Exploration or exploitation? Genome entropy and network centrality delineate pathogen evolution

Sheryl L. Chang [a] , Rebecca Rockett[b,c], Carl Suster[b,c], Adam J. Svahn[a,b], Oliver M. Cliff[d], Alicia Arnott[c], Qinning Wang[c,e], Rady Kim[e], Marc Ramsperger[e], Mailie Gall[e], Tania C. Sorrell[b,c], Vitali Sintchenko[b,c,e], Mikhail Prokopenko[a,d]

[a] *Centre for Complex Systems, Faculty of Engineering, The University of Sydney, Sydney, 2006, Australia*
[b] *The University of Sydney Institute of Infectious Diseases, The University of Sydney, Westmead, NSW, Australia*
[c] *Centre for Infectious Diseases and Microbiology–Public Health, Westmead Hospital, NSW, Australia*
[d] *School of Physics, Faculty of Science, The University of Sydney, Sydney, NSW, Australia*
[e] *NSW Enteric Reference Laboratory, Institute of Clinical Pathology and Medical Research NSW Health Pathology, Westmead, NSW, Australia*

Email (correspondence): sheryl.chang@sydney.edu.au

Modelling the evolutionary dynamics of foodborne pathogens is important in mitigation and prevention of epidemics. Here, we study 5-year surveillance data on Salmonella Typhimurium (STM) in New South Wales (NSW), Australia, applying network science and information theory to trace STM evolutionary pathways.

STM is a prominent serotype responsible for most of the infections acquired locally in NSW [1]. Public health surveillance undertakes routine whole-genome sequencing (WGS) from bacterial isolates. This study uses longitudinal WGS data for 3,939 STM isolates collected between 2016 and 2021 in NSW.

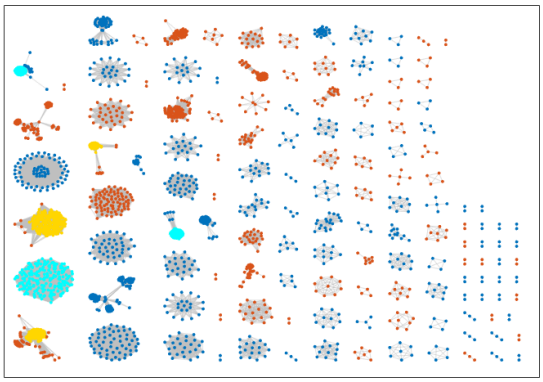

Figure 1. The undirected thresholded SNP subnetwork produced for the complete undirected SNP network using SNP distance threshold G_{max}=20 (N=3,939; M=318,654). Singletons (230 isolates) are removed. Total number of components: 135 (excluding singletons). Node colours: (blue) low-to-mid centrality with low prevalence; (cyan) low-to-mid centrality with high prevalence; (orange) mid-to-high centrality with low prevalence; (yellow) mid-to-high centrality with high prevalence.

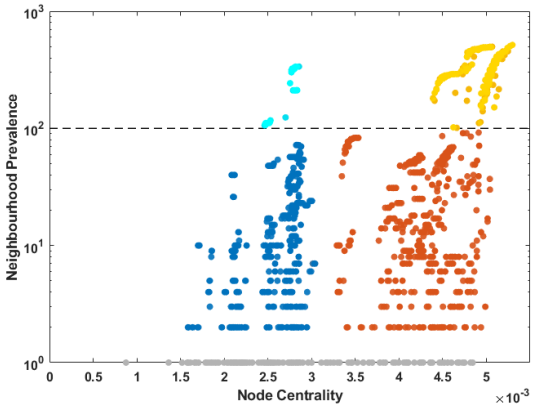

Figure 2. Centrality-prevalence space for the complete undirected SNP network. The neighbourhood prevalence (shown in log scale) of an isolate is measured by the number of neighbours within G_{max}=20. Blue/cyan and orange/yellow colours distinguish the groups of isolates with lower and higher centrality split at 3.1×10^{-3}. *The dashed line, set at 100 isolates, separates lower and higher neighbourhood prevalence (shown with lighter and darker colours).*

