Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 10004–10006, September 1996

Commentary

The minimal cell genome: ''On being the right size"

*Jack Maniloff**

Department of Microbiology and Immunology, University of Rochester, Medical Center Box 672, Rochester, NY 14642

''On being the right size'' is in quotation marks in the title because it has been used twice before–in each case to discuss the size of biological systems in terms of interest at that time. J. B. S. Haldane (1) originally used this title in the 1920s, for a paper describing the relationship between animal size and constraints like gravity, surface tension, and food and oxygen consumption. In the 1970s, N. W. Pirie (2) (acknowledging Haldane) used the title for a paper analyzing how factors like membrane properties, water structure, and the volume needed for ribosomes and other macromolecules set lower limits on cell size. Now, in the 1990s, the title is used to discuss efforts to define the minimal genome content necessary and sufficient for a living cell.

The search for the ''smallest autonomous self-replicating entity,'' which subsequently became the search for the smallest cell genome, was begun in the late 1950s by Harold Morowitz and coworkers (for review, see ref. 3). This led to studies of the mycoplasmas, showing that these microorganisms have the smallest reported cell (4) and genome (5, 6) sizes. The DNA sequence of the smallest known mycoplasma genome, that of *Mycoplasma genitalium*, recently was determined (7). Now, Mushegian and Koonin (8), in a paper in the current issue of the *Proceedings*, have carried out an intensive comparative analysis of the *M. genitalium* sequence with that of *Haemophilus influenzae* (9), the only other complete bacterial genome sequence presently available, to try to identify the minimal cell gene set and to suggest how this set might be reduced even further to reconstruct the genetic content of the primordial ancestral cell.

M. genitalium and *H. influenzae* provide a unique opportunity for such an analysis because both microorganisms evolved by repeated genome reductions from bacteria with larger genomes, *M. genitalium* from Gram-positive bacteria and *H. influenzae* from Gram-negative bacteria. Mushegian and Koonin (8) note that these genome losses and the fact that the Gram-positive and Gram-negative bacteria arose from an ancient phylogenetic divergence [at least 3200 million years (Myr) ago; ref. 10] suggest the genes conserved in both microorganisms should approximate the minimal gene set for a living cell. Using sequence data for the 469 putative protein genes encoded in the *M. genitalium* genome (8) and 1703 genes in the *H. influenzae* sequence (11), Mushegian and Koonin (8) constructed a minimal gene set by identifying genes with similar amino acid sequences and functions in both microorganisms (233 genes), adding in genes needed to fill gaps in essential metabolic pathways and not encoded in similar sequences in both microorganisms (23 genes), and subtracting functionally redundant and host-specific genes (6 genes). The resulting 250 genes are presented as an approximation of the minimal gene set necessary and sufficient for a ''modern-type'' cell. This corresponds to a genome size about 315 kb, using 1.25 kb per gene for small genomes based on the *M. genitalium* data (7). Mushegian and Koonin (8) also present a strategy for reducing the minimal gene set further, to try to reconstruct the genetic content of the last common ancestor of the Archaea, Bacteria, and Eucarya.

The Mushegian and Koonin (8) paper is the latest of several recent efforts to define a minimal genome. In considering these studies, it is important to remember that organisms with small genomes have arisen by two vastly different evolutionary pathways, one ''top down'' and the other ''bottom up.''

''Top down'' genomes evolved in organisms with increasing metabolic requirements. At each metabolic level, there can be a minimal genome: from photoautotrophs, able to grow in a medium of only CO_2, light, and inorganic salts; to simple heterotrophs, for which growth requires a basic medium containing an organic carbon and energy source (like glucose) and inorganic salts; to fastidious heterotrophs, for which growth requires a complex medium, frequently containing undefined components (like serum); to obligate intracellular parasites, only able to grow within eukaryotic host cells; and, finally, to organelles (i.e., mitochondria and chloroplasts) that have lost almost all bacterial genes and become eukaryotic cytoplasmic organelles.

