OutbreakFinder: a visualization tool for rapid detection of bacterial strain clusters based on optimized multidimensional scaling

Ming-Hsin Tsai[1,*], Yen-Yi Liu[2,*] and Chih-Chieh Chen[3,4]

[1] Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan
[2] Center for Research, Diagnostics and Vaccine Development, Centers for Disease Control, Ministry of Health and Welfare, Taichung, Taiwan
[3] Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung, Taiwan
[4] Rapid Screening Research Center for Toxicology and Biomedicine, National Sun Yat-sen University, Kaohsiung, Taiwan
* These authors contributed equally to this work.

ABSTRACT

With the evolution of next generation sequencing (NGS) technologies, whole-genome sequencing of bacterial isolates is increasingly employed to investigate epidemiology. Phylogenetic analysis is the common method for using NGS data, usually for comparing closeness between bacterial isolates to detect probable outbreaks. However, interpreting a phylogenetic tree is not easy without training in evolutionary biology. Therefore, developing an easy-to-use tool that can assist people who wish to use a phylogenetic tree to investigate epidemiological relatedness is crucial. In this paper, we present a tool called OutbreakFinder that can accept a distance matrix in csv format; alignment files from Lyve-SET, Parsnp, and ClustalOmega; and a tree file in Newick format as inputs to compute a cluster-labeled two-dimensional plot based on multidimensional-scaling dimension reduction coupled with affinity propagation clustering. OutbreakFinder can be downloaded for free at https://github.com/skypes/Newton-method-MDS.

Subjects Bioinformatics, Genomics, Microbiology
Keywords Multidimensional scaling, Microbiome clustering, Affinity propagation, Disease outbreak detection, MDS

Submitted 16 April 2019
Accepted 1 August 2019
Published 28 August 2019

Corresponding author
Chih-Chieh Chen,
chieh@imst.nsysu.edu.tw

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 7

DOI **10.7717/peerj.7600**

INTRODUCTION

Phylogenetic trees are widely used in public health to infer the molecular epidemiology of infectious diseases (*Hall & Barlow, 2006*; *Pybus, Fraser & Rambaut, 2013*). Many traditional methods such as the maximum likelihood, neighbor-joining, and Bayesian methods have been used to compute phylogenies for successful inference of epidemiological relationships (*Den Bakker et al., 2014*; *Grad et al., 2012*; *Leekitcharoenphon et al., 2016*; *Tettelin et al., 2014*). A phylogenetic-tree-like approach named "genetic relatedness tree," constructed using pulsed-field gel electrophoresis (PFGE) and multilocus sequence typing (MLST) data, is widely used in molecular

epidemiology (*Hunter et al., 2005*; *Maiden et al., 1998*). Deep learning has become more and more widely used to solve many of the most advanced issues in various fields. It can also be applied to bioinformatics to reduce the need for feature extraction and achieve high performance (*Le & Nguyen, 2019*). In epidemiological data analysis, the most crucial step in interpreting a phylogenetic tree is investigation of "clusters," which are usually considered probable outbreaks (*Ragonnet-Cronin et al., 2013*; *Vrbik et al., 2015*). Evolutionary biologists can easily interpret phylogenetic trees; however, epidemiologists find it difficult to use phylogenetic information to assist their research on outbreak detection. Therefore, we introduced a dimension reduction approach named multidimensional scaling (MDS) to transform a phylogenetic tree into a two-dimensional plot and subsequently apply affinity propagation (AP) (*Frey & Dueck, 2007*) for clustering. This approach can facilitate rapid identification of relationships among compared bacterial isolates for investigation of clusters. In this paper, we present a tool called OutbreakFinder that can assist in outbreak detection; this tool can help users to plot an MDS plot with a cluster labeled from a distance matrix or alignment files from Lyve-SET (*Katz et al., 2017*), Parsnp (*Treangen et al., 2014*), or ClustalOmega (*Sievers & Higgins, 2014*).

