model systems, such as the rat. Also required are many other reagents, resources, and technologies, such as full-length cDNAs; validated expression and protein arrays; additional recombination systems; and improved methodologies for tissue-specific expression, overexpression, or inducible expression of gene products.

Conclusions

The availability of the mouse genome sequence and the development of high-throughput, gene-based and phenotype-based mutagenesis paradigms constitute a turning point in biomedical research. We now set challenging goals for the next 10 years. Achieving these goals will require the biomedical research community to improve efficiencies, to reduce costs, and to coordinate international expertise and resources. The impact of these activities will be enormous—deeper insights into functions of genes individually and collectively; fundamental biological and disease processes; and ultimately improved diagnosis, prevention, and treatment of birth defects and adult diseases.

References and Notes

1. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
3. A mouse genome sequence is currently available to subscribers at www.celera.com/products/products.cfm. Information about publicly funded programs sequencing the mouse genome can be found at www.ncbi.nlm.nih.gov/Traces/trace.cgi
4. J. H. Nadeau *Nature Rev. Genet.*, in press.
5. M. J. Justice, *Nature Rev. Genet.* **1**, 109 (2000).
6. B. Zheng *et al.*, *Mol. Cell. Biol.* **20**, 648 (2000).
7. D. Metzger, R. Feil, *Curr. Opin. Biotechnol.* **10**, 470 (1999).
8. M. H. Hrabe de Angelis *et al.*, *Nature Genet.* **25**, 444 (2000).
9. P. M. Nolan *et al.*, *Nature Genet.* **25**, 440 (2000).
10. M. V. Wiles *et al.*, *Nature Genet.* **24**, 13 (2000).
11. www.imgs.org
12. www.nano.gov
13. www.nih.gov/science/models/mouse
14. www.informatics.jax.org/mgihome/nomen/allmut_form.shtml
15. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
16. K. P. Paigen, J. T. Eppig, *Mamm. Genome* **11**, 715 (2000).
17. J. H. Nadeau *et al. Nature Genet.* **24**, 221 (2000).
18. H. Su *et al. Nature Genet.* **24**, 92 (2000).
19. B. Zheng *et al.*, *Nature Genet.* **22**, 375 (1999).
20. We thank G. Duyk and K. Moore for their many thoughtful insights and suggestions during the preparation of this plan.
21. Members of the IMMC: R. Balling (President, International Mammalian Genome Society), German Center for Biotechnology, D-38124 Braunschwieig, Germany. G. Barsh, Departments of Pediatrics and Genetics, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. D. Beier, Genetics Division, Brigham and Women's Hospital, Boston, MA 02115, USA. S. D. M. Brown, MRC Mammalian Genetics Unit, UK Mouse Genome Centre, Harwell, OX11 ORD, UK. M. Bucan, Center for Neurobiology and Behavior, Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, USA. S. Camper (Secretariat, International Mammalian Genome Society), Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA. G. Carlson, McLaughlin Research Institute, Great Falls, MN 59405, USA. N. Copeland, National Cancer Institute–Frederick, Frederick, MD 21702, USA. J. Eppig, Jackson Laboratory, Bar Harbor, ME 04609, USA. C. Fletcher, The Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA. W. N. Frankel, Jackson Laboratory, Bar Harbor, ME 04609, USA. D. Ganten, Max Delbruck Center for Molecular Medicine, 13092 Berlin-Buch, Germany. D. Goldowitz, Department of Anatomy and Neurobiology, University of Tennessee, Memphis, TN 38163, USA. C. Goodnow, Medical Genome Centre, John Curtin School of Medical Research, The Australian National University, Canberra, ACT 2601, Australia. J.-L. Guenet (Secretariat, International Mammalian Genome Society), Unité de Génétique des Mammifères, Institut Pasteur, 75724 Paris Cedex 15, Paris, France. G. Hicks, MICB Center for Mammalian Functional Genomics, Manitoba Institute of Cell Biology, University of Manitoba, Winnipeg, Manitoba R3F 0V9, Canada. M. Hrabe de Angelis, Genome Project Group, GSF National Research Center for Environment and Health Institute of Experimental Genetics, D-85764 Neuherberg, Germany. I. Jackson (Vice President, International Mammalian Genome Society), MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, Scotland. H. J. Jacob, Human and Molecular Genetics Center, Milwaukee, WI 53226, USA. N. Jenkins, National Cancer Institute–Frederick, Frederick, MD 21702, USA. D. Johnson, Mammalian Genetics and Genomics, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. M. Justice (Secretariat, International Mammalian Genome Society), Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. S. Kay, Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA. D. Kingsley, Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA. H. Lehrach, Max-Planck Institute of Molecular Genetics–Berlin, D-14195 Berlin, Germany. T. Magnuson (Secretariat, International Mammalian Genome Society), Department of Genetics, University of North Carolina–Chapel Hill, Chapel Hill, NC 27599, USA. M. Meisler (Past President, International Mammalian Genome Society), Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA. J. H. Nadeau (corresponding author). A. Poustka, Division of Molecular Genome Analysis, German Cancer Research Center, 69120 Heidelberg, Germany. E. M. Rinchik, Mouse Genetics and Mutagenesis, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. J. Rossant, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada. L. B. Russell, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. J. Schimenti, The Jackson Laboratory, Bar Harbor, ME 04609, USA. T. Shiroishi (Secretariat, International Mammalian Genome Society), Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka-ken 411-8540, Japan. W. C. Skarnes, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA. P. Soriano, Program in Developmental Biology, Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. W. Stanford, Lunenfeld Gene Trap Laboratory, Program in Development and Fetal Health, Samuel Lunenfeld Research Institute, Toronto, Ontario M5G 1X5, Canada. J. S. Takahashi, Howard Hughes Medical Institute, Northwestern University, Evanston, IL 60208, USA. W. Wurst, Molecular Neurogenetics, Max-Planck Institute of Psychiatry, 80804 Munich, Germany. A. Zimmer, Department of Psychiatry, University of Bonn, Bonn 53113, Germany.