''Bottom up'' genomes refer to the genetic contents of the ''organisms,'' postulated by different investigators, to have arisen during the origin and evolution of cells. In the absence of a plausible alternative, small RNA oligonucleotides are believed to have been the primordial informational macromolecules (for review, see refs. 12 and 13). The first ''organism'' in the RNA world probably contained a population of similar, but not identical, self-replicating RNA oligonucleotides with relatively broad catalytic specificity. This presumably led to an organism in which genetically encoded RNA directed synthesis of a protein–designated the ''breakthrough organism'' by Steven Benner and coworkers (14). Subsequently, selection must have produced organisms with increasing numbers of different RNAs and proteins, but inaccurate replication and translation systems would have limited the size of these macromolecules to minigenes (perhaps 50–100 nt) and miniproteins (perhaps 20 residues or shorter) (12). This organism, containing populations of heterogeneous minigenes and miniproteins subject to natural selection, probably corresponds to the ''progenote'' hypothesized by Carl Woese (15): ''a theoretical construct, an entity that, by definition, has a rudimentary, imprecise linkage between its genotype and phenotype.'' Eventually, more accurate replication and translation systems must have evolved, allowing evolution of larger DNA genes and proteins, comparable in size to modern organisms. An organism containing these larger DNA genes and proteins would have been the last common ancestor of the Archaea, Bacteria, and Eucarya–designated the ''protogenote'' by Steven Benner and Andrew Ellington (16) and the ''ancestral cell'' by Christian de Duve (12). It is not evident exactly when DNA would have arisen during this progression. Assembly of minigenes into longer genes, and eventually genomes, can be visualized most easily in an RNA world in which RNA splicing could be utilized. DNA biochemistry may have been present as early as the breakthrough organism but not selected for until the transition from progenote to ancestral cell. As noted by Christian de Duve (12): ''In other words, life may have played with DNA for a long time in what might be called a half-hearted way and have adopted it definitely only when the need for it made itself felt through selective benefits.''

*e-mail: jman@medinfo.rochester.edu.

Therefore, small genomes in "top down" organisms arose by gene attrition, whereas small genomes in "bottom up" organisms arose by gene accretion. This means that small genomes in "top down" organisms are the result of an evolutionarily engineered downsizing, with the loss of selected genes from highly evolved, regulated, and integrated metabolic pathways (for review, see ref. 17). In contrast, small genome "bottom up" organisms are the result of evolutionary tinkering, with the recruitment of unregulated minigenes or genes into primitive metabolic pathways (for review, see refs. 18 and 19). The "bottom up" expansion and refinement of metabolic pathways is a fascinating example of François Jacob's observation that evolution works like a tinkerer, not like an engineer (20): "It works like a tinkerer–a tinkerer who does not know exactly what he is going to produce but uses whatever he finds around him whether it be pieces of string, fragments of wood, or old cardboards; in short it works like a tinkerer who uses everything at his disposal to produce some kind of workable object. For the engineer, the realization of his task depends on his having the raw materials and the tools that exactly fit his project. The tinkerer, in contrast, always manages with odds and ends. . . . In contrast with the engineer's tools, those of the tinkerer cannot be defined by project. What these objects have in common is 'it might well be of some use.' For what? That depends on the opportunities."

Experimental data on minimal genomes are limited to "top down" organisms, with the unexpected observation that the smallest known genomes are in certain free-living microorganisms rather than, as might have been expected, in microorganisms metabolically dependent on growth in eukaryotic host cells. In particular, although most mycoplasmas can be grown in axenic culture, they have smaller genomes than obligate intracellular parasites such as *Rickettsia* and *Chlamydia* (5, 6).