MATERIALS AND METHODS

Multidimensional scaling method

Multidimensional scaling is a widely used approach to projecting high-dimensional data onto a low-dimensional space, where the relationships among individual samples are preserved. This characteristic of MDS is often used to present high-dimensional data on a two-dimensional plane. A simple derivation of the MDS method is presented as follows:

Given n points in a p-dimensional space, the Euclidean distance between any two points, x_r and x_s, can be defined as $d_{rs}^2 = \sum_{i=1}^{p}(x_{ri} - x_{si})^2$. The MDS method aims to minimize the objective function as follows:

$$\text{tress} = \sqrt{\frac{\sum\left(d_{rs} - \hat{d}_{rs}\right)^2}{\sum {d_{rs}}^2}},$$

where $\hat{d}_{rs}$ is the predicted Euclidean distance. Let $B = XX^T$ be an $n \times n$ inner product matrix. In metric MDS, matrix B can be decomposed into QMQ^T, where Q is an eigenvector and M is a diagonal matrix of eigenvalues. The two forms of matrix B are combined as follows:

$$QMQ^T = Q\sqrt{M}\sqrt{M}Q^T = \left(\sqrt{M}Q^T\right)^T\left(\sqrt{M}Q^T\right) = X^TX$$

Then, we obtain $X = \sqrt{M}Q^T$. Therefore, the aforementioned derivation demonstrates that the traditional MDS method is designed to find an approximate solution rather than the optimal solution. For detailed derivation of the traditional MDS method, please refer to Supplemental Appendix A.

MDS by using Newton's method

The 100 points in Fig. S1 are generated by the function genSimulationData() as shown in Supplemental Appendix B, every 10 points is a group, 0–9, 10–19, 20–29…, and so on.

Figure S1A is an example of the distortion produced by the traditional MDS method, as shown in Supplemental Appendix C. Points 0–9 belong to the same group; however, points 9 and 3 evidently do not belong to the same group in traditional MDS (Fig. S1A). Therefore, we propose to perform MDS by using Newton's method (NMDS) to compensate for the deficiencies of the traditional MDS method. The proposed method has two parts; one is NMDS and the other is determination of the largest error edge to be amended. The detailed steps of the algorithm are as follows:

1. $\hat{d}_{jk} = \sqrt{\sum_{i=1}^{p} \left(x_{ji} - x_{ki}\right)^2}$
2. $r_{jk} = d_{jk}/\hat{d}_{jk}$, where d_{jk} is the observed distance
3. $\hat{x}_{ji} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ji} - x_{ki}\right)\left(r_{jk} - 1\right)$
4. Find the maximum error edge and force it to move to an ideal position (find the maximum ratio of $\hat{d}_{jk}/d_{jk}$ as the maximum error edge)

Here, $\hat{d}_{jk}$ represents the predicted distance between nodes j and k; r_{jk} represents the ratio of the observed and predicted distance between nodes j and k, namely $d_{jk}/\hat{d}_{jk}$; $\hat{x}_{ji}$ represents the new value of dimension i; and x_{ji} represents the current value of dimension i.

Steps 1–3 in NMDS minimize the objective function. Step 4 finds the maximum error edge and forces it to move to an ideal position. As shown in Fig. S1B, the proposed method can reduce grouping errors.

Multidimensional scaling by using Newton's method is implemented in OutbreakFinder, which provides parameters through which a user can change the distance scale between data points, such as using a logarithmic scale or power 2 scale for distances. Users can specify the color of each data point or use the AP algorithm to cluster and specify the color of each group. In addition, OutbreakFinder generates a text file of data point coordinates and color, as well as an MDS plot graph. Users can reproduce the MDS plot figure by using other tools. OutbreakFinder can directly read Lyve-SET, Parsnp, and multiple sequence alignment files and plot MDS graphs. If a user does not use the aforementioned tools, he or she can generate a distance matrix to plot an MDS graph.

Implementation

OutbreakFinder is written in Java and compiled into a standalone executable jar file that can be executed in Java Runtime Environment 1.8 or a later version. In addition, users can download the source code and compile it into a preferred version.

