene expression profiling "chips" and other types of "functional" genomics technologies will be unveiling many new features or behaviors of genes and protein sequences that will have to be taken into account if we are to fully understand and annotate their activities. But it will not be easy. To paraphrase Hayles (*8*, p. 22), annotations, insofar as they represent informational patterns abstracted from their instantiation in a biological substrate, "can never fully capture the embodied actuality, unless they are as prolix and noisy as the body itself."

The future may lie in a new vision of annotation that supersedes static, "repository biology" with a dynamic "virtual cell" (*13*) in which most properties and behaviors can be quantitatively modeled and dynamically represented in all of their interconnected complexity. Some progress toward such a goal has been made in recent work that elucidated the consequences of altered gene expression in heart failure (*14*) in a way that William Harvey, four centuries ago, could not have imagined but would surely appreciate as the first practitioner of "systems biology."

References and Notes

1. F. Sanger, *Annu. Rev. Biochem.* **57**, 1 (1988).
2. R. F. Doolittle, *J. Mol. Med.* **75**, 239 (1997).
3. F. S. Collins, *Nature Genet.* **9**, 347 (1995).
4. M. S. Boguski, *Trends Biochem. Sci.* **20**, 295 (1995).
5. S. J. Wheelan and M. S. Boguski, *Genome Res.* **8**, 168 (1998).
6. C. J. Jeffery, *Trends Biochem. Sci.* **24**, 8 (1999).
7. D. Fambrough, K. McClure, A. Kazlauskas, E. S. Lander, *Cell* **97**, 727 (1999).
8. N. K. Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Univ. of Chicago Press, Chicago, IL, 1999).
9. P. Hieter and M. Boguski, *Science* **278**, 601 (1997).
10. J. Weiner, *Time, Love, Memory: A Great Biologist and His Quest for the Origins of Behavior* (Knopf, New York, 1999).
11. E. Sober, *Philosophy of Biology* (Westview, Boulder, CO, 1993).
12. H. F. Judson, *The Eighth Day of Creation* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, expanded ed., 1996).
13. S. Shaw, transcript based on a presentation at the British Society for Immunology Annual Meeting, Harrogate, UK, 3 to 4 December 1998 (available at http://gryphon.jr2.ox.ac.uk/Harr98/SHAW/Shaw.htm).
14. R. L. Winslow, J. Rice, S. Jafri, E. Marban, B. O'Rourke, *Circ. Res.* **84**, 571 (1999).

VIEWPOINT

The Mammalian Gene Collection

Robert L. Strausberg,[1] Elise A. Feingold,[2] Richard D. Klausner,[1]* Francis S. Collins,[2]*

The Mammalian Gene Collection (MGC) project is a new effort by the NIH to generate full-length complementary DNA (cDNA) resources. This project will provide publicly accessible resources to the full research community. The MGC project entails the production of libraries, sequencing, and database and repository development, as well as the support of library construction, sequencing, and analytic technologies dedicated to the goal of obtaining a full set of human and other mammalian full-length (open reading frame) sequences and clones of expressed genes.

It is not yet routine to identify all possible mammalian genomic regions that are transcribed. This is in part because much of the DNA does not encode gene transcripts, and the rules of transcription and transcript processing are not yet fully understood. A particularly powerful material for studying gene expression, therefore, is cDNA, which is DNA reverse-transcribed from a complete RNA molecule that represents the full-length, expressed gene transcript. Indeed, one of the most effective and widespread manifestations of the genomics revolution has been the ready public access to cDNA libraries, sequences, and clones. The value of having such resources has been recognized since the early planning phases of the Human Genome Project (HGP) (*1*). However, it was also clear at that time that the development of an annotated and complete catalog of full-length human cDNAs (with sizes ranging from <1 to >10 kb for the array of human genes) would require advances in methodology and strategy, as well as improved reagents. Moreover, cost-effective DNA sequencing of tens to hundreds of thousands of full-length cDNAs would require technological advances not available at the start of the HGP.

In 1991, Venter and colleagues (*2*) developed a conceptually different approach to the establishment of systematic cDNA resources, termed the expressed sequence tag (EST) strategy. Although the sequence tags covered only a segment of the gene, and the clones were generally not full length, their utility for gene iden-

[1]National Cancer Institute, [2]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.