Mycoplasma species grow on the surfaces of a variety of hosts (e.g., humans, pumas, seals, insects, and plants) and have genome sizes ranging from 600 to 1700 kb (21). The nature of the selective pressure for repeated genome reductions during mycoplasma phylogeny is not known. Genome reductions in microorganisms have been suggested to be due to selection for faster (hence, smaller) replicating genomes to produce greater progeny cell yields, selection for smaller genomes to reduce the energy burden on cells growing in limiting nutrient environments, and loss of genome sequences with deleterious mutations if these mutations cannot be replaced by recombination with wild-type genomes (17, 22). Any or all of these factors could have affected mycoplasma genome phylogeny.

A surprising aspect of the phylogeny of the smallest mycoplasma genomes is that they evolved several different times (23). The ancestral mycoplasma arose from the *Streptococcus* phylogenetic branch about 590 to 600 Myr ago (unpublished data), probably from an organism with a genome size about 2000 kb (24). The mycoplasma phylogenetic tree later split into two major branches, about 450 Myr ago, probably from an organism with a genome size of 1700–2000 kb (unpublished data). Both branches then evolved to produce mycoplasma sublines with genome sizes of 1200–1700 kb (24). Mycoplasma species with small (600–1100 kb) genomes subsequently arose independently on several different sublines.

It is interesting that the smallest genome on each mycoplasma subline is 600–800 kb. This suggests that, on several independent phylogenetic branches, mycoplasmas have tested the 600- to 800-kb size limit and been unable to reduce their genome size further. Therefore, unless these smallest genome mycoplasmas are still evolving and undergoing genome reductions, 600–800 kb seems to be the lower limit for mycoplasma, and presumably cell, genome content.

Hence, the 469 genes of *M. genitalium* form the smallest gene set for a living organism. This is almost double the minimal 250 gene set derived by Mushegian and Koonin (8). At most 10% of *M. genitalium* genes appear to be host-specific (e.g., cytoadherence and surface antigen genes) (7). Even with these subtracted, the *M. genitalium* gene set is still appreciably larger than the minimal gene set of Mushegian and Koonin (8). This probably reflects the difference between a small (minimal) tinkered-together gene set, produced by a couple billion years of evolution, and a small (minimal) engineered gene set, produced by a computer. Even so, the numbers are surprisingly and encouragingly close.

There are two other efforts to determine minimal genome sizes. M. Itaya (25) has estimated the "minimal genome size required for life" by determining the fraction of the *Bacillus subtilis* genome containing dispensable loci following random mutagenesis. He calculated an average of 318 kb and a maximum of 562 kb for the minimal genome size. This corresponds to 254–450 genes, using 1.25 kb per gene for small genomes based on the *M. genitalium* data (7), in agreement with the Mushegian and Koonin minimal gene set and the *M. genitalium* gene set (8). In a very different approach, Siv Andersson, Charles Kurland, and coworkers are studying genome reductions and organizational changes in the transitions from free-living bacteria to obligate intracellular parasites to eukaryotic organelles (S. G. E. Andersson, personal communication). The DNA sequence of the *Rickettsia prowazekii* genome is being determined (S. G. E. Andersson, A.-S. Eriksson, A. K. Näslund, M.S. Anderson, and C.G. Kurland, personal communication). Rickettsia are α-proteobacteria and, therefore, belong to the same phylogenetic branch as the microbial ancestor of the mitochondria. Comparative sequence analysis should allow identification of genes lost in the transition of bacteria from a free-living to an obligate intracellular state and of general patterns in reductive genome evolution. Although the *R. prowazekii* sequence has not been completed, the available data have already provided examples of evolution of novel gene functions in the transition to intracellular growth in comparisons of *R. prowazekii*, *Chlamydia*, and mitochondria, and of reductive convergent and reductive divergent genome evolution in comparisons of *R. prowazekii* and *M. genitalium* (S.G.E. Andersson, personal communication).

Since the early days of molecular biology, the search for the minimal genome has been the "Holy Grail" in an effort to define the necessary and sufficient components for a living system. The Mushegian and Koonin minimal gene set (8) provides an important construct in this search and can be expected to be refined as more genome sequences are completed. It should be instrumental in organizing ideas about genomics, and if the 250 gene number is too small, the arguments for adding genes will demand understanding the difference between a tinkered-together genome and an engineered one. Also, because such a gene set may represent a minimal metabolic core of reactions for all cells, subtracting the minimal gene set from an organism's total gene inventory should reveal genes for the phenotypic characters that make each organism unique.