OutbreakFinder can parse the results of the Lyve-SET, Parsnp and Multi Alignment tools and generate MDS coordinates and graphs, as shown in Fig. 1. The user can define the color of each data point, and OutbreakFinder will export the corresponding image file according to the defined color map. Users can also use the AP module in OutbreakFinder for automatic clustering. The advantage of using an AP module for clustering is that there is no need to provide the number of groups, which makes cluster analysis easier.

Example analysis

To demonstrate how to use OutbreakFinder, we employed 59 *Salmonella* Heidelberg isolate genomes from three outbreaks with the same PFGE type (*Bekal et al., 2016*).

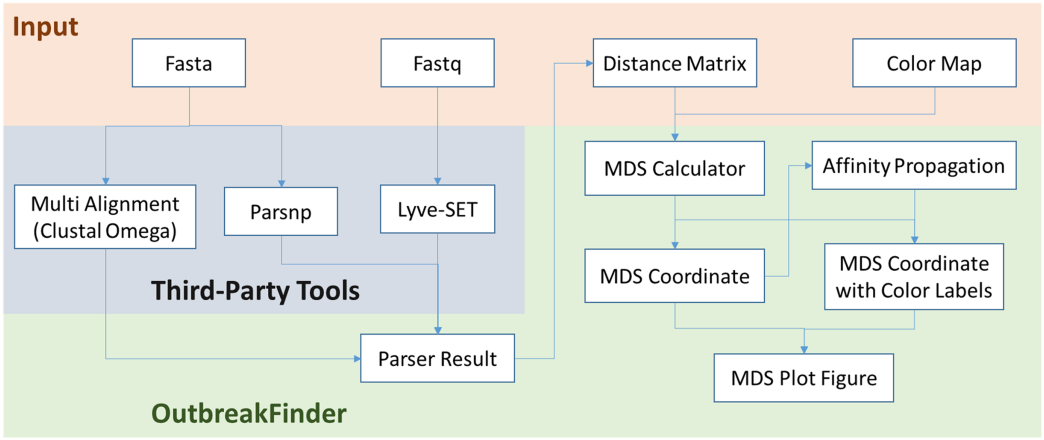

Figure 1 The schematic workflow of OutbreakFinder. Full-size DOI: 10.7717/peerj.7600/fig-1

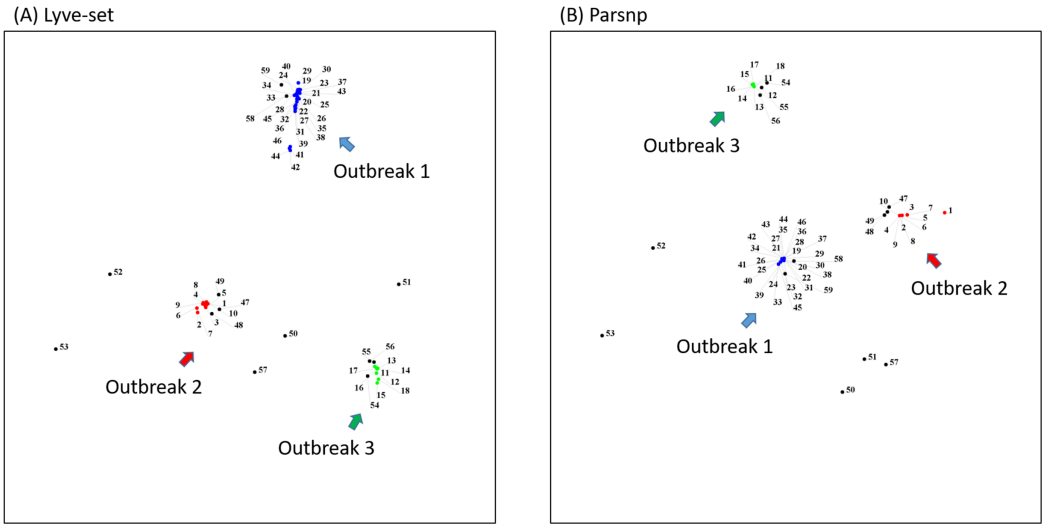

Figure 2 MDS plots of the 59 *Salmonella* Heidelberg isolates. The 59 *Salmonella* Heidelberg isolates from three outbreaks were identified as distinct clusters in the MDS plot by using Lyve-SET (A) and Parsnp (B). Outbreak 1, Outbreak 2, and Outbreak 3 are colored blue, green, and red, respectively.
Full-size DOI: 10.7717/peerj.7600/fig-2

