out of Africa (*5*, *19*, *20*). The possibility that human history has been characterized by genetically relatively homogeneous groups ("races"), distinguished by major biological differences, is not consistent with genetic evidence.

How, then, does genetics explain the stereotypic features of "races": skin color, hair color and texture, and facial traits? These traits are quite literally superficial, in that they affect exposed surfaces of the body. It is reasonable to suggest that variation in these traits may reflect differential selection by climate in various parts of the world. Recent analysis of the melanocortin-stimulating hormone receptor gene (MC1R) suggests that various alleles of this single locus may underlie much of observed human variation in skin and hair color (*21*, *22*). This variation is largely due to varied amounts of eumelanin (brown and black melanins) and phaeomelanin (red and yellow melanins) produced by melanocytes. Eumelanin protects against ultraviolet (UV) radiation, whereas phaeomelanin may contribute to skin damage, including melanoma, induced by UV. The balance of melanins is regulated by melanocyte-stimulating hormone, which acts through its receptor. Amino acid sequence variants occur at multiple sites in the second transmembrane domain, the first extracellular domain, and the seventh transmembrane domain of the MC1R protein. Variation at these sites was found in more than 80% of individuals with red hair and fair skin that burns rather than tans, but in less than 4% of British or Irish individuals with skin that tans without burning, and in no African individuals. Among Asians, still other amino acid substitutions in MC1R are common. Nucleotide diversity at MC1R is several times higher than the average nucleotide diversity in human populations. High nucleotide diversity, coupled with common variation at nonsynonymous sites, suggest that MC1R variation is an adaptive response to selection for different alleles in different environments, possibly to differences in day length and hence available sunlight at different latitudes. If true, variation at this locus, which encodes evolutionarily important but superficial traits, has been the cause of enormous suffering. Variation in other traits popularly used to identify "races" is likely to be due to similarly straightforward mechanisms, involving limited numbers of genes with very specific physiological effects. Of course, prejudice does not require a rational basis, let alone an evolutionary one, but the myth of major genetic differences across "races" is nonetheless worth dismissing with genetic evidence.

References and Notes

1. C. P. Snow, *The Two Cultures and the Scientific Revolution, The Rede Lecture* (Cambridge Univ. Press, New York, 1959).
2. H. Greely, *Annu. Rev. Anthropol.* **27**, 473 (1998). _____, *Houston Law Rev.* **33**, 1431 (1997).
3. A. C. Wilson and R. L. Cann, *Sci. Am.* **266**, 68 (April 1992).
4. L. Jin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3796 (1999).
5. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1994); A. Piazza *et al.*, *ibid.* **92**, 5836 (1995).
6. M. T. Seielstad, E. Minch, L. L. Cavalli-Sforza, *Nature Genet.* **20**, 278 (1998).
7. G. Barbujani *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4516 (1997).
8. L. Excoffier, P. E. Smouse, J. M. Quattro, *Genetics* **131**, 479 (1992).
9. D. Comas *et al.*, *Am J. Hum. Genet.* **63**, 1824 (1998).
10. A. Perez-Lezaun *et al.*, *ibid.* **65**, 208 (1999).
11. E. S. Poloni *et al.*, *ibid.* **61**, 1015 (1997).
12. M. Krings *et al.*, *ibid.* **64**, 1166 (1999).
13. Second International Symposium in Memory of Richard M. Goodman, National Foundation for Jewish Diseases and Tel Aviv University, Israel, June 1999.
14. N. Risch *et al.*, *Nature Genet.* **9**, 152 (1995).
15. A. G. Motulsky, *ibid.*, p. 99; A. de la Chapelle and F. A. Wright, *Proc. Natl. Acad. Sci. U.S.A.*, **21**, 12416 (1998).
16. J. Martinez-Laso *et al.*, *Tissue Antigens* **47**, 63 (1996).
17. A. B. Spurdle and T. Jenkins, *Am. J. Hum. Genet.* **59**, 1126 (1996).
18. K. Skorecki *et al.*, *Nature* **385**, 32 (1997).
19. L. B. Jorde *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3100 (1997).
20. H. Kaessmann *et al.*, *Nature Genet.* **22**, 78 (1999).
21. P. Valverde *et al.*, *ibid.* **11**, 328 (1995).
22. B. K. Rana *et al.*, *Genetics* **151**, 1547 (1999); H. B. Schioth *et al.*, *Biochem. Biophys. Res. Commun.* **260**, 488 (1999).