The minimal gene set will also be a useful paradigm for thinking about the origin of cells, from the RNA world to the ancestral cell. However, it must be remembered that any minimal gene set that is defined will be a subset of the larger tinkered-together gene set in the original ancestral cell genome, and that ancestral cell gene set arose by natural selection for growth in an environment of high temperature, reducing atmosphere, and sulfur metabolism.

1. Haldane, J. B. S. (1928) *Possible Worlds and Other Papers* (Harper & Brothers, New York), pp. 20–28.
2. Pirie, N. W. (1973) *Annu. Rev. Microbiol.* **27,** 119–132.
3. Morowitz, H. J. (1984) *Isr. J. Med. Sci.* **20,** 750–753.
4. Carson, J. L., Hu, P.-C. & Collier, A. M. (1992) in *Mycoplasmas: Molecular Biology and Pathogenesis*, eds. Maniloff, J., McElhaney,

R. N., Finch, L. R. & Baseman, J. B. (Am. Soc. Microbiol., Washington, DC), pp. 63–72.
5. Krawiec, S. & Riley, M. (1990) *Microbiol. Rev.* **54,** 502–539.
6. Cole, S. T. & Saint Girons, I. (1994) *FEMS Microbiol. Lett.* **14,** 139–160.
7. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270,** 397–403.
8. Mushegian, A. R. & Koonin, E. V. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10268–10273.
9. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269,** 496–512.
10. Stackebrandt, E. (1995) in *Molecular Basis of Viral Evolution*, eds. Gibbs, A. J., Calisher, C. H. & García-Arenal, F. (Cambridge Univ. Press, New York), pp. 224–242.
11. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996) *Curr. Biol.* **6,** 279–291.
12. de Duve, C. (1991) *Blueprint for a Cell: The Nature and Origin of Life* (Neil Patterson, Burlington, NC).
13. Orgel, L. E. (1994) *Sci. Am.* **271,** 77–83.
14. Benner, S. A., Ellington, A. D. & Tauer, A. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7054–7058.
15. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
16. Benner, S. & Ellington, A. D. (1990) *Science* **248,** 943–944.
17. Andersson, S. G. E. & Kurland, C. G. (1995) *Biochem. Cell Biol.* **73,** 775–787.
18. Jensen, R. A. (1976) *Annu. Rev. Microbiol.* **30,** 409–425.
19. Jensen, R. A. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, eds. Neidhardt, F. C., Curtiss, R., III, Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M. & Umbarger, H. W. (Am. Soc. Microbiol., Washington, DC), Vol. 2, pp. 2649–2662.
20. Jacob, F. (1977) *Science* **196,** 1161–1166.
21. Razin, S. (1992) in *Mycoplasmas: Molecular Biology and Pathogenesis*, eds. Maniloff, J., McElhaney, R. N., Finch, L. R. & Baseman, J. B. (Am. Soc. Microbiol., Washington, DC), pp. 3–22.
22. Samuelsson, T. & Borén, T. (1992) in *Mycoplasmas: Molecular Biology and Pathogenesis*, eds. Maniloff, J., McElhaney, R. N., Finch, L. R. & Baseman, J. B. (Am. Soc. Microbiol., Washington, DC), pp. 575–591.
23. Maniloff, J. (1992) in *Mycoplasmas: Molecular Biology and Pathogenesis*, eds. Maniloff, J., McElhaney, R. N., Finch, L. R. & Baseman, J. B. (Am. Soc. Microbiol., Washington, DC), pp. 549–559.
24. Maniloff, J. (1983) *Annu. Rev. Microbiol.* **37,** 477–499.
25. Itaya, M. (1995) *FEBS Lett.* **362,** 257–260.