A sequence of changes

With more than 1,000 bacterial and archaeal genome sequences now available, we look at the progress that has been made over the past 15 years.

2010 marks the fifteenth anniversary of the availability of the first complete bacterial genome sequence, that of *Haemophilus influenzae*[1]. At 1.8 Mb and with ~1,750 protein-coding genes, this sequencing effort seems modest today, but it was a crucial genomics milestone. The *H. influenzae* sequence was followed a few months later by the sequence of *Mycoplasma genitalium*, and the first archaeal genome sequence — that of *Methanocaldococcus jannaschii* — was published the following year. 5 years later, 30 complete sequences were available. From then on, the pace of completion began to accelerate, and on 21 October 2009 a landmark in microbial genome sequencing was reached with the completion of the thousandth sequence, that of *Methanocaldococcus vulcanius* M7.

This data explosion was fuelled by improvements in the Sanger-based sequencing technologies and, more recently, the development of the so-called 'next-generation' ultra-high-throughput sequencing methods, which have substantially lowered the costs and increased the speed of genome sequencing. Microbial genome sequencing is now a routine part of many grant applications, and although dedicated sequencing centres still carry out most genome sequencing projects, sequencing facilities are now available in many individual academic departments; indeed, a bench-top sequencing platform should soon be commercially available.

The completed genomes include those of all the important model and reference organisms, and multiple sequences are available for many of the key human and animal pathogens. However, over the past 15 years microorganisms have been selected for sequencing on an *ad hoc* basis. Consequently, the phylogenetic distribution of the completed sequences is understandably biased towards organisms of specific medical or economic interest, and more than 80% of the available sequences represent just three major lineages: Proteobacteria, Firmicutes and Actinobacteria.

> "the phylogenetic distribution of the completed sequences is ... biased towards organisms of specific medical or economic interest"

This bias was noted in a recent report (Large-Scale Sequencing: The Future of Genomic Sciences?) from the American Academy of Microbiology (AAM). The report represents the output of an AAM colloquium that was held in September 2008 to assess the issues associated with large-scale sequencing and review how best to proceed. The colloquium participants recommended that a large-scale, taxonomically driven effort to redress the balance in the sequenced genomes is warranted and should begin with the targeting sequencing of cultured isolates, in particular those with detailed metadata, and then move towards sequencing of non-cultured species and genera as single-cell isolation methods improve.

The AAM report concludes: "A well-coordinated large-scale effort to target the breadth and depth of microbial diversity would result in the greatest impact." Interestingly, the results of a pilot project that aims to do just that have recently been published[2]. The paper presents the first results from a US Department of Energy-funded project known as the Genomic Encyclopaedia of Bacteria and Archaea (GEBA), the ultimate aim of which is to use targeted sequencing to generate "a phylogenetically balanced genomic representation of the microbial tree of life" (REF. 2). In the pilot project, the authors used a phylogenetic tree constructed using small subunit ribosomal RNA genes to identify major branches for which cultured isolates were available that had not yet been sequenced. A total of 159 isolates were selected for sequencing, and this publication presents the analysis of the first 56 of those. The results indicate that such a systematic approach generates substantial additional information compared with the sequencing of isolates that have been selected at random. Moreover, the authors calculated that sequencing just ~1,500 isolates in this targeted manner should effectively sample 50% of the diversity of known cultivated species, a vast improvement on the current sampling.

The availability of genome sequence data dramatically increased our understanding of numerous areas of microbiology, from basic microbial physiology to the identification of the genes underpinning pathogenicity. The steady availability of microbial genome data in the late 1990s kick-started the 'omics' revolution and led us into the realms of functional genomics, comparative genomics and metagenomics, as well as other areas of research such as systems (micro)biology, reverse vaccinology and, more recently, synthetic biology. However, for those interested in microbial diversity, at times it has seemed that the increasing amount of genome sequence data has served only to emphasise that we are just 'scratching the surface' in our understanding. The promising results for GEBA suggest that this, too, may be about to change.

1. Fleischmann, R. *et al. Science* **269**, 496–512 (1995).
2. Wu, D. *et al. Nature* **462**, 1056–1060 (2009).

© 2010 Macmillan Publishers Limited. All rights reserved