GENOME WATCH

Adding genomic 'foliage' to the tree of life

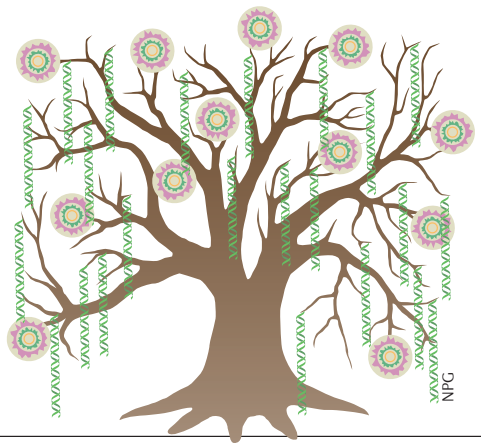
Alan Walker

This month's Genome Watch highlights recent studies that used metagenomics and single-cell genomics to gain insights into previously uncultivated and poorly characterized microbial lineages.

At least 60 bacterial or archaeal phyla are currently recognized on the basis of small subunit ribosomal RNA or metagenomic data from environmental surveys. However, many phyla have either no or very few cultured representative species, as current culture collections are biased towards microorganisms of medical or industrial importance. Therefore, our understanding of microbial life has large gaps. The Genomic Encylopaedia of Bacteria and Archaea Project, which improved reference databases by sequencing genomes from a targeted, phylogenetically diverse array of cultured isolates[1], was an important advance. However, by focusing on cultured isolates, it did not include microorganisms that have never been grown in the laboratory, which represent the great majority of all species. Single-cell genomics and metagenomics offer two approaches to address this problem: with single-cell genomics, individual microbial cells are isolated from environmental samples and their DNA is amplified and then sequenced, whereas with metagenomics, DNA is extracted directly from environmental samples, shotgun-sequenced and then assembled.

Rinke *et al.*[2] used single-cell genomics to characterize 201 uncultivated cells, which were derived from 29 mostly undefined bacterial and archaeal lineages that were isolated from nine environmental habitats. The average genome coverage of the isolated cells was around 40%, which reflects a bias that was introduced during DNA amplification — a major limitation of single-cell genomics. Nonetheless, this study is the first sizeable characterization of several candidate phyla, helped to resolve the phylogenetic placement of the isolated cells and revealed many examples of potential lateral gene transfer across domains. The genomes also yielded nearly 20,000 new hypothetical protein families, which increases the total number of identified families by more than 8%. An additional benefit of this novel genomic data is that it broadened the diversity of reference databases, which enabled the authors to improve the phylogenetic placement of sequence data from metagenomic studies.

Kantor *et al.*[3] recently used metagenomics to recover genomes from four uncultivated candidate phyla — SR1, WWE3, OD1 and TM7 — that were present in aquifer sediment. Metagenomics does not require DNA amplification prior to sequencing, meaning that the amplification bias that limits single-cell genomics can be avoided. As a result, the authors were able to construct complete genomes for three species and an almost complete genome for the fourth. As with Rinke *et al.*[2], these genomes yielded many novel hypothetical protein families. A key finding was that all of the genomes were small, ranging from 0.7 to 1.17 Mb, and lacked many common biosynthetic pathways. This suggests that these cells rely on other organisms in the environment for growth. In support of this, all four genomes encode type IV pili — which enhance mobility, adhesion and biofilm formation — as well as a number of other factors, such as nucleases, proteases and transporters, that might be useful for scavenging resources from the environment.



A limitation of metagenomics is that it can be challenging to assemble complete genomes from rare members of microbial communities, and there are also concerns that sequences from multiple divergent species can be wrongly incorporated into single assemblies. Albertsen *et al.*[4] recently developed an improved data-assignment approach that is based on the cross-comparison of metagenomic sequence data from the same sample prepared in two different ways, and they used it to assemble 31 genomes derived from an activated sludge bioreactor. Four of these genomes were from the uncultivated candidate phylum TM7. As with the TM7 genome that was generated by Kantor *et al.*[3], they showed that the average genome size was small at just 1 Mb. The authors therefore propose that compact genomes are a common feature of the TM7 phylum.

The approach that is outlined by Albertsen *et al.*[4] is just one of many improvements in the rapidly developing fields of metagenomics and single-cell genomics. With these techniques, it is now possible to better understand all branches of the tree of life, even those that have been most reluctant to grow in the laboratory and reveal their secrets.

Alan Walker is at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro3203

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. Kantor, R. S. *et al.* Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* **4**, e00708-13 (2013).
4. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnol.* **31**, 533–538 (2013).

Competing interests statement
The author declares no competing financial interests.