*To whom correspondence should be addressed.

tification was immediately recognized. By 1993, vigorous EST sequencing efforts were under way in the private sector, but in general these databases were unavailable to academic researchers. Thus, the stage was set for the implementation of public EST sequencing projects that have now contributed over 1.5 million human ESTs (and additional ESTs for many other organisms) to a GenBank division specifically devoted to managing EST sequences (dbEST). These ESTs have been produced in many laboratories in a worldwide effort, with major contributions being made through the efforts of the Merck Gene Index (*3*), the Cancer Genome Anatomy Project (*4*) (http://www.ncbi.nlm.nih.gov/ncicgap/), The Institute for Genomic Research (*5*), and the Howard Hughes Medical Institute (*6*).

An important development that was also key for the widespread use of ESTs was the formation of the IMAGE consortium (*7*) (http://bbrp.llnl.gov/bbrp/image/), led by the Lawrence Livermore National Laboratory, to ensure that collections of clones as well as sequences would be accessible by the biomedical research community. Through the efforts of this consortium, clones from which the EST sequences derive are available at modest cost through a public-private–sector partnership.

One of the greatest challenges for EST databases and data users is to understand the relationships of these relatively short sequences to each other and to other genes. Toward that end, the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) has developed algorithms to assign ESTs with sequence similarity to clusters, forming the basis of the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/). Recently, more than 30,000 of these UniGenes were systematically mapped (*8*), allowing ready integration of the EST database into positional cloning projects, often trimming months or even years off of the search for a disease gene.

Although the public EST resource has been a remarkably productive interim solution, there are limitations in accuracy resulting from the single-pass sequencing that has been used to identify ESTs and the possibility that the UniGene clustering may mix closely related genes. There are also many applications for which partial sequences are not adequate. For example, accurately predicting the function or structure of a gene product or isolating the protein product, requires a full-length sequence.

Full-length sequences for only about 6000 of the 80,000 to 100,000 human genes are in the current database. Furthermore, a physical clone of the full-length sequence is needed for additional experimentation and the actual clones have not been accessible in the form of an organized public collection.

However, cDNA technology has advanced substantially during the past few years. For example, reagents such as enzymes with improved fidelity and processivity have been engineered and applied to the production of cDNA libraries, as in the *Drosophila* genome project (*9*) (http://www.fruitfly.org/), so that libraries with a majority of full-length cDNAs in the size range up to at least 3 kb are being generated routinely. Moreover, initial application of those technologies to RNA derived from human cell lines has generated very similar results (*10*).

These advances, in combination with dramatic improvements in DNA sequencing technology, lead us to believe that a highly effective pipeline can now be established to obtain representative full-length coding sequences and clones for human genes encoding transcripts at least up to the 3- to 4-kb size range. Furthermore, methodological improvements on a number of fronts have created the expectation that a very high fraction of full-length cDNAs will be identifiable in the near future. For example, size selection methods for transcripts and cDNAs are currently being developed for the isolation of large cDNAs (*11*). Additional advances in cDNA library methodology, including procedures to remove common or already-detected sequences (*12*), are now refined to the point where they can start to be applied to isolation of full-length cDNAs derived from rare transcripts. Additional specialized technologies, such as approaches that select for the 5′ transcript cap (*13–15*) are also being successfully applied to cDNA library construction.

Therefore, we believe that the time is right to initiate a program, MGC, designed to provide to the research community representative sequences and clones for all human and mouse genes, and ultimately those of other mammalian species. This program results from discussions at several National Institutes of Health (NIH) planning meetings, including those leading up to the most recent 5-year plan of the HGP (*16*).

The MGC program is sponsored by 16 NIH institutes and the NLM (*17*) and will be led by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The program includes components for (i) production, analysis, and distribution of libraries, clones, and sequences; and (ii) technology development. In its first year, approximately $10 million has been set aside for this effort, with an expectation that this amount will grow in future years.

