Combined visualization of genomic and epidemiological data for outbreaks

Carl J. E. Suster[1,2,3] 🔟, Anne E. Watt[2,3,4], Qinning Wang[1,4], Sharon C.-A. Chen[1,2,3,4], Jen Kok[1,2,3,4] and Vitali Sintchenko[1,2,3,4]

[1]Centre for Infectious Diseases and Microbiology – Public Health, Westmead Hospital, Westmead, NSW, Australia; [2]Sydney Infectious Diseases Institute, The University of Sydney, Sydney, NSW, Australia; [3]School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia and [4]Centre for Infectious Diseases and Microbiology Laboratory Services, Institute of Clinical Pathology and Medical Research, NSW Health Pathology, Westmead, NSW, Australia

Abstract

In epidemiological investigations, pathogen genomics can provide insights and test epidemiological hypotheses that would not have been possible through traditional epidemiology. Tools to synthesize genomic analysis with other types of data are a key requirement of genomic epidemiology. We propose a new 'phylepic' visualization that combines a phylogenomic tree with an epidemic curve. The combination visually links the molecular time represented in the tree to the calendar time in the epidemic curve, a correspondence that is not easily represented by existing tools. Using an example derived from a foodborne bacterial outbreak, we demonstrated that the phylepic chart communicates that what appeared to be a point-source outbreak was in fact composed of cases associated with two genetically distinct clades of bacteria. We provide an R package implementing the chart. We expect that visualizations that place genomic analyses within the epidemiological context will become increasingly important for outbreak investigations and public health surveillance of infectious diseases.

Report

High-resolution characterization and clustering of pathogen genomes using whole genome sequencing (WGS) has proven to be a powerful complementary tool to investigations of infectious disease outbreaks [1, 2]. An epidemiological investigation will often involve establishing a case definition to demarcate the outbreak, geospatial and temporal clustering, case follow-up and further investigations to establish transmission pathways or common reservoirs [3]. Genomic epidemiology draws together bioinformatic analysis from pathogen genomic sequences and epidemiological data to provide context and support to inferences of transmission. Synthesis of genomic and epidemiological data is required to perform these inferences, however, this can be challenging given the complexity of the underlying data [4]. These challenges can be compounded when genomic and epidemiological investigations are conducted by discrete organizations (e.g., public health laboratories and local public health services respectively) with different information contexts and objectives.

The epidemic curve, a histogram of cases binned along the time axis, is a key representation relied upon in disease control. It is straightforward to prepare from line listing data, supports both data exploration and statistical epidemiology [5], and is straightforward to interpret. Genomic epidemiology instead relies upon phylogenomic trees, which display a detailed summary of the relationships between the included genomes under the particular assumptions of the underlying evolutionary models. Phylogenomic trees are prepared using specialized tools that implement the evolutionary models and efficiently search for the optimal tree (e.g., IQ-TREE, http://www.iqtree.org/). Trees are an essential representation of the salient information contained in a set of genomes that support genomic epidemiology and public health surveillance. Visually extracting this information relies upon knowledge of the conventions of tree drawing, and some familiarity with the process of their preparation. New users of trees may therefore require dedicated training so as to minimize the risk of misinterpretation.

We propose a new visualization called the 'phylepic' chart that synthesizes epidemic curves and phylogenomic trees. The structure of the phylepic chart facilitates the visual linkage of the molecular time represented in the tree with the epidemiological time represented in the epidemic curve [6]. For illustration, we adapted a dataset from the investigation of a point-source foodborne outbreak in New South Wales (NSW), Australia. Clinical cases presenting with gastroenteritis were epidemiologically linked to the same food product, sparking an investigation. Bacterial isolates were collected during the outbreak period from clinical samples, food samples, and food manufacturing facilities, and all underwent WGS. The resulting sequences were analysed alongside historical sequences collected in NSW using conventional bioinformatic

pipelines to produce a core genome phylogenomic tree. For the purpose of illustration, we have pruned the resulting phylogeny and edited some details in the scenario. The chart is presented in Figure 1.

Relying on the epidemiological data alone, the silhouette of the epidemic curve in Figure 1 is consistent with a single point-source outbreak. The epidemic curve is coloured by genomic cluster membership. The presence of two genomic clusters suggests that there are two distinct strains associated with cases occurring in the same period. The phylogenomic tree provides the more complete genomic context of the outbreak by providing detailed information about the inferred relatedness of bacterial sequences. The tree reveals that the most recent common ancestor of the two outbreak clusters (Clust-04 and Clust-05) is also the common ancestor of several other clusters and singletons in the historical dataset. This provides evidence that Clust-04 and Clust-05 arose from separate introductions of the pathogen into the food product. From the tree alone the epidemiological link between the two clusters is not

evident and there is no representation of the temporal proximity of cases. Only the combination of the two lines of evidence leads to the conclusion of distinct introductions to the same food product.

