GENOMICS

Defining Genes in the Genomics Era

Michael Snyder and Mark Gerstein

A genome is defined as the entire collection of genes encoded by a particular organism. But what is a gene? Historically, the term gene, attributed to Johansson, first appeared in the early 1900s as an abstract concept to explain the hereditary basis of traits (*1, 2*). Phenotypic traits were ascribed to hereditary factors even though the physical basis of those factors was not known. Subsequently, early genetic studies by Morgan and others associated heritable traits with specific chromosomal regions. In the 1930s, Beadle introduced the concept of "one gene, one enzyme," which later became "one gene, one polypeptide."

With the advent of recombinant DNA and gene cloning, it became possible to combine the assignment of a gene to a specific segment of DNA and the production of a gene product. Although it was originally presumed that the final product was a protein, the discovery that RNA has structural, catalytic, and even regulatory properties made it evident that the end product could be a nucleic acid (*3*). Thus, we now define a gene in molecular terms as "a complete chromosomal segment responsible for making a functional product." This definition has several logical components: the expression of a gene product, the requirement that it be functional, and the inclusion of both coding and regulatory regions. According to this definition, it should be possible to use straightforward criteria to identify genes in the DNA sequence of a genome. Five such criteria are in common use, but their application is not straightforward.

Open reading frames (ORFs). An ORF is a string of codons bounded by start and stop signals, where codons are nucleotide triplets encoding amino acids. An obvious way to find protein-coding genes is through identifying large ORFs in the genome. This is particularly applicable to prokaryotes and other organisms with few introns (the regions spliced out of RNA) in their genes. Even so, many genes are short and difficult to identify in this way. Moreover, organisms with genes that undergo an appreciable amount of RNA splicing often have small exons sandwiched between large introns, making ORFs especially difficult to find.

Sequence features. Once an ORF is identified, codon bias often is used to determine whether the ORF is a gene (*4*). The value of this measure stems from the fact that genes, particularly highly expressed genes, exhibit biased nonrandom use of codons. However, for many genes, the bias is weak, and small ORFs (or exons) contain too few codons to exhibit statistically significant bias. Beyond overall bias, one can also look for specific patterns in the DNA sequence such as splice sites to help locate genes (*5*). Computer programs that use DNA sequence features alone predict fewer than 50% of exons and 20% of complete genes (*5*). Moreover, while both the existence of an ORF and favorable sequence features may imply the presence of a gene product, they say nothing about that product's function.

Sequence conservation. In contrast to focusing on an individual DNA sequence, genes can be identified by comparing multiple sequences among organisms (*4, 5*). DNA sequence conservation among species is an excellent method to gauge the importance of the gene product. However, conserved sequences could be nontranscribed regulatory elements. Another problem with using conservation to find genes is that it requires sequences of related organisms that are separated by appropriate evolutionary distances. A current estimate of the number of genes in an organism can never be an absolute, unchanging number, because it is contingent on the specific related organisms used for comparison.

Evidence for transcription. A non–sequence-based approach for identifying genes is to search for RNA or protein expression, the hallmark of a gene product. This is commonly accomplished using microarray hybridization, serial analysis of gene expression (SAGE), cDNA mapping, or sequencing of expressed sequence tags (*6–8*). Large-scale tagging of genes with transposons reveals many new regions in the yeast genome that are capable of producing proteins (*9*) (see the figure). Likewise for humans, hybridization of labeled cDNAs to microarrays containing sequences of entire chromosomes shows that sizable fractions of the chromosomes are stably expressed (*10, 11*). However, the function, if any, of many of these transcribed regions is not known. Conversely, there appear to be conserved ORFs that are not transcribed and whose RNA or protein products have not yet been identified (see the figure).

Gene inactivation. One method for ascertaining a gene's function is to mutate or inactivate its product (*12*). This can be accomplished by direct gene disruption or RNA interference. However, many coding sequences make products whose inactivation does not result in an obvious phenotype. For instance, only one-sixth of yeast genes are essential, and mutations in the remainder usually do not cause an obvious phenotype as long as the yeast are grown in rich medium (*13*) (see the figure). Presumably, this reflects functional redundancy among gene products, assay sensitivity, or the failure to find conditions under which the product is useful. Thus, many, if not most, genes are difficult to identify solely by inactivation.