FUTURE DIRECTIONS: GENE NUMBER

What If There Are Only 30,000 Human Genes?

Jean-Michel Claverie

The confirmation that there might be fewer than 30,000 protein-coding genes in the human genome is one of the key results of the monumental work presented in this issue of *Science* by Venter *et al.* (*1*). That a mere one-third increase in gene numbers could be enough to progress from a rather unsophisticated nematode [*Caenorhabditis elegans*, with about 20,000 genes (*2*)] to humans (and other mammals) is certainly quite provocative and will undoubtedly trigger scientific, philosophical, ethical, and religious questions throughout the beginning of this new century. By the same token, humans appear only five times as complex as a bacterium like *Pseudomonas aeruginosa* (*3*). Although a significant uncertainty is still attached to this low number (see below), it was not totally unexpected, after the downward trend initiated by the analysis of the first two complete human chromosomes (*4*, *5*), as well as two independent statistical studies (*6*, *7*), and the unexpectedly low (14,000) *Drosophila* gene number (*8*).

After the older *C* value paradox (*9*), we now have an apparent *N* value paradox on our hands: Neither the cellular DNA content (in mass) nor its gene content appears directly related to our intuitive perception of organismal complexity. However, logic taught us that paradoxes often arise from the use of imprecise or ambiguous terminology. In a quick (admittedly nonrepresentative) survey among people in my laboratory, the answers to the question: "How much more complex is a human compared to a nematode?" ranged from a mere 100 to near infinity. Those widely different opinions were mostly the result of the lack of an objective (physical) measurement of what we mean by "biological complexity." Some only considered the diversity of cell types, others considered brain circuitry, and others went as far as including the cultural achievements of the human species as a whole. Thus, 30,000 human genes is not equally surprising to everybody.

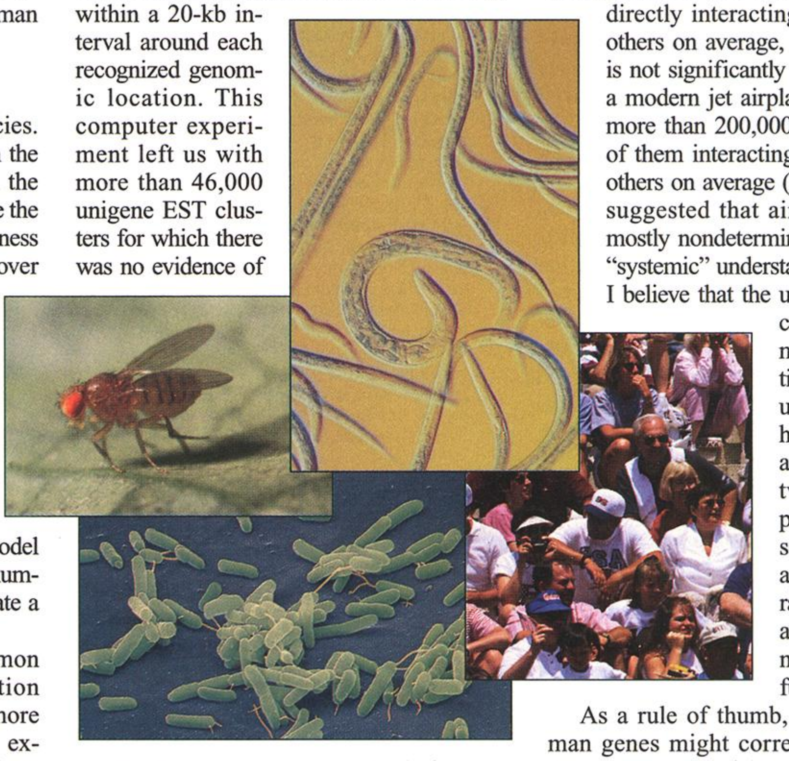
Furthermore, any personal estimate of biological complexity K can be fitted to the gene number N, by arbitrarily choosing a suitable functional relationship $K = f(N)$: proportional: $K \sim N$, polynomial: $K \sim N^a$, exponential: $K \sim a^N$, or even factorial: $K \sim$