The calendar grid in the lower right of Figure 1 shows a visual projection of the tree onto a time axis representing the date of sample collection. For this particular chart, the time axis was restricted to the investigation period, meaning that for isolates collected before mid-August, there is no calendar tile. To distinguish these from cases with missing metadata, triangles are drawn at the left edge of the calendar to indicate out-of-range values. The epidemic curve and calendar grid are binned by epidemiological week. The phylepic chart is most readable when the tree is relatively small and there are not too many date bins, both of which are reasonable restrictions for the intended application in outbreak investigation. The bin width can be decreased to a single day or increased to another interval depending on the appropriate resolution for an investigation.
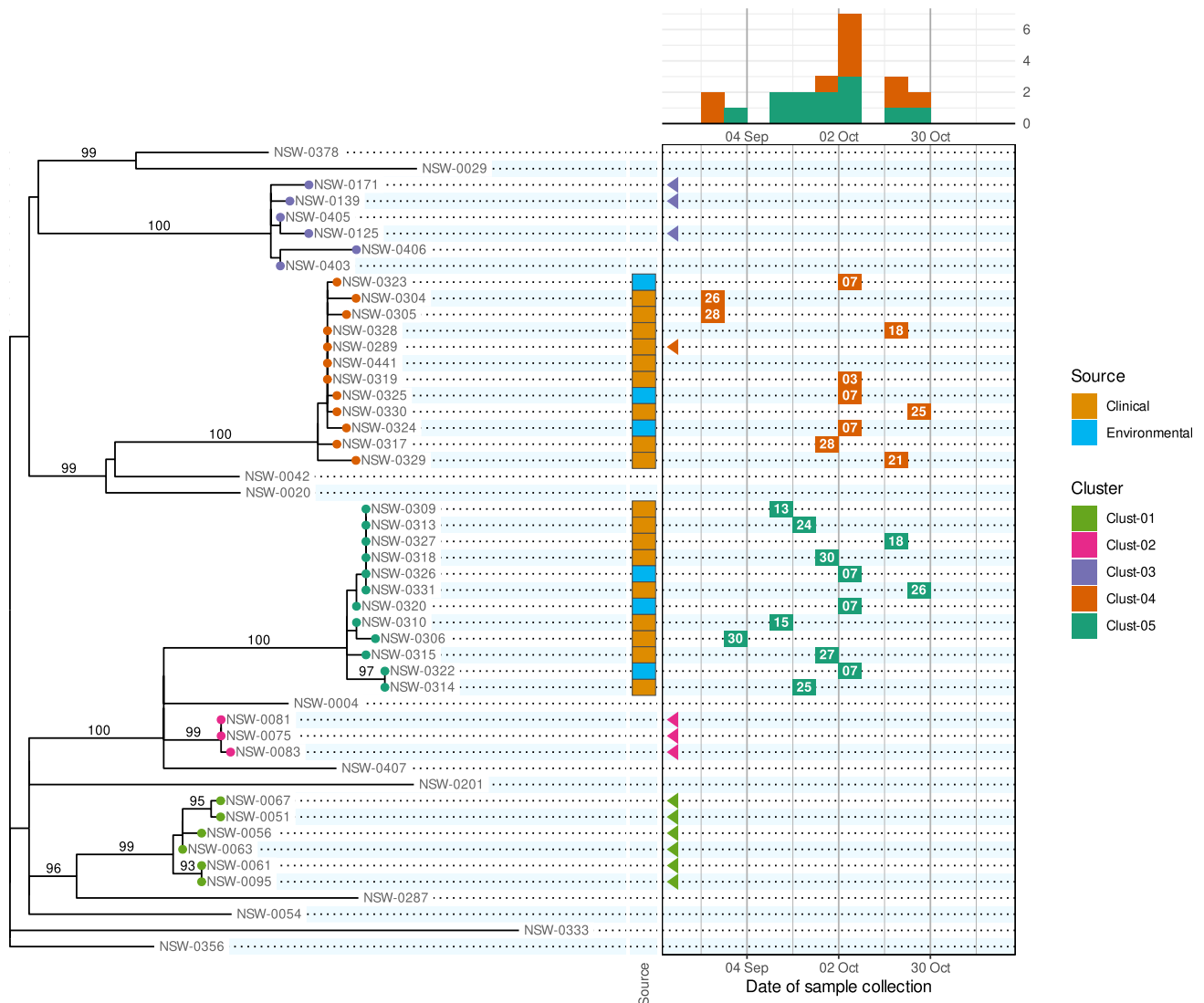

Figure 1. A phylepic chart using an illustrative dataset adapted from a foodborne enteric pathogen outbreak in NSW. The panel on the left shows the phylogenomic tree with tips coloured by genomic cluster membership. Numbers on branches are bootstrap support values expressed as percentages. The panel on the upper right shows the epidemic curve using the date of sample collection. The calendar grid (boxed) in the lower right connects the tree to the epidemic curve, and each tile shows the day of the month of the sample collection date. Samples with a collection date prior to the outbreak period are shown with triangles at the left edge of the date scale. Colours on tree tips, calendar tiles, and the epidemic curve refer to the same genomic clusters.