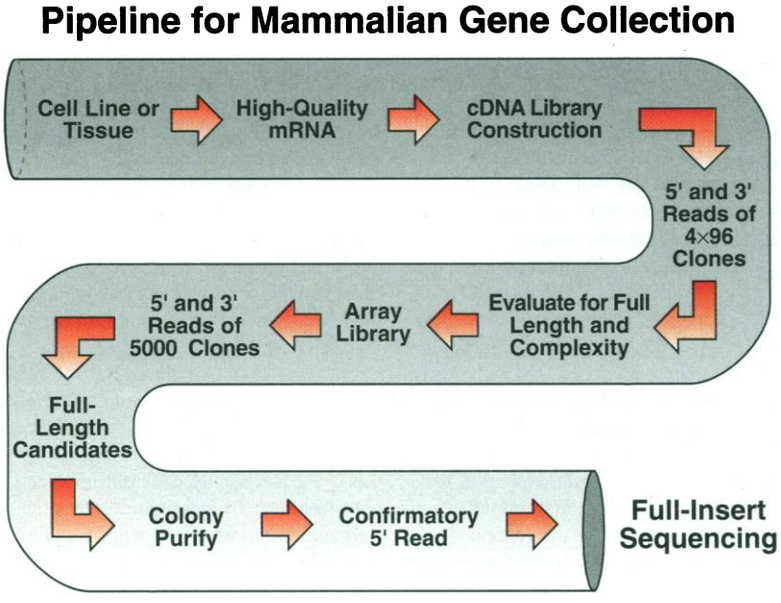

Fig. 1. Pipeline for the MGC.

Library and Sequence Production Pipeline

Complementary DNA libraries and clones. We believe that the goal of producing high-quality libraries with good representation of full-length cDNAs (greater than 50% of clones with full open reading frames) can initially be achieved through the application of well-established cDNA methodologies and the use of RNA derived from human cells in culture, both primary and immortalized. Our initial approach is to establish a pipeline for cloning and sequencing transcripts up to 3 to 4 kb in size, and also to rigorously test methods such as size selection to assess their effectiveness for isolation of longer full-length transcripts.

Individual steps in the MGC project pipeline are outlined in Fig. 1. After initial library evaluation, all arrayed clones will be accessible through the IMAGE consortium, and the 5′ and 3′ EST sequences will be immediately deposited in GenBank. Thus, even before the full-length sequencing has been performed, clones and tag sequences will be available to the community.

Full-length sequencing pipeline. The sequencing goal is to establish a pipeline capable of determining at least 20,000 highly accurate full-length sequences per year. In the first-year feasibility phase, the goal is to sequence approximately 5000 to 7000 full-length cDNAs with insert sizes of up to 3 to 4 kb. It is anticipated that multiple laboratories will participate in this effort, and various sequencing strategies (for example, based on transposons, concatenation, primer-walking, and traditional shotgun sequencing) will be explored. In addition, we will continually assess the productivity of the MGC sequencing with respect to cost-effectiveness, throughput, and sequence quality. With respect to sequence quality, the standards established for the HGP (http://www.nhgri.nih.gov:80/Grant_info/Funding/Statements/RFA/quality_standard.html) will be applied to the MGC. Currently, the standard is that finished sequences are to be at least 99.99% accurate. The sequence validation process includes sequencing of selected clones by multiple laboratories. The MGC is initially emphasizing sequencing of full-length cDNA from human libraries. However, mouse libraries will also be prepared early in the project, and 5′ and 3′ ESTs and the clones will be accessible. As sequencing capacity increases, the mouse clones will also be subjected to full-length sequence analysis.

Informatics. Key to the success of the project will be the establishment of robust informatics tools. Because the full coding sequence for most of the human genes is not known, one of the key initial challenges will be the development and refinement of algorithms for the selection of clones potentially encoding complete sequences. These algorithms will be continually refined as the project progresses and we learn more about human DNA sequences in the form of cDNAs and genomic DNA.

Progress reports, from library preparation to complete sequence analysis, will be accessible through the MGC Web site (http://www.ncbi.nlm.nih.gov/MGC) and through the databases maintained by the NCBI. Annotation of the sequences and clones (for example, homology with other genes, gene families, tissue expression patterns, and polymorphism identification) will be facilitated through analysis tools developed for this and other projects by NCBI, NCI, and NHGRI.

Technology development. Although we are confident that technology is available to initiate production of full-length clones and sequences, completion of this project will require the development of new technologies designed to identify (i) rare transcripts, (ii) very long transcripts, and (iii) transcripts with especially challenging structures. In addition, new methods are needed to assemble high-quality libraries directly from human (and other) tissues, especially those that are available in small quantity and those from which RNA extraction is challenging. Therefore, we will support the development of libraries designed to be specifically enriched for full-length cDNAs, such as through size selection, for the more difficult human cDNAs.