Structural & Genetic Information Laboratory, CNRS-AVENTIS UMR 1889 31 Chemin Joseph Aiguier, 13402, Marseille, France. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr

N!. Which relationship is a reasonable one? I personally favor defining the complexity of an organism as the number of theoretical transcriptome states that its genome could achieve, where the transcriptome represents the universe of transcripts for the genome. According to the simplest model, in which each gene is either ON or OFF, a genome with *N* genes can (theoretically) encode 2^N states. According to this model, the human species appears

$$2^{30,000}/2^{20,000} = 2^{10,000} \cong 10^{3000}$$

more complex than the nematode species. This very big number (much bigger than the total number of elementary particles in the known universe) can indeed accommodate the most idealistic opinions about the uniqueness of human beings and their superiority over worms! More seriously, because genes are not independently expressed but are redundant and/or co-regulated in subsets, and also because many of these theoretical transcriptome states would be lethal, the exponents in the above formula would have to be reduced by one or two orders of magnitude. However, gene expression exhibits more than two states. A trivial mathematical model can thus illustrate how a relatively small number of genes could be sufficient to generate a tremendous biological complexity.



It is also consistent with the common view (*10*) that biological sophistication evolves through the development of more individually and finely regulated gene expression mechanisms, rather than a sheer increase in the number of genes. Accordingly, metazoan promoters do obey more intricate (and mostly unknown) triggering rules than their microbial counterparts, by making a combinatorial use of an expanded repertoire of transcription factors (*11*).

The vertebrate immune system is another example—this time real—of a biological system capable of generating a quasi-infinite repertoire of specific responses, by using a simple combinatorial logic involving a few hundred different genes that are regulated in a relatively straightforward manner.

If we can be convinced that 30,000 genes might be compatible with our perception of human complexity, this number has still to be reconciled with the much higher number of mRNA species—at least 85,000—as inferred from various assemblies of expressed sequence tags (ESTs) (*12–14*). Alternative polyadenylation is an obvious explanation for this discrepancy. However, the latest estimate (*15*) only predicts about 39,000 different "endings" from 30,000 genes (*16*). Alternative splicing is the next mechanism that can be invoked and could account for up to 48,000 different cDNAs (*16*) according to published statistics (*17*). Combining the detailed probabilities of both mechanisms in a simultaneous and independent manner could account for a maximum of 66,000 total different transcripts (albeit unlikely to generate as many nonoverlapping EST clusters). Using another approach, we mapped each of the 82,000 Unigene (release 116) clusters to the available human genome draft sequence in GenBank and searched for any significant protein homology within a 20-kb interval around each recognized genomic location. This computer experiment left us with more than 46,000 unigene EST clusters for which there was no evidence of protein-coding potential (*16*). Aside from the artifactual contamination by intron sequences (from unspliced heterogenous nuclear RNAs), the large excess of cDNA/EST clusters over identified protein-coding genes could thus be explained by two main factors: the presence of numerous alternative forms of protein-coding transcripts, together with a significant number of transcripts from uncharacterized (regulatory) "genes" not encoding proteins (such as Xist or H19). We must remember that genes of the latter category are not detected by the current ab initio gene-finding programs and are usually discovered by chance. The thorough investigation of the nagging discrepancy between protein-coding gene and apparent mRNA numbers might thus still reveal some important biological discoveries.