While we have chosen a foodborne bacterial pathogen for this example, phylepic charts are equally compatible with any other pathogens such as respiratory viruses. The tree need not necessarily be built from the consensus genome or a core genome as in our example; it is sufficient that a tree of some sort is prepared that is meaningful in the context of the investigation. It remains important to convey any caveats and limitations of the phylogenomic analysis alongside the results to ensure correct interpretation.

To communicate relevant epidemiological context, the most popular tools for visualizing trees (e.g., the ggtree R package, https://yulab-smu.top/treedata-book/, or the ETE toolkit Python library, http://etetoolkit.org/) allow tips to be annotated with an array of coloured boxes or symbols. This can work well for categorical or binary data such as the isolation source of samples, results of laboratory assays, or the presence of antimicrobial resistance, toxins, or virulence factors. These annotations can be incorporated into the phylepic chart, as shown with the coloured tiles indicating clinical and environmental sample sources in Figure 1. Additional columns can be incorporated to represent other categorical or quantitative data as required. With some care, it is possible to use existing software libraries such as those mentioned above to produce a chart similar to Figure 1.

Dates associated with each case (reporting, sample collection, symptom onset) are key data for generating epidemiological hypotheses. These dates do not directly relate to the branch lengths in phylogenomic trees. Instead, the axis extending from the root of the tree to its tips can be thought of as molecular time, reflecting the evolutionary distances predicted by the model. It is possible to express branch lengths in units of time using an inferred molecular clock rate, and by incorporating temporal information as explicit constraints in the model [7]. With a time-scaled tree of this sort, an alternative visualization is made possible where an epidemic curve is vertically aligned with the tree [6]. This can be helpful to show how population-level dynamics over a longer period correspond to structural features of the tree, however, it is less suitable for high-resolution outbreak investigation since the dates associated with individual cases are not shown.

We have implemented the phylepic chart in an R package. Our package uses the ggplot2 library [8] as a framework, the ggraph library (https://ggraph.data-imaginist.com) for drawing the tree, and the cowplot library (https://wilkelab.org/cowplot) for aligning panels. The package is designed so that minimal code is required to implement visualizations such as Figure 1, while full customisation and annotation of the plots are possible using the underlying frameworks. Manual curation and careful consideration of how data are represented remain important in crafting an effective visualization of the data. The package takes as input a prebuilt phylogeny (in any commonly used tree format such as Newick) and tabular data containing at minimum a column of dates associated with each sequence in the tree. Compared to existing software libraries for annotating phylogenetic trees, the package implements routines for date handling (including automatic binning by week) and ensuring data matching and consistency across panels.

As genomic epidemiology becomes a routine part of public health outbreak investigations, effective integration of genomic and epidemiological evidence to support collaboration between relevant public health and clinical teams remains a key challenge. Communication and visualization tools will assist in conveying complex genomic data into investigation and response contexts [9, 10]. Our experience with using phylepic charts in public health reporting has shown that they can help epidemiologists and other public health professionals to accurately interpret phylogenomic trees by relating genomic features to more familiar epidemiological details of the cases under investigation.

References

[1] **Armstrong GL**, et al. (2019) Pathogen genomics in public health. *New England Journal of Medicine Massachusetts Medical Society* **381**, 2569–2580.

[2] **Baker KS** (2020) Microbe hunting in the modern era: Reflecting on a decade of microbial genomic epidemiology. *Current Biology* **30**, R1124–R1130.

[3] **World Health Organization** (2021) *Framework and Toolkit for Infection Prevention and Control in Outbreak Preparedness, Readiness and Response at the National Level*. Geneva: World Health Organization.

[4] **Watt AE**, et al. (2022) State-wide genomic epidemiology investigations of COVID-19 in healthcare workers in 2020 Victoria, Australia: Qualitative thematic analysis to provide insights for future pandemic preparedness. *The Lancet Regional Health – Western Pacific* **25**, 100487.

[5] **Nishiura H** (2016) Methods to determine the end of an infectious disease epidemic: A short review. In Chowell G and Hyman JM (eds), *Mathematical and Statistical Modeling for Emerging and re-Emerging Infectious Diseases*. Cham: Springer International Publishing, pp. 291–301.

[6] **Sintchenko V and Holmes EC** (2015) The role of pathogen genomics in assessing disease transmission. *BMJ* **350**, h1314.

[7] **Sagulenko P**, **Puller V and Neher RA** (2018) TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**, vex042.

[8] **Wickham H** (2016) *ggplot2: Elegant Graphics for Data Analysis*, 2nd Edn. Cham: Springer International Publishing .

[9] **Arnott A**, et al. (2021) Documenting elimination of co-circulating COVID-19 clusters using genomics in New South Wales, Australia. *BMC Research Notes* **14**, 415.

[10] **Argimón S**, et al. (2016) Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics Microbiology Society* **2**, e000093.