VIEWPOINT

Biosequence Exegesis

Mark S. Boguski

Annotation of large-scale gene sequence data will benefit from comprehensive and consistent application of well-documented, standard analysis methods and from progressive and vigilant efforts to ensure quality and utility and to keep the annotation up to date. However, it is imperative to learn how to apply information derived from functional genomics and proteomics technologies to conceptualize and explain the behaviors of biological systems. Quantitative and dynamical models of systems behaviors will supersede the limited and static forms of single-gene annotation that are now the norm. Molecular biological epistemology will increasingly encompass both teleological and causal explanations.

The sequences of proteins (and a few nucleic acids) had slowly been accumulating in the literature since Sanger's seminal work on the structure of insulin in the 1950s. However, the real catalyst for the expanding depth and scope of our knowledge of macromolecular sequences was the development of rapid DNA-sequencing technologies in 1975 (*1*). At that point, the direct determination of protein sequences began to be supplanted by conceptual translations (using the genetic code) of DNA sequences into their cognate gene products. For the next 15 years, sequence data continued to be so novel and revealing that each new example was reported in a peer-reviewed publication accompanied by a richly detailed interpretation derived from the biological hypothesis or context that led to the cloning of a particular gene in the first place. [In one case, which seems remarkable in hindsight, the cloning of the first mammalian messenger RNA (mRNA) in 1977 merited three separate publications in the journal *Cell*, one describing the 5′ untranslated region, another the 3′ untranslated region, and another the coding sequence of rabbit β globin mRNA.] Often there were many follow-up publications about particularly important genes or proteins that resulted in an even richer and more complete "annotation" (in the literature) of the biological processes in which a sequence was involved. This was the era of "functional cloning," the hallmark of which was "function first, sequence later." In that era, database similarity (or "homology") searching was a hit-or-miss activity, characterized by frequent misses and rare hits, the latter usually being extremely rewarding and informative and engendering the palpable excitement of ground-breaking discoveries (*2*). On the order of 10,000 mammalian genes have been functionally cloned over the past two decades.

In the late 1980s and early 1990s, the method of functional cloning was enriched by a powerful new approach, positional cloning (*3*). In this approach, one begins with a phenotype, proceeds through genetic linkage and physical

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

mapping, and then obtains the sequence underlying the phenotype solely by virtue of its location (position) in the genome. (Studies to elucidate the biochemistry or pathophysiology of the gene come afterward.) Again, the manner in which these cloning experiments take place produces informative annotation of the sequences in the literature and archival sequence databases. On the order of several hundred mammalian genes have been isolated with this approach or the related approach of "positional candidate" cloning.

With the genesis of the Human Genome Project and the development of associated strategies for rapid gene discovery made possible by the automation of sequencing technology, the tide began to turn in the early 1990s (*4*). Studies started to appear in the literature that reported the cloning of hundreds to thousands of new genes in a single paper. The only "annotation" for these new sequences consisted of computational similarity studies that segregated the new genes into categories of completely novel sequences or sequences that were identical or related (homologous) to known, functionally cloned genes already in the database. As tens of thousands, then hundreds of thousands, then millions of expressed gene fragments from a variety of organisms were produced over the next 8 years, the frequency of "hits" in database searches increased dramatically. However, the average informativeness of these hits was substantially decreased, and, consequently, annotation-by-homology began to degrade in quality, if not quantity. Adding to these circumstances was the increasing availability of complete genome sequences for both single-cell and metazoan organisms containing thousands to tens of thousands of predicted genes, most of which were accompanied by no independent evidence of functionality or even expression. The sequence databases began to accumulate vast numbers of entries merely labeled as "ORF" (open reading frame, conceptual translation) or "hypothetical protein," oftentimes further annotated by an automated, computational assessment of similarity to some other known gene or gene product. Although this wealth of new data has been of immense value to biology and has created a thriving field of comparative genomics, it has also created daunting challenges for both experimental and computational biologists.

There was a turning point in 1995, when the number of genes in the database began to exceed the number of papers in the literature (Fig. 1). The resultant gap between sequences and classical approaches to discover and annotate their functions is diverging to such an extent that it is now widely accepted that we cannot realistically expect traditional experimental methods to scale up and have a substantial impact on bridging this gap.