Whole-genome sequencing raw reads (Table S1) were downloaded from the National Center for Biotechnology Information's Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/sra). The downloaded sra files were converted into fastq files; then, Lyve-SET was used to extract SNPs from these fastq files and produce a distance matrix. OutbreakFinder uses the distance matrix to generate an MDS plot, as shown in Fig. 2A. Subsequently, we used the same dataset to test Parsnp. OutbreakFinder can directly parse the results of Parsnp output and generate an MDS plot, as shown in Fig. 2B. We use published benchmark datasets (*Timme et al., 2017*) to verify the availability of OutbreakFinder. This dataset contains four major foodborne bacterial pathogens, such as *Listeria monocytogenes*, *Campylobacter jejuni*, *Escherichia coli*, and *Salmonella enterica*. These pathogens are listed in Tables S2–S5. The procedure is the same as described above,

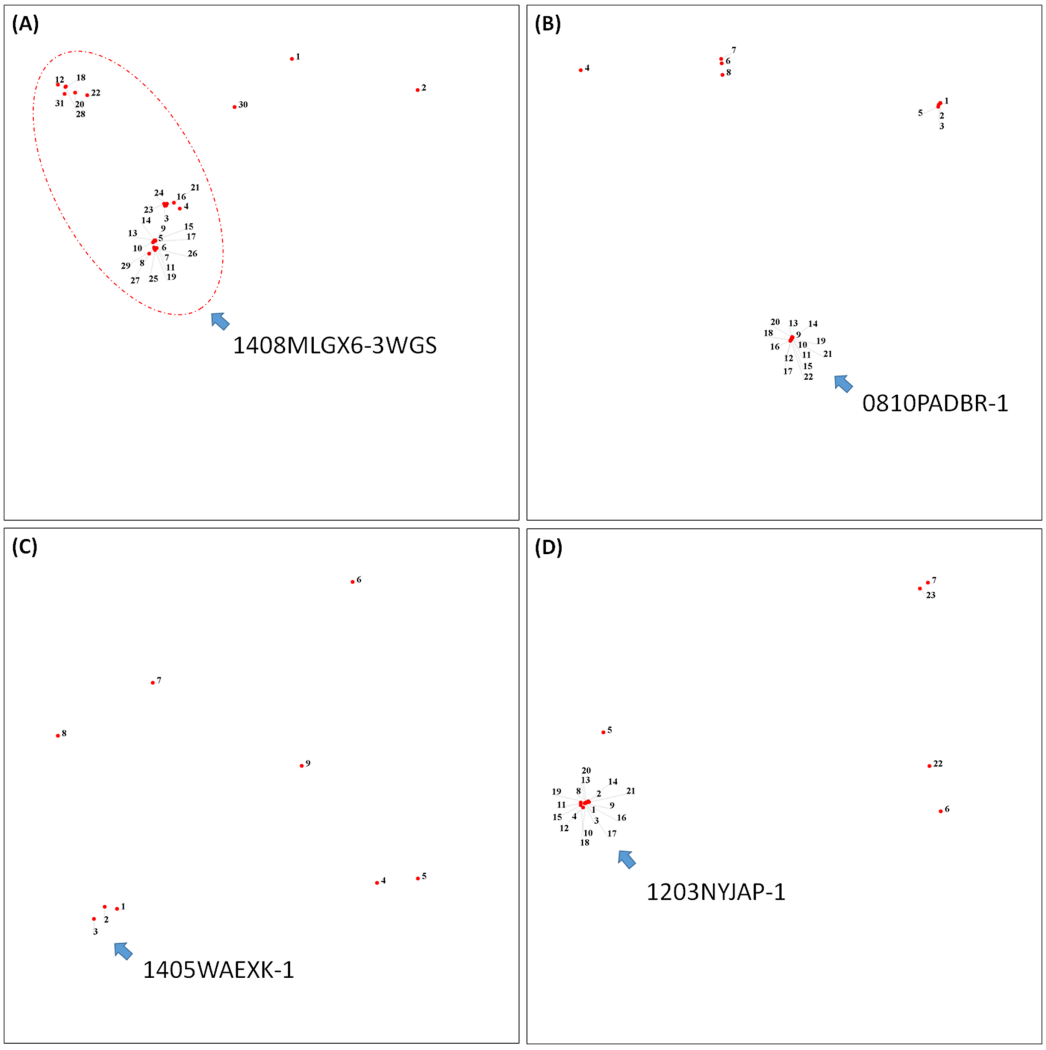

Figure 3 Results of the benchmark datasets. (A) Thirty-one *Listeria monocytogenes* isolates, (B) 22 *Campylobacter jejuni* isolates, (C) Nine *Escherichia coli* isolates, and (D) 23 *Salmonella enterica* isolates from outbreaks and outgroups were identified as distinct clusters in the MDS plot.
Full-size DOI: 10.7717/peerj.7600/fig-3

Lyve-SET was used to extract SNPs and produce the distance matrixes, and then OutbreakFinder uses the distance matrixes to export the MDS plot. As shown in Fig. 3, outbreaks and outgroups can be clearly distinguished from the MDS plot. Although there are two outbreak clusters in Fig. 3A, the outbreak and outgroup can still be clearly distinguished.

Performance

Finally, we used a large dataset to stress OutbreakFinder, which included 38,360 16S rRNA sequences downloaded from rrnDB (*Stoddard et al., 2015*). Although 38,360 sequences constitute a very large dataset, the results revealed that OutbreakFinder could still output results normally, as shown in Fig. 4. By contrast, if the traditional MDS method is used for dimension reduction analysis, more computational resources are required because the time

