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: $K \sim Na$, exponential: $K \sim aN$, or even factorial: $K \sim N!$. 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.

Figure 1

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

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 polymorphisms (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) [Medline].
4. I. Dunham *et al., Nature* **402,** 489 (1999) [Medline].
5. M. Hattori *et al., Nature* **405**, 311 (2000) [Medline].
6. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000) [Medline].
7. H. Roest Crollius *et al., Nature Genet*. **25**, 235 (2000) [Medline].
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) [Medline].
10. L. Huang, R. J. Guan, A. B. Pardee, *Crit. Rev. Eukaryotic Gene Expr.* **9**, 175 (1999) [Medline].
11. J. W. Fickett, W. W. Wasserman, *Curr. Opin. Biotechnol.* **11**, 19 (2000) [Medline].
12. D. L. Wheeler *et al*., *Nucleic Acids Res.* **29**, 11 (2001) [Medline].
13. F. Liang *et al., Nature Genet.* **25**, 239 (2000) [Medline].
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
15. human gene index (HGI release 6.0).
16. E. Beaudoing *et al. Genome Res.* **10**, 1001 (2000) [Medline].
17. 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.
18. A. A. Mironov *et al*., *Genome Res*. 9, 1288 (1999) [Medline].

19. H. H. McAdams, A. Arkin, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 814 (1997) [Medline].
20. See www.airbus.com/; thanks to P. Emrich, Airbus Industry Provisional Services Manager.
21. S. Brenner, *Science* **287**, 2173 (2000).
22. 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/).
23. J. D. Watson, F. H. C. Crick, *Nature* **171**, 737 (1953).

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

Volume 291, Number 5507, Issue of 16 Feb 2001, pp. 1255-1257.