GENOME WATCH

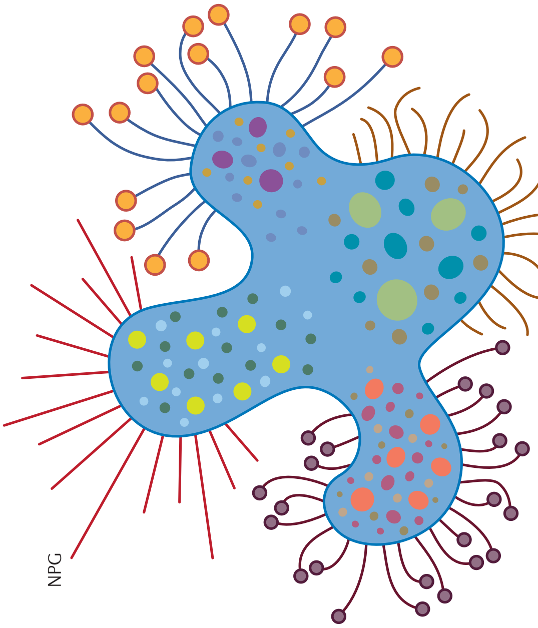
Species Mash-up

Silvia Argimón and David M. Aanensen

This month's Genome Watch describes how Mash can be used to tackle comparisons between large amounts of genomic and metagenomic sequence data for taxonomic applications.

The identification of the causative agent of a disease remains a fundamental premise in the field of infectious disease research and epidemiology. Originally, Koch's postulates proposed criteria to link a specific microorganism to a particular disease, and with great advances at the genetic, molecular and microbial community levels, the revised molecular version aimed to provide guidelines to identify particular genes that are involved in bacterial virulence[1], and, more recently, dysbiosis has been implicated in disease causation. However, whether from bacterial isolates or complex microbial populations, the organism or organisms that are associated with disease need to be taxonomically identified.

Classic clinical diagnostics based on phenotypic and biochemical properties yielded PCR-based methods, which, in turn, were outcompeted by 16S rRNA sequencing[2], a technique that has dominated the field in recent years. With the recent outpour of whole-genome sequences, methods that looked at the similarity over the entire sequence were developed, which either rely on k-mer counts or string matching. These are computationally challenging, because of the large volumes of sequence data. This issue was recently addressed using Mash[3], which is a toolkit that implements the MinHash technique to reduce large sequence sets to compressed sketch representations. MinHash was originally developed to quickly estimate the similarity between two datasets, and has been used in several areas of computer science; for example, in web search engines (remember AltaVista?) for the determination of very similar webpages and their removal from search results.



NPG

Mash can rapidly match genome sequences to a RefSeq sketch, both from raw reads and assemblies. Furthermore, reference sketches can be defined by a user, which enables tailored searches. Novel and useful features of Mash include a significance test to account for chance matches and the Mash distance metric, which estimates the mutation rate between two sequences directly from their MinHash sketches and which correlates with the average nucleotide identity (ANI), an alignment-based metric. An ANI of 95% is currently accepted as a threshold to assign strains to the same species based on core genomic sequences. The authors determined that this is equivalent to a Mash distance of 0.05 for *Escherichia coli*, which indicates the potential application of Mash for taxonomic assignment.

This application is further demonstrated by the authors by the progress towards strain-level identification. Although much work is required in this respect, Mash offers a tantalising glimpse into the scale and speed that could be applied for the identification of pathogenic lineages. A hybrid approach of quick binning, through Mash, into species, with finer-scale comparisons between similar lineages would enable a much faster identification of genomic distance for outbreak analysis and global surveillance. Another application could address the untangling of mixed samples in microbial genomic sequencing projects: comparing reads to reference sketches could fine-tune sequence data filtering prior to phylogenetic inference.

Mash can also cluster metagenomic sequences, with the potential for future metagenomic sequence classification. This could lead to its application for the culture-independent identification of pathogens from clinical samples, which would enable the identification of multiple pathogenic species and/or strains simultaneously.

Although every method has its limitations, the development of Mash represents an important step the towards rapid, scalable binning of sequences into boundaries that are represented by reference sequences. However, species designations and genetic distance are not always perfectly correlated, and the issue of whether these species boundaries are consistent with realistic biological boundaries is a topic for another discussion.

Silvia Argimón and David M. Aanensen are at
The Centre for Genomic Pathogen Surveillance,
The Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambridge CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro.2016.175
Published online 11 Nov 2016

1. Falkow, S. Molecular Koch's postulates applied to bacterial pathogenicity — a personal recollection 15 years later. *Nat. Rev. Microbiol.* **2**, 67–72 (2004).
2. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA* **82**, 6955–6959 (1985).
3. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

Competing interests statement
The authors declare no competing interests.