Beyond these five criteria, there are additional issues in gene identification: overlap, alternative splicing, and pseudogenes. There are now examples of overlapping reading frames of protein-coding genes, overlapping transcriptional units (for example, where the exon of one gene is encoded within the intron of another), and even overlapping protein-coding and RNA-coding genes (*14, 15*). In all cases of gene overlap, each gene has a unique functional sequence and thus is distinct.

What about products from alternatively spliced genes? In the human genome, more than half the genes have spliced isoforms, and this is likely to be an underestimate because not all variants have been identified (*16, 17*). Gene products from alternatively spliced messenger RNAs (mRNAs) have functionally unique and distinct sequences. A comprehensive system for describing such variants is lacking. Ultimately, it may be better to define a molecular parts list based on functional protein domains (the protein "domainome") rather than whole genes.

The definition of a gene is also inextricably linked with the definition of a pseudogene (or dead gene) (*18*). Pseudogenes are similar in sequence to normal genes, but they usually contain obvious disablements such as frameshifts or stop codons in the middle of coding domains. This prevents them from producing a functional product or having a detectable effect on the organism's phenotype. Pseudogenes occur in a wide variety of animals, fungi,

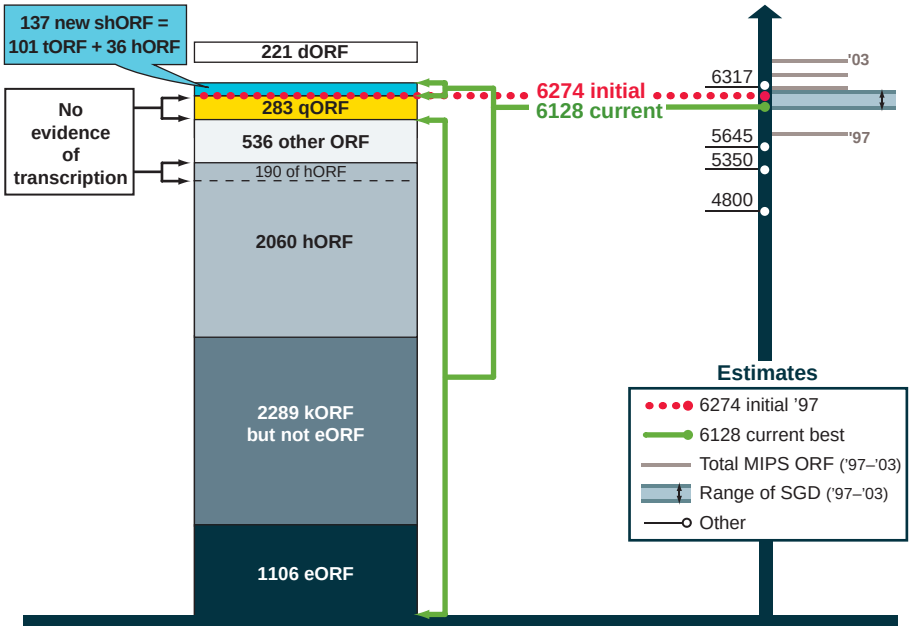
M. Snyder is in the Department of Molecular, Cellular, and Developmental Biology, and both authors are in the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.

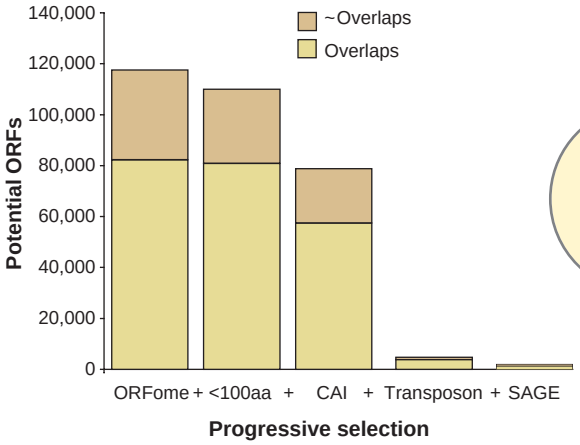
plants, and bacteria. They can be quite prevalent; for example, there are 80 ribosomal protein genes in the human genome, versus >2000 associated pseudogenes (*19*).