Project management. Overall direction of the MGC is the responsibility of the directors of the lead institutes: NHGRI and NCI. An External Steering Committee (ESC) composed of members of the scientific community external to NIH (*18*) will provide oversight for all aspects of the program, including the current projects, future plans, and production efforts.

Oversight of the library, clone, and sequence production pipeline is the responsibility of an Implementation Working Group (IWG) composed of the scientists building these resources (*19*). The IWG, whose membership includes intramural and extramural scientists, is responsible for ensuring that this pipeline meets or exceeds the stated goals.

A competitive contract solicitation will be issued for full-length cDNA sequencing. This pipeline will be funded through a flexible contract mechanism, overseen by a team of program directors (*20*) from the participating NIH institutes [the Inter-Institute Coordinating Committee (IICC)]. This mechanism was chosen to ensure that the project can continually take advantage of new scientific opportunities.

The technology development efforts will be funded primarily through research grant mechanisms overseen by the IICC. Already one solicitation has been issued by this group (http://www.nih.gov/grants/guide/rfa-files/RFA-CA-99-005.html), and it is expected that there will be additional opportunities to participate in the technology development efforts in support of the project.

Overall Vision

A major thrust of contemporary biological and biomedical research is to determine and understand the genetic contribution to disease and other biological phenomena. Complete catalogs of genes (both sequences and clones) will be essential for thorough genetic analysis. The MGC program is designed to generate these critical resources, which will be widely used and of inestimable value to biological researchers. Because the value of the information, clones, and analytical tools produced will be most fully realized only if they are broadly available to the academic and industrial communities, it is necessary and appropriate that they be developed by a public-sector effort. Beyond the identification of the coding sequences themselves, these resources will be the basis for experiments within the academic and industrial communities, targeted at understanding gene expression at many levels, and the nature and properties of the gene products. Thus, it will be important to develop and use vector systems that will be well-suited to efficient transfer of the coding sequences to various mammalian and nonmammalian expression vectors.

However, it is also important to realize that, while the sequences and clones for a representative cDNA for each human gene will be of immediate utility to the community, they will represent only a piece of the puzzle. The transcript catalog will actually not be complete for some time because the development of strategies for identifying and cataloging alternatively processed transcripts will remain a great challenge.

We welcome the participation of other groups, including the international community, to work in concert to meet this ambitious goal. Through the MGC Web site, the ESC, and other venues, the program will seek input about ways to enhance the resources produced, including interfacing with future high- throughput genomic, proteomic, and other biological research projects.