From a different point of view, a small number of human protein-coding genes means that the potential of functional genomics may be realized more easily and faster than anticipated. In the last few years, an increasing number of researchers [including Venter *et al.* (*1*)] have been saying that the old and classical "reductionist" approach would be totally inadequate to figure out the function of all genes. Instead, they propose that complex genetic networks should be studied as a whole, using "new theoretical approaches," according to the premise that nondeterministic and/or chaotic phenomena might govern the functioning of the human genome (*18*). Unfortunately, these new approaches (reminiscent of the old general system theory) are still poorly developed, and have no track record of significant discovery in molecular biology. In fact, with only 30,000 genes, each directly interacting with four or five others on average, the human genome is not significantly more complex than a modern jet airplane [which contains more than 200,000 unique parts, each of them interacting with three or four others on average (*19*)]. Yet, it is rarely suggested that airplane behavior is mostly nondeterministic and requires a "systemic" understanding. Accordingly, I believe that the use of simple hierarchical regulatory models in conjunction with the spectacular development of high-throughput analyses (microarray, two-hybrid system, proteomics, chemical screening, etc.) will again be sufficient to rather quickly generate most of the significant results in functional genomics.

As a rule of thumb, about 10% of human genes might correspond to potential drug targets related to diseases of socio-economical importance. With only 3000 candidate genes to work from, i.e., 30 for each of the top 100 companies throughout the world, the pharmaceutical industry is now facing a new challenge. If the high-throughput approaches cited above are used, developing leads for all of these candidates should only take a few years of fierce competition. In this context (and if patents on genes are destined to hold), one can seriously question the long-term sustainable growth and economic viability of the whole industry, as well as the future of a pharmaceutical R&D strategy consisting of developing new leads for the same targets over and over again. The "end of the beginning" (*20*) of the genomic era, might thus be followed by the "beginning of the end" very quickly, if new ways of designing and marketing medicines are not found.

Although still heralded as economically unrealistic by many, the development of personalized treatments based on genomic polymorphisms and individual transcriptome patterns might thus quickly become a necessary driving force of pharmaceutical innovation. By reporting the generation and mapping of 2.3 million new single nucleotide polymor-

CREDITS: G. BERNARD/ANIMALS ANIMALS (FLY); SINCLAIR THOMAS/PHOTO RESEARCHERS (*C. ELEGANS*); CORBIS (PEOPLE); OLIVER MECKES/PHOTO RESEARCHERS (*P. AERUGINOSA*)

phisms (SNPs) [a number comparable to what is already publicly available (*21*)] Venter *et al.* (*1*) show that these new opportunities—to paraphrase another milestone article—"have not escaped their notice" (*22*).

References and Notes

1. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
3. C. K. Stover *et al.*, *Nature* **406**, 959 (2000).
4. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
5. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
6. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
7. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
8. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
9. The observed 40,000-fold variation in eukaryote haploid DNA content ("*C* value") is unrelated to organismic complexity or to the numbers of protein-coding genes; see T. Cavalier-Smith, *J. Cell Sci.* **34**, 247 (1978).
10. L. Huang, R. J. Guan, A. B. Pardee, *Crit. Rev. Eukaryotic Gene Expr.* **9**, 175 (1999)
11. J. W. Fickett, W. W. Wasserman, *Curr. Opin. Biotechnol.* **11**, 19 (2000).
12. D. L. Wheeler *et al.*, *Nucleic Acids Res.* **29**, 11 (2001)
13. F. Liang *et al.*, *Nature Genet.* **25**, 239 (2000).
14. F. Liang *et al.*, *Nature Genet.* **26**, 501 (2000). The cited mRNA number does not take into account ESTs only sampled once and not overlapping with any others ("singletons"). There are about 300,000 singletons in the The Institute for Genomic Research human gene index (HGI release 6.0).
15. E. Beaudoing *et al. Genome Res.* **10**, 1001 (2000).
16. S. Audic, J.-M. Claverie, "The first draft of the human genome: An academic and industrial perspective," workshop at the Max-Planck-Institut für Molekulare Genetik, Berlin, 1 to 2 October 2000.
17. A. A. Mironov *et al.*, *Genome Res.* 9, 1288 (1999).
18. H. H. McAdams, A. Arkin, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 814 (1997).
19. See www.airbus.com/; thanks to P. Emrich, Airbus Industry Provisional Services Manager.
20. S. Brenner, *Science* **287**, 2173 (2000).
21. There are 2,558,564 SNPs as of 8 December 2000 in dbSNP (www.ncbi.nlm.nih.gov/SNP/), including 801,776 mapped SNPs generated by the SNP Consortium (http://snp.cshl.org/data/).
22. J. D. Watson, F. H. C. Crick, *Nature* **171**, 737 (1953).