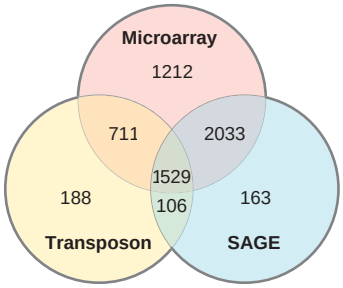
The boundary between living and dead genes is often not sharp. A pseudogene in one individual can be functional in a different isolate of the same species. For example, *FLO8* is active in one strain of yeast but inactive in another (*20*), and so technically is a gene only in one strain. Moreover, pseudogenes can be transcribed (*21*). Conversely, there are other pseudogenes that have entire coding regions without obvious disablements but do not appear to be expressed, such as, human ribosomal pseudogenes (*19*); presumably, they lack the regulatory elements required for transcription.

As a practical example of the current state of defining genes, consider the genome of the budding yeast *Saccharomyces cerevisiae*. This genome was one of the first to be sequenced, and it remains the best characterized in terms of functional genomics (which defines the functions of each gene product). Furthermore, its genes undergo only a small amount of splicing. Consequently, it is the organism for which we have the clearest grasp of which DNA sequences are genes. When the yeast genome was first sequenced, all ORFs longer than 100 codons were named, resulting in 6274 possible genes (*22*). This number has been considerably revised since then (see the figure). More small genes have been identified (*9*) either through new homologies found in databases or through evidence of transcription. In addition, 283 genes have been moved into the realm of "questionable ORFs" because they lack any evidence of transcription, function, or sequence conservation (*23*). Finally, a small number of pseudogenes have been found in the laboratory strain of *S. cerevisiae*, some of which may be functional in other yeast strains (*22*).

For yeast, the assignment of short ORFs has been particularly difficult. From the raw genome sequence, one can systematically define the universe of all possible (potentially overlapping) ORFs—what we call the "ORFome"—and then examine the evidence that each encodes a protein (see the figure). Overall, there are >100,000 possible ORFs that are longer than 15 codons. This number is constrained only slightly by codon bias, but it drops dramatically when evidence of transcription is included. However, each transcription experiment does not provide information about every possible gene in a genome. Thus, one obtains the strongest signal when one combines multiple different sources of information. That is, the likelihood that a gene encodes a functional product is best weighed using multiple criteria.







Genes, ORFs, and 'omes (**Top**) The initial published yeast genome claimed 6274 genes (*22*), but this has been revised many times since then. The time series data on numbers of genes are based on the SGD and MIPS databases: http://genome-www.stanford.edu/Saccharomyces and http://mips.gsf.de/proj/yeast/CYGD/db. These databases use different criteria for ORF inclusion: MIPS adds all candidate ORFs whereas SGD limits inclusion. Also shown are other estimates for the number of genes in the yeast genome (*26*–*29*). The central column shows the types of ORFs in the current yeast annotation. These include eORF (essential ORF) (*13*), kORF (known ORF with a well-characterized function), hORF (ORF validated by homology only), shORF (short ORF), tORF (transposon identified ORF), qORF (questionable ORF), and dORF (disabled ORF or pseudogene) (*21*). (The numbers are based on the ORF classes defined in the MIPS database.) Compared with the initial annotation, the current estimate of 6128 ORFs reflects two opposing trends: (i) the addition of new shORFs (*9*) found either through transcription experiments (tORFs) or from sequence comparisons with proteins newly deposited in the databases (hORFs); (ii) the removal of qORFs with no evidence of being transcribed (that is, lacking SAGE or transposon tags, and not expressed on microarrays) and with no sequence similarity to any other protein. (For simplicity, we include in the qORFs 10 ORFs associated with Ty elements in the original annotation. Further information is at http://bioinfo.mbb.yale.edu/genome/yeast/orfome.) (**Bottom left**) The explosion in defining shORFs. The first bar depicts the potential ORFs in the raw DNA sequence of the yeast genome that are >15 codons. The second bar shows the large number that are also <100 codons in length. The third bar demonstrates that the number of ORFs is not reduced by requiring a high codon adaptation index (CAI > 0.11). The remaining bars illustrate how the number of potential ORFs is radically reduced by selecting only those shORFs that show evidence of transcription (transposons and SAGE). (**Bottom right**) Functional genomics information is best used in a combined fashion. Illustrated is the number of ORFs in the yeast genome that are transcribed according to data from microarray hybridization, SAGE, and transposon tagging.

