microbiology Comment

Two novel methods for using genome sequences to infer taxonomy

Next generation sequencing (NGS) is catalysing a host of new developments across microbiology. Two papers recently published in *Microbiology* (Jolley *et al.*, 2012; Bennett *et al.*, 2012) describe methods that exploit NGS genome data to classify bacterial genomes based on core gene sequences. In general, these methods agree with 16S rRNA phylogenetic trees, but the novel methods have the added advantage of providing strain resolution within a given species. Furthermore, these approaches are scalable for large numbers of genomes, do not depend on a reference genome and can use as input genomes from different formats: finished sequences or genome assemblies in multiple contigs. Both papers focus on a set of 'core genes' to use; 53 genes encoding ribosomal proteins in the case of ribosomal multi-locus sequence typing (rMLST; Jolley *et al.*, 2012) or a set of core genes defined through comparative genomics (Bennett *et al.*, 2012).

The introduction of MLST for strain identification (Maiden *et al.*, 1998) provided the first sequence-based approach to strain resolution for many bacterial species. However, with the advent of inexpensive whole-genome sequencing technologies that now allow sequencing a bacterial genome for close to the same price as sequencing the seven or so genes for MLST, many have wondered about expanding the set of genes to be used and indeed which genes might be optimal (for an example see Leekitcharoenphon *et al.*, 2012). The rMLST method (Jolley *et al.*, 2012) uses 53 genes encoding bacterial ribosomal proteins, which are found in nearly all bacteria. rMLST provides combined taxonomy and typing data, which has obvious advantages. The authors conclude that 'the ribosome occupies the interface between genotype and phenotype that is a required focus of microbiology in the post-genomic era of research', and hence the choice of using ribosomal proteins is a logical extension for typing.

The second approach (Bennett *et al.*, 2012) focussed on the genus *Neisseria*, which contains some members that can be difficult to classify by using 16S rRNA. A set of 246 genes was found to be conserved across all the 55 *Neisseria* genomes in the database, and these core genes were used to construct a tree for the sequenced *Neisseria* strains. There were seven groups, consistent with other known data. The authors also propose that in some cases, the current names of the organisms are not consistent with their distance, based on their genome sequence. The resulting core gene tree is robust and is similar to that found by just using the 53 ribosomal proteins as in the rMLST method. However, the additional information obtained by knowing the set of 'core genes' as well as the variable 'accessory genes' for a given set of organisms can be quite useful in better understanding their underlying biology. Again, the advantage for both methods is that they will allow rapid, reproducible classification of bacterial groups based on genome sequences.

These methods offer novel ways to study the ever-increasing breadth and depth of bacterial genome data available to us; an obvious application is to metagenomic analyses of bacterial communities. The underpinning infrastructure provided by the Bacterial Isolate Genome Sequence Database (BIGSdb; http://pubmlst.org/software/database/bigsdb/) provides a ready platform for users to interrogate allele definitions and strain data (Jolley & Maiden, 2010). The rMLST and core gene approaches will be useful tools for mining the bacterial genome data mountain.

David W. Ussery[1] and Stephen V. Gordon[2]

[1]Microbial Genomics Group, Technical University of Denmark, Center for Biological Sequence Analysis, BioCentrum-DTU, Bld 208, Lyngby DK-2800, Denmark

[2]Schools of Veterinary Medicine, Medicine and Medical Science, Biomolecular and Biomedical Science, UCD Conway Institute, University College Dublin, Ireland

Correspondence: David W. Ussery (dave@cbs.dtu.dk)

Bennett, J. S., Jolley, K. A., Earle, S. G., Corton, C., Bentley, S. D., Parkhill, J. & Maiden, M. C. (2012). A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology* **158**, 1570–1580.

Jolley, K. A. & Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K. & other authors (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**, 1005–1015.

Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M. & Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* **13**, 88.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140–3145.

DOI 10.1099/mic.0.059816-0