Using the WGS data, we derived undirected genotype networks in which nodes represent individual isolates and edges represent genetic proximity (Figure 1), defined by distances between the genomes quantified by differences between their single nucleotide polymorphisms (SNPs) [2]. We then built a 2-dimensional space in terms of two quantities measured for each node: the closeness centrality and the prevalence defined as the number of close neighbours within G_{max} = 20. This centrality-prevalence space relates a structural property of the pathogen population (network centrality) to its functional property (neighbourhood prevalence) [3,4], with the function interpreted as the average fitness of a pathogen. Figure 2 reveals a structure with at least 2 groups distinguished by centrality: isolates with low-to-mid centrality ($\leq 3.1 \times 10^{-3}$), and isolates with mid-to-high centrality ($> 3.1 \times 10^{-3}$). It is worth pointing out that Figure 1 shows a thresholded network including only the edges with SNP distance smaller than G_{max}=20, whereas Figure 2 is based on a complete network. Previous studies argued that a mid-to-high centrality region characterises higher-risk STM pathogens [3,4], however these findings have not been verified with WGS data.

In order to explain the separation between two groups in terms of evolutionary pathways leading to high-risk pathogens, we computed normalised entropy of the shell genome for each group. The pangenome comprises the entire set of genes from all isolates, and shell genes are identified as those that appear in 10% to 95% of the observed genomes. The entropy is computed using gene presence/absence data over a fixed number of time-ordered isolates (n = 10, note that collection date is not uniformly spaced) and normalised [5]. The isolates with low-to-mid centrality have a higher normalised shell genome entropy, which demonstrates a greater genetic diversity (Figure 3.a), while the normalised shell genome entropy

for isolates with mid-to-high centrality is smaller, showing a reduced genetic diversity (Figure 3.b). This difference highlights an exploration-exploitation distinction between the evolutionary pathways: the isolates with low-to-mid centrality tend to explore the adaptive landscape, while the isolates with mid-to-high centrality are more stable in their shell genome, thus exploiting their niches.

The exploration-exploitation distinction is also supported by analysis of a directed network constructed by giving each edge of the original genotype network a direction, determined by the collection date of the connected nodes (isolates). To trace genetic changes over time, we identified sufficiently long directed paths formed by connecting a sequence of directed edges ($m > 5$). For each path we compared changes of the neighbourhood prevalence between the start and end nodes, distinguishing positive ("successful") and negative ("unsuccessful") changes. Following [3], the identified paths were projected on the centrality-prevalence space of the original undirected network, and for each point (i.e., each node) the prevalence changes were averaged across all paths originating from this node, thus separating nodes tending to evolve to higher or lower prevalence. Figure 4 shows two regions: a dense "transition" region with a majority of nodes starting successful paths, and a smaller "bottleneck" region with nodes starting unsuccessful paths (in blue circle). Increasing prevalence is only observed in the group of isolates with mid-to-high centrality (Figure 4; shown in red).