The yeast genome is, of course, far simpler than the human genome, and we expect many of the problems evident in yeast to be greatly magnified in human. First, we expect the human genome to contain a vast number of potential ORFs given the small size of exons (average size ~140 base pairs) and the complexity of mRNA splicing (*16, 19*). It is doubtful that we will be able to find true genes among these ORFs solely by analyzing their raw nucleotide sequences. In fact, initial estimates of the number of genes in the human genome ranged from 20,000 to >100,000 (*17, 23–25*).

One solution for annotating genes in sequenced genomes may be to return to the original definition of a gene—a sequence encoding a functional product—and use functional genomics to identify them. Moreover, if we add conservation information obtained from cross-genome comparisons, we can streamline the process. Ultimately, we believe that identification of genes based solely on the human genome sequence, while possible in principle, will not be practical in the foreseeable future. Only through large-scale systematic functional genomics experiments and through careful sequence comparisons against related organisms will we be able to convincingly arrive at a definitive annotation of the human genome.

References and Notes

1. M. Morange, *The Misunderstood Gene* (Harvard Univ. Press, Cambridge, MA, 2001).
2. R. Falk, *Stud. Hist. Philos. Sci.* **17**, 133 (1986).
3. S. Eddy, *Cell* **109**, 137 (2002).
4. C. Burge, S. Karlin, *Curr. Opin. Struct. Biol.* **8**, 346 (1998).
5. M. Zhang, *Nature Rev. Genet.* **3**, 698 (2002).
6. C. Horak, M. Snyder, *Funct. Integ. Genomics* **2**, 171 (2002).
7. P. Brown, D. Botstein, *Nature Genet.* **21**, 33 (1999).
8. V. Velculescu *et al.*, *Cell* **88**, 243 (1997).
9. A. Kumar *et al.*, *Nature Biotechnol.* **20**, 58 (2002).
10. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
11. J. Rinn *et al.*, *Genes Dev.* **17**, 529 (2003).
12. P. Coelho *et al.*, *Curr. Opin. Microbiol.* **3**, 309 (2000).
13. G. Giaever *et al.*, *Nature* **418**, 387 (2002).
14. P. Coelho *et al.*, *Genes Dev.* **16**, 2755 (2002).
15. K. T. Tycowski *et al.*, *Nature* **379**, 464 (1996).
16. B. Modrek, C. Lee, *Nature Genet.* **30**, 13 (2002).
17. E. Lander *et al.*, *Nature* **409**, 860 (2001).
18. P. Harrison, M. Gerstein, *J. Mol. Biol.* **318**, 1155 (2002).
19. Z. Zhang *et al.*, *Genome Res.* **12**, 1466 (2002).
20. H. Liu *et al.*, *Genetics* **144**, 967 (1996).
21. P. Harrison *et al.*, *J. Mol. Biol.* **316**, 409 (2002).
22. H. Mewes *et al.*, *Nature* **387** (suppl.), 7 (1997).
23. P. Harrison *et al.*, *Nucleic Acids Res.* **30**, 1083 (2002).
24. J. Venter *et al.*, *Science* **291**, 1304 (2001).
25. M. Das *et al.*, *Genomics* **77**, 71 (2001).
26. M. Kowalczuk *et al.*, *Yeast* **15**, 1031 (1999).
27. P. Mackiewicz *et al.*, *Yeast* **19**, 619 (2002).
28. C. Zhang, J. Wang, *Nucleic Acids Res.* **28**, 2804 (2000).
29. G. Blandin *et al.*, *FEBS Lett.* **487**, 31 (2000).

