Key variables affecting genetic distance calculations in genomic epidemiology

Published Online
August 6, 2021
https://doi.org/10.1016/S2666-5247(21)00183-X
See Online/Articles
https://doi.org/10.1016/S2666-5247(21)00149-X

The benefits of using whole-genome sequencing for outbreak detection and surveillance of bacterial pathogens have been well documented in recent years.[1,2] However, the bioinformatics methods used to analyse and interpret genomic data need to be standardised and validated if genomic epidemiology is to be widely implemented and routinely used in public health investigations. This means that genome analysis methods should yield robust, reproducible and generalisable results regardless of the geographical origin, laboratory, or bacterial population of study.

Key to transmission inference is the determination and interpretation of genetic distances from bacterial genomic data. The simplest way to establish genetic relatedness between bacterial isolates is to count the number of nucleotide differences (ie, the number of single-nucleotide polymorphisms [SNPs]) between their genome sequences. Genetic relatedness thresholds have been proposed above which recent bacterial transmission can be ruled out, while distances below indicate probable transmission. It is increasingly acknowledged that epidemiological follow-up is needed to confirm definite transmission.

In their study published in *The Lancet Microbe*, Claire Gorrie and colleagues[3] quantified the effect of key methodological variables on the calculation of SNP distances, which in turn influenced how many isolates fell below specific relatedness thresholds. For this investigation, they used a large collection of multidrug-resistant pathogens sourced from a 15-month prospective study done in eight hospitals in Melbourne (VIC, Australia). They included isolates from four major nosocomial pathogens: meticillin-resistant *Staphylococcus aureus*, vancomycin-resistant *Enterococcus faecium*, and extended-spectrum β-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae*.

The authors assessed the effect of three key variables: the inclusion or omission of prophage and recombination regions, the choice of the reference genome, and the number and overall diversity of samples being compared at any given time.

It is common practice to detect and mask recombinogenic regions before calculating SNP distances, to restrict the analysis to SNPs arisen by point mutation only, which are known to accumulate at a constant rate, and avoid including regions with high density of SNPs brought by homologous recombination. Of note, the authors recommend not to mask recombination. In cases where isolates were closely related, which are the ones most likely to be recently transmitted, masking recombination had little or no effect on pairwise SNP distances. In line with this observation, another study[4] published in 2021 found that recombination rarely contributed to SNP differences among recently diverged *Enterococcus faecium* isolates (differing by 0 to 20 whole-genome SNPs), and that it became a bigger contributor of SNP differences as strains diverged. Gorrie and colleagues also noted that scanning for and masking prophages had no extra benefit, as prophage regions were enclosed within regions already detected as recombination or already excluded from analysis because they were not part of the core-genome.

The authors noted that using a reference genome that is closely related to the group of isolates being compared resulted in bigger core-genome alignments (ie, the fraction of the genome shared by all isolates and used to count SNPs), providing a greater degree of resolution. Consistent with these findings, a study evaluating the performance of SNP-calling pipelines found that, irrespective of pipeline, a major determinant for reliable SNP calling was the choice of the reference genome.[5]

Gorrie and colleagues also observed that the greater the number and diversity of strains being compared over time, the smaller the core-genome alignments became, with concomitant reductions in resolution. To overcome this limitation, they recommended a so-called sliding window approach, by which only isolates sampled within a given time window (ie, 3 months) were compared, in order to maintain longer and more stable alignment sizes over time.

Alternative approaches to obtain stable and generalisable SNP distances have been proposed. One option consists in calling SNPs on the species core genome, which is the fraction of the genome shared by all, or the vast majority, of strains of the same species.

The advantage of using the species core genome is that the length of alignments remains constant regardless of the number and diversity of strains being compared, making SNP distances directly comparable.[6,7] The main limitation is that a significant portion of the genome (ie, the accessory genome) is ignored in comparisons, limiting the resolution that can be achieved.

To overcome the limitations of reference-based and core genome approaches, direct whole-genome comparisons can be done instead. One such approach would consist in reconstructing the whole genome of individual isolates by de-novo assembly and using the resulting draft assemblies as references to map the reads of other isolates against, to ultimately obtain pairwise SNP distances. The length of genome portions used in comparisons would need to be considered to compute normalised SNP distances that are comparable. The use of k-mers has also been proposed to calculate pairwise genetic distances,[8] with the advantage of detecting variations in both the core and accessory genomes.

Beyond the use of genetic distances, probabilistic approaches that combine genetic and spaciotemporal data[9] and clustering phylogenetic methods[10] have also been proposed to identify putative transmission chains. Regardless of the approaches used, methods will need to be benchmarked. Robust retrospective datasets of well characterised outbreaks could be used to validate methods. Factors such as computational resources and facility of interpretation by non-experts will also determine what methods become ultimately established in genomic epidemiology.

FC reports funding from the Wellcome Trust (LSHTM/Wellcome Institutional Strategic Support Fund Fellowship [204928/Z/16/Z] and Wellcome Trust Sir Henry Postdoctoral Fellowship [201344/Z/16/Z]); consultancy fees from Next Gen Diagnostics LLC; and payment from Wellcome Genome Campus Advanced Courses for the development of FutureLearn course "Bacterial Genomes: Antimicrobial Resistance in Bacterial Pathogens".

Francesc Coll
francesc.coll@lshtm.ac.uk

London School of Hygiene & Tropical Medicine, Faculty of Infectious and Tropical Diseases, Department of Infection Biology, London WC1E 7HT, UK

1 Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* 2017; **19:** 9–20.
2 Tang P, Croxen MA, Hasan MR, Hsiao WW, Hoang LM. Infection control in the new age of genomic epidemiology. *Am J Infect Control* 2017; **45:** 170–79.
3 Gorrie CL, Gonçalves Da Silva A, Ingle DJ, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. *Lancet Microbe* 2021; published online Aug 6. https://doi.org/10.1016/S2666-5247(21)00149-X.
4 Gouliouris T, Coll F, Ludden C, et al. Quantifying acquisition and transmission of Enterococcus faecium using genomic surveillance. *Nat Microbiol* 2021; **6:** 103–11.
5 Bush SJ, Foster D, Eyre DW, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *Gigascience* 2020; **9:** 1–21.
6 Coll F, Raven KE, Knight GM, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant *Staphylococcus aureus*: a genomic epidemiology analysis. *Lancet Microbe* 2020; **1:** e328–35.
7 Aggelen HV, Kolde R, Chamarthi H, et al. A core genome approach that enables prospective and dynamic monitoring of infectious outbreaks. *Sci Rep* 2019; **9:** 7808.
8 Harris SR. SKA: split kmer analysis toolkit for bacterial genomic epidemiology. *bioRxiv* 2018; published online Oct 25. https://doi.org/10.1101/453142 (preprint).
9 Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *Mol Biol Evol* 2019; **36:** 587–603.
10 Lees JA, Harris SR, Tonkin-Hill G, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* 2019; **29:** 304–16.