GENOME WATCH

Tracing outbreaks with machine learning

Nicole E. Wheeler

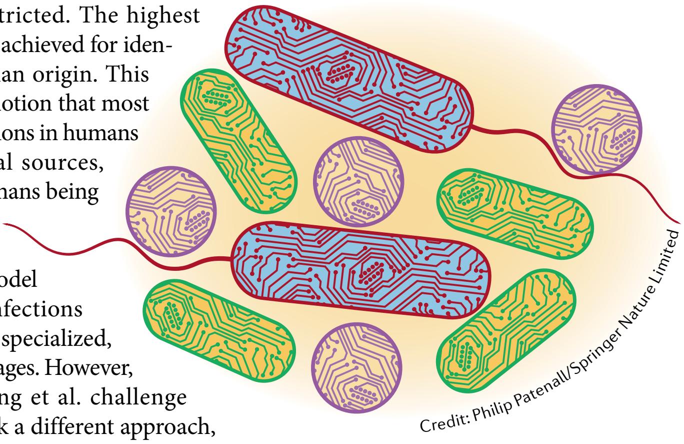
This Genome Watch article discusses the application of machine learning algorithms to predict the source of food-borne infections.

Tracing the source of food-borne disease is challenging. Outbreak investigations require detailed epidemiological analysis to trace infections to a common source. The routine use of whole-genome sequencing for the surveillance of food-borne illness is extending to more organizations[1], which provides an opportunity to leverage the data to better attribute cases of food poisoning. A major cause of food poisoning worldwide is *Salmonella enterica* subsp. *enterica* serovar Typhimurium. This serovar has a broad host range and can persist in various environments; however, some sublineages are more commonly associated with particular hosts[2]. Recently, two studies have used machine learning to identify *S.* Typhimurium molecular markers that are associated with different hosts, which could be used to trace the source of infections.

In 2017, Lupolova et al.[3] trained a support vector machine to predict host of origin based on the accessory genome content of *S.* Typhimurium. Their model was able to distinguish human, avian, swine and bovine strains with an overall accuracy of 81%. Recently, Zhang et al.[4] addressed the problem of source prediction using a different data set and machine learning algorithm. To test their algorithm, they gathered *S.* Typhimurium genomes from a wide range of sources, including isolates from eight zoonotic outbreaks. They trained a random forest model on core genome variation and accessory genome content of only the animal isolates, and then tested it on all isolates. The algorithm achieved 83% accuracy, and attributed 7 out of 8 zoonotic outbreaks to the correct source.

Despite the similar accuracy of the two models, they yielded contradictory conclusions. The model built by Lupolova et al. predicted that a surprising proportion of their strains were host-restricted. The highest accuracy (at 90%) was achieved for identifying strains of human origin. This seems to counter the notion that most *S.* Typhimurium infections in humans originate from animal sources, with adaptation to humans being associated with more severe disease outcomes[2]. Instead, the model suggests that these infections originate mostly from specialized, human-associated lineages. However, observations by Zhang et al. challenge this finding. They took a different approach, assuming that human isolates were likely to be zoonoses and therefore did not include humans as a possible source. Their algorithm confidently attributed one-third of human strains to a specific animal. To better compare the two approaches, they built another predictor that included humans as a source, which still assigned only one-third of human strains to human origins. Zhang et al. argue that this discrepancy is due to differences in strategies used to build their strain collections, specifically the care they took to exclude closely related isolates that were epidemiologically linked. Roughly one-third of the human isolates in their data set shared a nearest neighbour with another human strain, compared to 85% of the isolates in Lupolova et al.'s training data, consistent with the difference in prediction accuracies. Further, Zhang et al. found that by simply matching isolates to their nearest neighbour in the data set, around 70% of animal isolates, and 6 out of 8 zoonoses, could be correctly attributed.

These studies illustrate ongoing progress in building machine learning algorithms that could inform public health. However, they also point to the importance of data set design and the influence that clonally related samples sharing a phenotype can have on the patterns identified by machine learning models[5]. As described above, study design has affected the conclusions drawn. In one study the model suggested that human isolates were predominantly restricted to the human niche, whereas the model from the other study contradicts this conclusion, suggesting mainly zoonotic sources. The application of machine learning methodologies in bacterial genetics is a rapidly growing field. As approaches in this area mature, they may be applied to predict the source of food-borne disease outbreaks.



Credit: Philip Patenall/Springer Nature Limited

Nicole E. Wheeler
Centre for Genomic Pathogen Surveillance, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK.
e-mail: microbes@sanger.ac.uk

https://doi.org/10.1038/s41579-019-0153-1

1. Allard, M. W. et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* **54**, 1975–1983 (2016).
2. Branchu, P., Bawn, M. & Kingsley, R. A. Genome variation and molecular epidemiology of *Salmonella enterica* Serovar Typhimurium pathovariants. *Infect. Immun.* **86**, e00079–18 (2018).
3. Lupolova, N. et al. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* **3**, e000135 (2017).
4. Zhang, S. et al. Zoonotic source attribution of *Salmonella enterica* serotype Typhimurium using genomic surveillance data, United States. *Emerg. Infect. Dis.* **25**, 82–91 (2019).
5. Falush, D. Bacterial genomics: microbial GWAS coming of age. *Nat. Microbiol.* **1**, 16059 (2016).

Competing interests
The author declares no competing interests.