References and Notes

1. *Report of the Committee on Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
2. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
3. A. R. Williamson, *Drug Discov. Today* **4**, 115 (1999).
4. R. L. Strausberg, C. A. Dahl, R. D. Klausner, *Nature Genet.* **15**, 415 (1997).
5. M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995).
6. M. Marra *et al.*, *Nature Genet.* **21**, 191 (1999).
7. G. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
8. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
9. G. M. Rubin, *Trends Genet.* **14**, 340 (1998).
10. _____, D. Harvey, L. Hong, personal communication.
11. M. B. Soares, unpublished data.
12. M. F. Bonaldo, G. Lennon, M. B. Soares, *Genome Res.* **6**, 791 (1996).
13. M. Seki, P. Carninci, Y. Nishiyama, Y. Hayashizaki, K. Shinozaki, *Plant J.* **15**, 707 (1998).
14. P. Carninci *et al.*, *DNA Res.* **4**, 61 (1997).
15. I. Edery, L. L. Chu, N. Sonenberg, J. Pelletier, *Mol. Cell Biol.* **15**, 3363 (1995).
16. F. S. Collins *et al.*, *Science* **282**, 682 (1998)
17. The participating institutes include NCI; NHGRI; National Institute of Neurological Disorders and Stroke; National Institute on Alcohol Abuse and Alcoholism; National Institute on Aging; National Institute of Arthritis and Musculoskeletal and Skin Diseases; National Institute of Child Health and Human Development; National Institute of Mental Health; National Eye Institute; National Institute of Allergy and Infectious Diseases; National Institute on Deafness and Other Communication Disorders; National Institute on Drug Abuse; National Institute of Dental and Craniofacial Research; National Institute of Diabetes and Digestive and Kidney Diseases; National Institute of Environmental Health Sciences; National Heart, Lung and Blood Institute; and NLM.
18. Initial members of the ESC, to be chaired by Barbara Wold, include Connie Cepko, Ronald Davis, Geoff Duyk, Edward Harlow, Stewart Scherer, Philip Sharp, and Lincoln Stein.
19. Members of the IWG, chaired by Bob Strausberg, include James Battey, Narayan Bhat, Gerard Bouffard, Michael Brownstein, Francis Collins, Jeffrey Derge, Michael Emmert-Buck, Elise Feingold, Weiniu Gan, Eric Green, Lynette Grouse, Mark Guyer, Ralph Hopkins, Richard Klausner, David Lipman, Louis Staudt, Carolyn Tolstoshev, Jeff Touchman, Ted Usdin, and Lukas Wagner.
20. Members of the IICC, chaired by Elise Feingold and Bob Strausberg, include Susan Banks-Schlegel, Peter Clepper, Jennifer Couch, Hemin Chin, Maria Giovanni, Mark Guyer, Robert Karp, Gabrielle Leblanc, Anne McCormick, Jonathan Pollack, Vicki Seyfert, William Sharrock, Judy Small, Rochelle Small, Philip Smith, Jose Velazquez, and Michael Whalin.

21. We thank M. Guyer for valuable comments during the preparation of this manuscript.

References and Notes

[2] **Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project**

Mark D. Adams; Jenny M. Kelley; Jeannine D. Gocayne; Mark Dubnick; Mihael H. Polymeropoulos; Hong Xiao; Carl R. Merril; Andrew Wu; Bjorn Olde; Ruben F. Moreno; Anthony R. Kerlavage; W. Richard McCombie; J. Craig Venter

Science, New Series, Vol. 252, No. 5013. (Jun. 21, 1991), pp. 1651-1656.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819910621%293%3A252%3A5013%3C1651%3ACDSEST%3E2.0.CO%3B2-6

[8] **A Physical Map of 30,000 Human Genes**

P. Deloukas; G. D. Schuler; G. Gyapay; E. M. Beasley; C. Soderlund; P. Rodriguez-Tomé; L. Hui; T. C. Matise; K. B. McKusick; J. S. Beckmann; S. Bentolila; M.-T. Bihoreau; B. B. Birren; J. Browne; A. Butler; A. B. Castle; N. Chiannilkulchai; C. Clee; P. J. R. Day; A. Dehejia; T. Dibling; N. Drouot; S. Duprat; C. Fizames; S. Fox; S. Gelling; L. Green; P. Harrison; R. Hocking; E. Holloway; S. Hunt; S. Keil; P. Lijnzaad; C. Louis-Dit-Sully; J. Ma; A. Mendis; J. Miller; J. Morissette; D. Muselet; H. C. Nusbaum; A. Peck; S. Rozen; D. Simon; D. K. Slonim; R. Staples; L. D. Stein; E. A. Stewart; M. A. Suchard; T. Thangarajah; N. Vega-Czarny; C. Webber; X. Wu; J. Hudson; C. Auffray; N. Nomura; J. M. Sikela; M. H. Polymeropoulous; M. R. James; E. S. Lander; T. J. Hudson; R. M. Myers; D. R. Cox; J. Weissenbach; M. S. Boguski; D. R. Bentley

Science, New Series, Vol. 282, No. 5389. (Oct. 23, 1998), pp. 744-746.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819981023%293%3A282%3A5389%3C744%3AAPMO3H%3E2.0.CO%3B2-T

LINKED CITATIONS

[16] **New Goals for the U.S. Human Genome Project: 1998-2003**

Francis S. Collins; Ari Patrinos; Elke Jordan; Aravinda Chakravarti; Raymond Gesteland; LeRoy Walters; Members of the DOE and NIH Planning Groups

Science, New Series, Vol. 282, No. 5389. (Oct. 23, 1998), pp. 682-689.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819981023%293%3A282%3A5389%3C682%3ANGFTUH%3E2.0.CO%3B2-%23