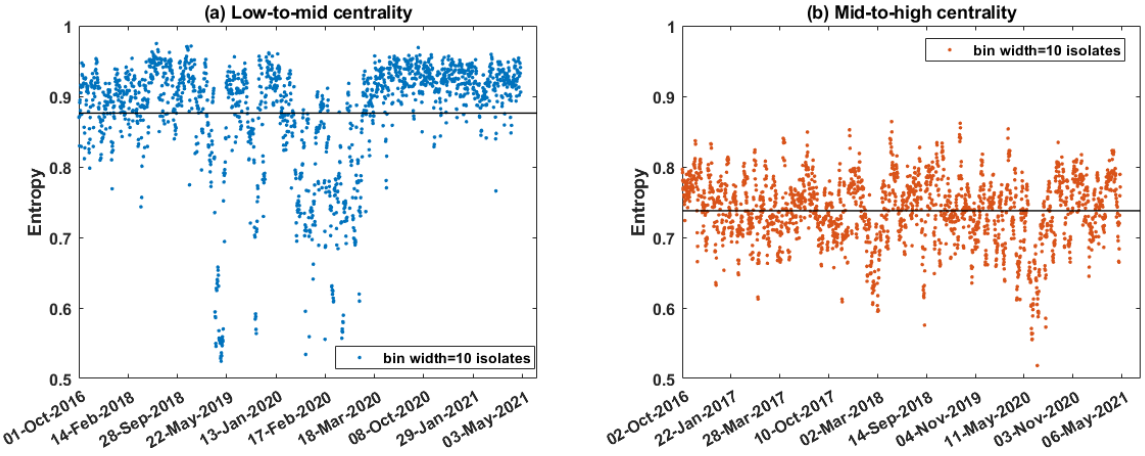

Figure 3. Normalised shell genome entropy (713 genes, in bits) over time for the two groups identified in Fig. 2. The normalised entropy is higher for (a) isolates with low-to-mid centrality (mean=0.876, SD=0.082) than for (b) isolates with mid-to-high centrality (mean=0.738, SD=0.048). Mean is marked by solid black line. Dates are indicative. Note that time intervals are not evenly distributed: each bin contains the same number of isolates (i.e., 10) but time intervals have different duration.

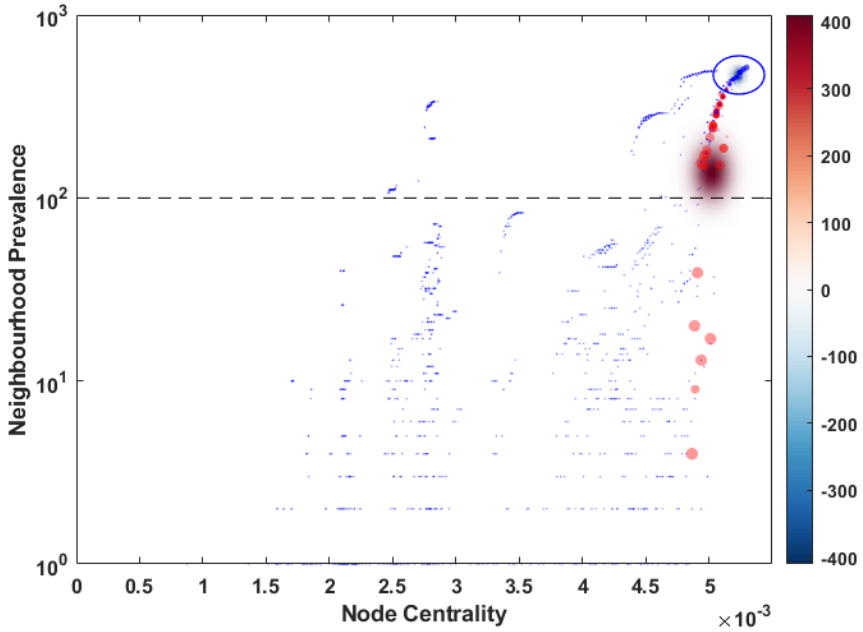

Figure 4. Estimated probability density of transition and bottleneck regions in the centrality-prevalence space, revealed by the expected values of both successful and unsuccessful paths. For every isolate, the point size is proportional to the average change in neighbourhood prevalence. Positive change in prevalence (i.e., increasing number of connections) is shown in red, and negative change (i.e., decreasing number of connections) is shown in blue. Only paths longer than 5 steps (m>5) are considered. The dashed line, set at 100 isolates, separates lower and higher neighbourhood prevalence.

These observations demonstrate, now at the WGS level, (i) a clear distinction between explorative and exploitative groups of pathogens, and (ii) an evolutionary pathway along which pathogens exploit the search-space by increasing their network centrality, before encountering a bottleneck. In this case, such a bottleneck can be attributed to the national Foodborne Illness Reduction Strategy [1] which successfully contained STM outbreaks in NSW in the last five years. These findings further suggest that population of pathogens is self-organising in the centrality-prevalence space, under both evolutionary and epidemic containment constraints.

References

[1] Davis, B.P.F, et al. Salmonellosis in Australia in 2020: possible impacts of COVID-19 related public health measures, *Communicable Diseases Intelligence* (2022) 46.

[2] Svahn, A.J., et al. Genome-wide networks reveal emergence of epidemic strains of Salmonella Enteritidis, *International Journal of Infectious Diseases* (2022) 117:65-73.

[3] Cliff O.M., et al. Inferring Evolutionary Pathways and Directed Genotype Networks of Foodborne Pathogens. *PLOS Computational Biology* (2020) 16:e1008401.

[4] Cliff O.M., et al. Network Properties of Salmonella Epidemics. *Scientific Reports* (2019) 9:6159.

[5] Wang, Y. and Shang, P. Analysis of Shannon-Fisher information plane in times series based on information entropy. *Chaos* (2018) 28:103107.