A microbial minimalist

Barry R. Bloom

WHAT'S in a genome? Lots of genes, of course, but the first complete bacterial genome sequences to be published[1–3] reveal that each genome has a unique story to tell and adds its own new mysteries to be explained. The announcement in *Science* last month by Fraser *et al.*[1] of the *Mycoplasma genitalium* sequence comes on the heels of the *Haemophilus influenzae* (strain Rd) sequence released earlier this year[2,3] by scientists from university and government and from the Institute of Genomic Research.

Mycoplasma genitalium[1] was chosen because it has the smallest genome known for a self-replicating organism (580 kilobases), with the hope that it might provide insight into the minimal functional gene set for a living organism. *Haemophilus influenzae* has a larger but still relatively small genome (1.8 megabases), and serves well as a model bacterium; the major human pathogen of this species is the type b, a major cause of otitis media, for which a most effective vaccine has recently become available. Although these are by no means the first complete genomes sequenced, the first being that of the phage ΦX174 (5,386 base pairs) by Sanger in 1977 and the most complex being that of cytomegalovirus (229 kilobases), they are the largest to date.

Sequencing

As physical maps did not exist for either bacterial genome, they were sequenced by a random shotgun strategy applied to the whole genome using 3–9-fold coverage and high throughput, and analysed with sophisticated assembly and alignment software[2,3]. This means that, for *H. influenzae*, 28,000 sequencing reactions were performed; for *M. genitalium*, the entire genome was sequenced on eight sequencing machines by five individuals working for two months. Gaps were then closed by sequencing overlapping fragments obtained from independent libraries. This assault was remarkably rapid, error-free (1 base in 5,000–10,000), and cost-effective. A comparable amount of time was spent analysing the sequence data and sieving existing bacterial sequence information for comparison: likely genes were assigned and the genetic map annotated with putative operons, regulatory regions, starts and stops. The sequencing cost on this scale works out at 30 cents per base, or about $200,000 for the complete sequence of the *Mycoplasma* genome.

What has been learned about the genes? In the case of *H. influenzae*, 1,743 possible coding regions were identified, 1,007 of which could take on various functions in the cell, including amino-acid and lipid metabolism, biosynthesis of cofactors and the cell envelope, energy production, transport, nucleotide and protein synthesis, and DNA replication and transcription. The distribution of the respective open reading frames in *H. influenzae* is nicely shown in colour-coded maps, which give a handy picture of its genome organization and an indication of the amount of DNA a typical bacterium chooses to invest in different functions — for example, 10% goes to energy metabolism, 17% to transcription and translation, 12% to transport and 8% to cell envelope proteins.

In the case of the 470 predicted coding regions of *M. genitalium*, an organism that lives in association with mammalian cells, the investment is quite different. For example, whereas *H. influenzae* has 68 genes for amino-acid biosynthesis, *M. genitalium* has only one. It has no genes for cytochromes or enzymes of the tricarboxylic acid cycle. *Mycoplasma genitalium* is by no means the earliest eubacterium and it has obviously learned to simplify its life by association with its mammalian hosts, so it is fascinating to learn what it could afford to shed and still survive. Yet it has committed almost 5% of its genome to repeated elements encoding an adhesin, which presumably allows it to stick to the cells that nurture it and which can probably be altered by recombination to enable antigenic variants to elude the immune responses of its host.

This leads to some of the mysteries, the untold stories. Foremost is the fact that despite the wealth of information in the sequence databases from microorganisms, fully a third of the open reading frames of both genomes predict sequences that cannot be assigned any biological function. Does this mean that each organism has evolved some very specialized proteins, not common to others, to carry out important or unique functions, or simply that insufficient information is available to account for the species diversity? And what do the 90 genes found in *M. genitalium*, but not in *H. influenzae*, actually do? Finally, how does *M. genitalium* survive without a transcription factor for the stress response? Perhaps in simplifying its lifestyle it has learned how to avoid or cope with stress — perhaps a lesson for us all.

What does all this mean for microbiologists? At best, it means that they will be freed to do more biology, rather than just molecular biology. The need for random mutagenesis and screening — very inefficient approaches to defining gene functions even when transposable elements are available, and overwhelming when not — will be superseded by amplification using the polymerase chain reaction of specific genes of interest. These will be of known or unknown function, readily mutated, and rapidly deleted from or inserted into the chromosome in order to approach the mysteries of uniqueness and diversity.

Availability

For pathogens, this means defining the genes that control virulence. It means that complete information on these — and scientists must insist that this applies to *all* sequenced genomes — should be available to scientists everywhere, not only in print, but in usable databases as in the work[4] discussed here, to allow scientists to pursue the implications of this knowledge.

As with all microbial genetics, we can predict that some of the fall-out from knowing the complete genome sequences will enable organisms to be modified for medical, agricultural and economic purposes. And we must anticipate the more sinister application to biological warfare, calling for vigilant international surveillance.

Finally, in light of widespread public interest in the threat of infectious diseases and emerging pathogens, it is curious that in the initial formulation of the Human Genome Project not a single pathogen was included. Neither are the two organisms discussed here significant pathogens. Given the desperate need for public understanding and support of science, one could argue that, in the words of Tallyrand upon the assassination of the Duc de Broglie, "It was worse than a crime: it was a blunder".

The power and cost-effectiveness of modern genome sequencing technology mean that complete genome sequences of 25 of the major bacterial and parasitic pathogens could be available within five years. For about 100 million dollars (only 500 times the investment in the *Mycoplasma* genome), we could buy the sequence of every virulence determinant, every protein antigen and every drug target. It would represent for each pathogen a one-time investment from which the information derived would be available to all scientists for all time. And we could then think about a new, post-genomic era of microbe biology. □

Barry R. Bloom is at the Howard Hughes Medical Institute, Albert Einstein College of Medicine, Bronx, New York 10461, USA.

1. Fraser, C. M. *et al. Science* **270**, 397–403 (1995).
2. Fleischmann, R. D. *et al. Science* **269**, 496–512 (1995).
3. Smith, H. O., Tomb, J.-F., Dougherty, B. A., Fleischmann, R. D. & Venter, J. C. *Science* **269**, 538–540 (1995).
4. World Wide Web site http://www.tigr.org