FUTURE DIRECTIONS: SEQUENCE INTERPRETATION

Making Sense of the Sequence

David J. Galas

In this issue of *Science* on page 1304 and this week's issue of *Nature* appear versions of the sequence of the human genome (*1, 2*) that signal the dawn of a new era. For the research biologist, it is easy to think about the advantages of having the sequence of every gene of potential interest, but another thing altogether to think about how to find all of them and to validate their identities and structures. The use of genome sequences to solve biological problems has even been afforded its own label; for better or worse, it's called "functional genomics." This new way of doing biology means some real changes, many of which are well under way in the community.

Since the publication of the *Saccharomyces cerevisiae* genome in 1996 (*3*), we have become familiar with the use of the full genome sequence in investigations of gene expression patterns and controls, protein-protein interaction networks, and other biological problems (*4–6*). These investigations are marked by a global point of view that was simply not possible before we had the sequence. Although we still do not know the function of about a third of the yeast genes, we do know that all possible protein and RNA participants in cellular function are encoded in the sequence we have.

As simple as it sounds, to know that there are no other unknown genetic components that can provide alternative explanations of experimental results is a fundamental shift of perspective. This shift is beginning to transform our approach to science, enabling researchers to face the challenge of identifying all the molecular components of the cell, as well as understanding how they are controlled, interact, and function. From a picture of the "software" of the single cell, we can look to the future when researchers will begin building, with as fine a degree of resolution, an integrated view of the universe of cell-cell interactions, differentiation, and development from single cell to organism. The availability of complete sequences of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (*7–9*) is already beginning to revolutionize such studies, and this list may soon include significant sequence from other biological models of metazoan development.

Estimates from genes analyzed to date suggest that the average number of alternates spliced from the transcript of a single mammalian gene might be in the range of two to three or more. As the present sequence yields estimates of about 30,000 genes (*1, 2*), this would give us an estimated 90,000 or more distinct proteins encoded by the human genome, without considering proteolytic processing or posttranslational modifications. Thus, the complexity of the mammalian genome relative to that of yeast still presents formidable technical obstacles.

So how can the working biologist take advantage of all this new information and bring about the advances predicted? The first step is to understand that the present form of the available sequence information of the human genome is not a complete, fully annotated inventory of the human genes in each chromosome. Nor is the available sequence a single continuous and exact sequence for each chromosome. The reported genome sequence is represented by a set of sequences that cover the genome in a statistical sense but have a very large number of interruptions and gaps. Although the completeness and continuity will continue to improve, there are significant uncertainties when inferences are made from these data. The concept of the "contig" is essential to our understanding of this limitation. A contig is a contiguous piece of sequence information inferred by assembling sequence reads from single reactions (usually 400 to 800 bases in length). The number of contigs reported in the sequence data and their spectrum of sizes are important parameters in the analysis of genes. As of 12 December 2000, the public database at the U.S. National Center for Biotechnology Information (NCBI) reported that the largest contig in the entire available sequence was 28.5 megabase-pairs (Mb) in size; there were 43 contigs larger than 1 Mb, 566 contigs between 250 kb and 1 Mb, and 1628 contigs between 100 and 250 kb in size. This represented a total of approximately 600 Mb in contigs larger than 100 kb—less than 20% of the full sequence of the genome. As illustrated in figure 8 of IHGSC (*2*), half of the sequence lies in contigs 22 kb or smaller, though they can be joined to form larger contigs. We must distinguish here "initial sequence contigs," derived from sequenced clones, and "merged sequence contigs," derived by merging sequence contigs from overlapping sequenced clones [see figures 6 and 7 in (*2*)]. Because Venter *et al.* (*1*) assemble sequence contigs, not from sequenced clones, but from the entire collection of sequence reads, this distinction is not necessary in their report.

Because the average gene is of the same order of magnitude or larger than many of the contigs (a good estimate might be about 30,000 base pairs), this means that a significant fraction of human genes are unlikely to be represented on a single sequence contig in these data sets. The likelihood of finding one of the largest genes, such as Titin [~250 kb in size with >200 exons (*1*)] on a single contig is much smaller than for small, simple genes like the olfactory receptor genes, which average less than 2 kb (*2*). It will be a while before the gaps get filled in and the contigs are joined together.

Therefore, in the near future, many genes will have to be synthesized from an inferred organization of the contigs into a gapped mosaic of assemblies called "scaffolds." This

The author is at the Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA. E-mail: David_Galas@kgi.edu