Faced with this dilemma, many researchers resorted to bioinformatics approaches alone in an attempt to add "value" to this data. Unfortunately, in one sense, computational sequence analysis, although one of the most powerful and important tools in molecular biology for the past 17 years, is now approaching a wall in terms of its ability to reveal detailed and reliable inferences about the biological implications of sequence data. This situation has little to do with the speed and sensitivity of search and alignment programs or accessibility issues relating to the World Wide Web. Rather, it has to do with the accuracy and organization of the data and the reliability, consistency, and up-to-dateness of the annotation (experimental or computational) associated with it. At present, it is prudent to view most annotations as hypotheses with some probability of being incomplete, misleading, or even incorrect.

Even for functionally cloned genes, published interpretations of sequence data today are often quite speculative, unsupported by adequate documentation of statistical significance, or lacking details necessary for reproducibility. In any event, such annotations are almost instantly outdated by the tsunami of new data that washes over the community daily. Thus, annotations of uncertain or time-limited value become enshrined in secondary archives and are propagated by transitivity whereby new sequences are compared against old and annotation is transferred as a chain of weak inferences to many other (usually "hypothetical") proteins. Gene prediction technology is still not reliable enough (in the absence of corresponding full-length mRNA sequences or comprehensive comparative genomic data from another species, for example, human and mouse) to produce accurate models of transcription units. In principle, all of these problems are soluble given enough time and resources to improve and expand the infrastructure of experimental validations and to provide periodic reassessments and revisions of computed features with the new information and improved methods (*5*).

Apart from these issues, which are essentially issues of progressive and vigilant quality assurance, there are more fundamental and difficult problems to face. One of these is the fact that a single gene or protein may have multiple forms and functions that are context-dependent and that can never be fully understood by sequence analysis alone. [The phenomenon of "one protein—many functions" has been referred to as "moonlighting" in a recent publication (*6*).] Furthermore, one of the important lessons from gene knockout experiments is that some genes may appear to have no recognizable phenotype or function at all. There are several explanations for this. One may simply not be assaying for phenotype or function under the precise environmental or developmental circumstances under which it would be revealed. Additionally (and particularly in more complex organisms), there may be so many "fail safe" systems, redundancies, or convergent pathways (*7*) in the genome blueprint that surrogates are often able to fill in for the missing or damaged players. Also, some genes may be skeuomorphs (a term from archaeology), that is, an adaptation that is no longer functional but that was functional at an earlier time (*8*, p. 17). Other genic entities may be pseudogenes or even protogenes that may be awaiting an evolutionary opportunity to be called into new service.

Besides bioinformatics, another response to the challenge of massive amounts of sequence data was the development of "functional genomics" and "proteomics" technologies in the mid-1990s. Functional genomics refers to the development and application of global (genome-wide or system-wide) experimental approaches to assess gene function or activity by making use of the information and reagents provided by genome mapping and sequencing projects (*9*). It is characterized by high-throughput or large-scale experimental methodologies, and the fundamental strategy is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion. In contrast to the intimate details of function that traditional biological disciplines provide, functional genomics and proteomics produce much broader but shallower information about large numbers of genes and proteins. These approaches often result in conclusions such as

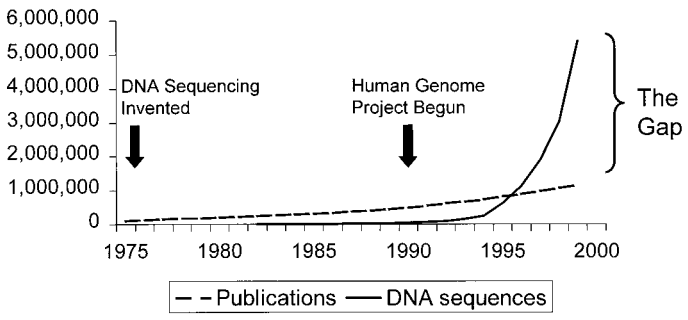

Fig. 1. Cumulative increases of published articles in molecular biology and genetics and DNA sequence records in GenBank. Accumulations of articles in the "G5" subset of MEDLINE are charted alongside GenBank records over the same time period. The former data were obtained with the help of R. M. Woodsmall, and values for 1999 were extrapolated based on the first 9 months of the year. Data on GenBank growth are available in GenBank Release Notes at www.ncbi.nlm.nih.gov. Growth since release 113.0 (15 August 1999) to the end of the year was extrapolated by M. Cavanaugh. No attempt was made to eliminate redundancies in the information content of either GenBank records or articles in the literature.

"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.