30. We thank A. Kumar, M. Vidal, S. Karlin, C. Burge, P. Harrison, Z. Zhang, M. Zhang, W. Summers, M. Cherry, R. Lifton, M. Muensterkoetter, M. Seringhaus, and A. Sali for helpful comments.

PLANETARY SCIENCE

A Liquid Core for Mars?

Veronique Dehant

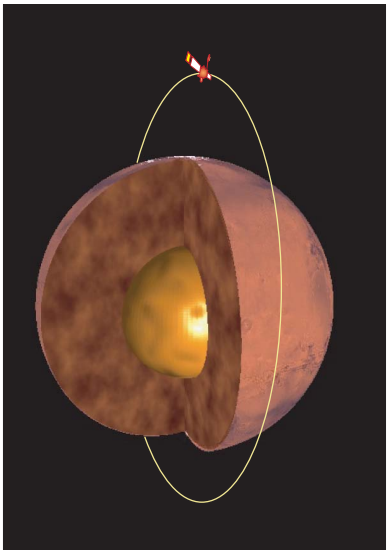
Mars is a planet very similar to Earth. Early in their evolution, both planets must have been sufficiently hot to be molten. Earth still has a liquid core, but the smaller size of Mars would favor faster cooling. Extrapolation from Earth suggests that Mars today should therefore not have a liquid core. However, small differences in elemental composition between the two planets prevent our simply extrapolating from knowledge of Earth's properties (*1*). On page 299 of this issue, Yoder *et al*. (*2*) present evidence that the iron core of Mars is liquid, with important implications for martian geology.

Enhanced online at www.sciencemag.org/cgi/content/full/300/5617/260

There are a few constraints on Mars' deep interior based on analysis of martian meteorites (*3*, *4*), observation of the absence of a global magnetic field (*5*), and knowledge of the planet's mass and moments of inertia (*6*). Moments of inertia quantify the global mass repartition within Mars. They provide evidence for the existence of a denser martian core and can be used to constrain the core dimension (*7*). However, the uncertainty of the core's density and dimension remains large because they depend on the temperature profile and light element abundance, and these properties are still unknown.

The author is at the Observatoire Royal de Belgique, Bruxelles, B-1180 Belgium. E-mail: veronique.dehant@oma.be

Scientists interested in modeling the martian interior are therefore looking for other kinds of complementary data. As for Earth, the Sun's gravitational attraction induces global phenomena on Mars—namely, tides and precession-nutation (the motion of the rotation axis in space). Tides are deformations induced by the gravitational pull of the Sun. They are related to surface displacements, surface gravity changes (such as those that would be measured by a gravimeter on the martian surface), and mass repartitioning inside the planet. These changes are periodic, with periods related to Mars's orbit around the Sun (and, to a minor extent, to the orbits of the two martian moons, Phobos and Deimos, around Mars).



The physical state of the martian core observed by MGS orbit tracking.

To study these phenomena, long-term observations—for example, of the annual or semiannual periods—are needed. Surface gravity data, surface displacements, and nutations cannot yet be observed because their measurement requires a network of geophysical stations on the martian surface (*8*). But some information can be obtained from a Mars orbiter such as Mars Global Surveyor (MGS), which is (in addition to the classical steady-state self-gravity of the planet) subject to gravitational forces resulting from the mass redistributions induced by the tides. Hence, information on the planet's response to the tidal force may be deduced from the precise reconstruction of the MGS orbit. Because this response depends on the internal structure of Mars, it is possible to infer properties of the core.

The mass repartitioning induced by the tides is usually described by a set of dimensionless numbers called "Love numbers," which express the nonrigidity of the planet. The value of the *k*-Love number (the Love number relevant for the perturbation of the orbit) will be much larger if the core is liquid than if it is solid (liquid versus solid core values change by ~50%) (*9*). Observational constraints on this *k*-Love number would allow the physical state of the core to be determined.

The long time series of Mars Global Surveyor DSN (Deep Space Network) tracking data provides such constraints. Smith *et al*. (*10*) have used these data to deduce the *k*-Love number directly from the position of the spacecraft orbiting Mars. However, the main term of the gravitational potential was unfortunately not very accurate.

Yoder *et al*. now use another indirect observation of the gravitational effect in-