GENOME WATCH

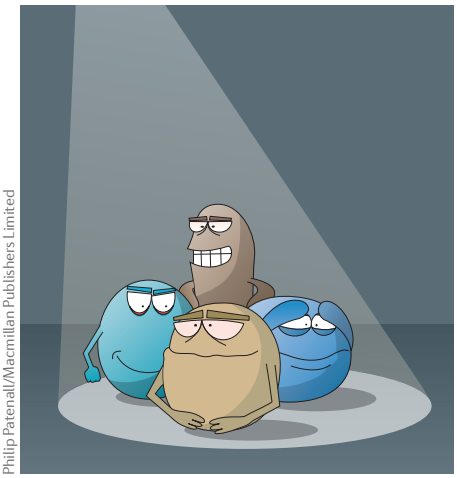
Illuminating microbial diversity

Samuel C. Forster

This month's Genome Watch discusses the release of 1,003 bacterial and archaeal genomes, and describes how they could increase our understanding of the diversity of microbial biological functions and contribute to improved metagenomic analyses.

Almost every environment on Earth, from soils to oceans to the human gastrointestinal tract, has its own unique community of microbial species. The majority of these species have never been grown in the laboratory and many of those that have been cultured and characterized have not been sequenced. Advances in sequencing technologies have led to a substantial increase in the availability of microbial genomes; however, this research has been typically applied to species that are pathogenic or have important roles in industrial applications. The relatively narrow phylogenetic diversity that has been captured by these studies has limited the potential for understanding the vast functional biology within microbial communities.

In 2009, with less than 1,000 bacterial and archaeal genomes sequenced globally, the Genomic Encyclopaedia of Bacteria and Archaea (GEBA) initiative sequenced 56 genomes from phylogenetically diverse species. This work highlighted the immense range of bacterial genes and functions that exist in nature, even within known species, and demonstrated the importance of further investigation into this rich and unexplored resource[1]. With the advent of single-cell genomics, Rinke *et al.* further expanded this resource in 2013 by sequencing 201 uncultivated archaeal and bacterial cells from diverse environments, which led to the discovery of more than 60 phyla[2]. In the latest GEBA study, Mukherjee *et al.* sequenced 974 bacterial and 29 archaeal reference strains across 21 phyla to further increase our understanding of the phylogenetic diversity within the prokaryotic branch of the tree of life[3]. This study represents the most diverse set of genome data that has been released to date and includes 845 previously unsequenced species. Alone, this one dataset increased known protein diversity by 10.5% and includes many novel regulatory proteins and biosynthetic gene clusters. Further studies of these newly identified proteins and functional pathways have the potential to improve our understanding of many diverse biological processes and to provide new opportunities for medicine and industrial applications.



Philip Patenall/Macmillan Publishers Limited

Improvements in sequencing technologies have also revealed unprecedented levels of diversity in microbial communities. Although 16S ribosomal RNA (rRNA) sequencing provided important insights into community diversity, this approach is frequently unable to achieve biologically meaningful resolutions. Methods for sequencing the entire DNA from microbial communities (metagenomic sequencing) have the capacity to provide strain level resolution; however, these approaches are regularly hindered by the numerous unsequenced microbial taxa that are contained within the samples. Approaches such as *de novo* assembly and metagenomic species analysis have been developed to computationally determine the taxonomic composition of these datasets[4]. Although these approaches can identify abundant species in the sample, it is often challenging to identify less abundant species and strain differentiation remains problematic. The increasing number of phylogenetically diverse genome datasets will make genome-guided metagenomic analysis increasingly more feasible and the preferred method for researchers.

Research into microbial communities is progressing from genotype–phenotype association studies to include experimental validation of bacterial species and functions. High-quality genome sequences of reference strains will be essential to achieve the high-resolution sequence-based species identification that is needed to support this work. The availability of these genome sequences will enable the refinement of computational models and support experimental validation of computationally predicted biological functions. Analyses of these reference genomes could also provide novel insights into the genetic basis for phenotypic characteristics. In time, this expanded collection of genome sequences may also provide a basis to advance metagenomic studies to include experimental validation of microbial communities and predicted functions.

Samuel C. Forster is at the Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK and the Hudson Institute of Medical Research, Clayton, Victoria 3168, Australia.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro.2017.106
Published online 30 Aug 2017

1. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
2. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
3. Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).
4. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).

Competing interests statement
The author declares no competing interests.