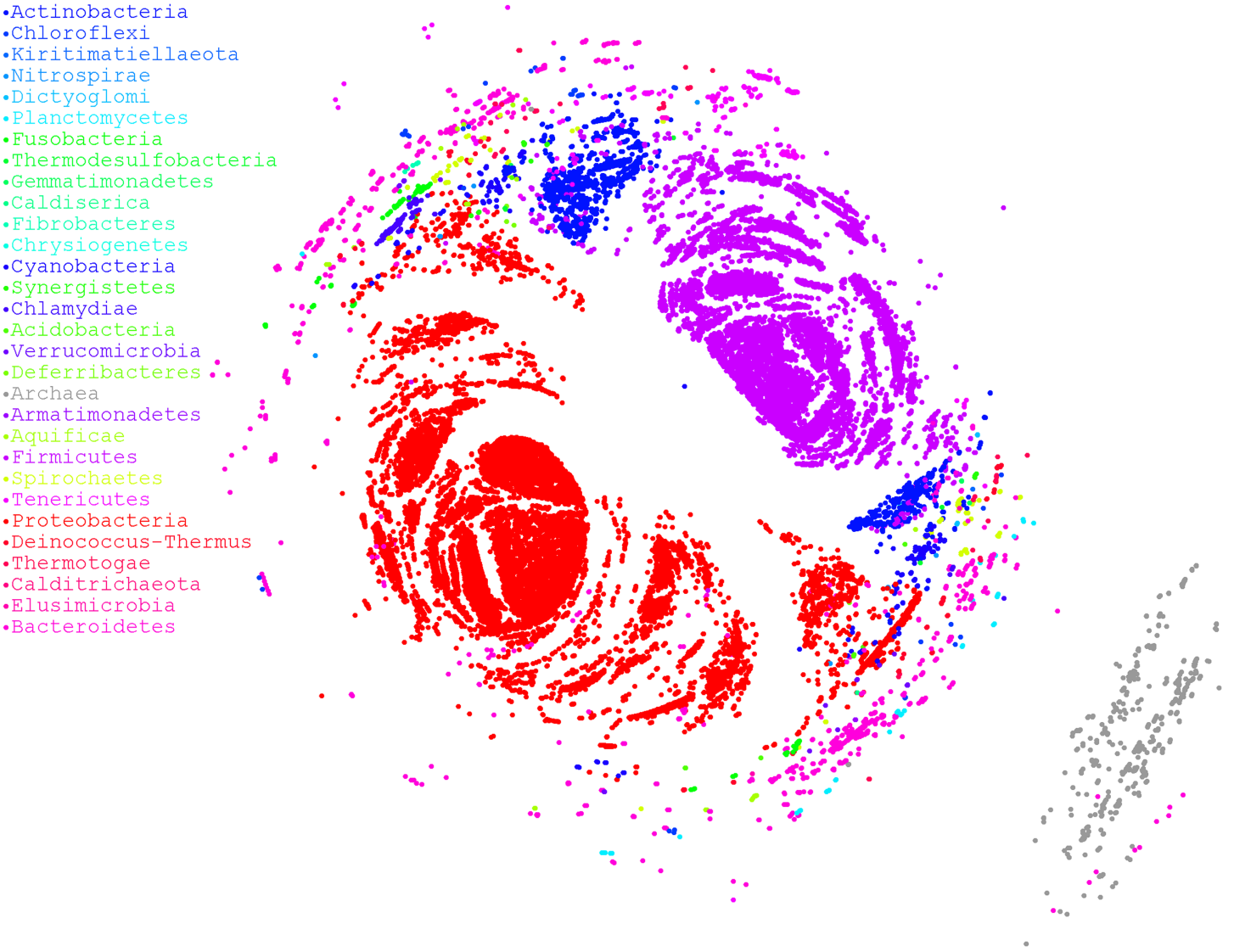

Figure 4 MDS ordination plot based on alignment distances among microorganisms. Different colors represent different phyla. A total of 38,360 sequences of 16S rRNA were downloaded from rrnDB, which contains 8,223 bacterial records representing 2,913 species and 262 archaeal records representing 201 species.
Full-size DOI: 10.7717/peerj.7600/fig-4

complexity of traditional MDS is $O(n^3)$. In the case of NMDS, the time complexity is only $O(cn^2)$, where c is the number of iterations and usually does not exceed 1,000. Therefore, OutbreakFinder can calculate the MDS of 38,360 16S rRNA in 1 day using a single core and 30 G memory computing resources. In contrast, the manifold.mds cannot handle the same dataset, even if the computing resources are increased to 20 core and 120 G memory, the work cannot be completed. Manifold.mds is a method in the scikit-learn suite in Python. To confirm that the proposed method is robust enough, we performed 1,000 simulations to compare the performance of NMDS and manifold.mds. As shown in Fig. S2, in most of the simulated fn is smaller than fc, fn, and fc are the values of the error function of NMDS and manifold.mds, respectively, which means that the performance of NMDS is better than that of manifest.mds.

DISCUSSION

Application of phylogenetic information for epidemiological inference is increasingly popular in the next generation sequencing era. Data from several approaches such as PFGE, single nucleotide polymorphism (SNP), and MLST can be used to construct a phylogenetic tree. However, interpreting a phylogenetic tree without training in evolutionary biology is seldom easy. An epidemiologist's primary goal is to determine whether bacterial isolates form a cluster that might indicate a probable outbreak. Therefore, development of a simple-to-use visualization tool that can help epidemiologists interpret relationships among compared isolates is crucial. In this paper, we present a visualization tool named OutbreakFinder that can show relationships among compared bacterial isolates on a two-dimensional MDS plot from distance matrix calculation or tree transformation (in Newick format). With the assistance of OutbreakFinder, epidemiologists can easily investigate probable outbreaks without experiencing the difficulty of "reading" phylogenetic trees.

CONCLUSIONS

The analysis of differences among biological individuals based on inferred phylogenetic trees is one of the main methods. It is also often used to analysis outbreaks in epidemiology. Relative to the phylogenetic tree, the graph may be more suitable for the presentation of outbreak analysis. We recommend using the MDS method to present the results of the analysis, because MDS is more suitable for use in clustering problems. We use empirical data to verify the performance of OutbreakFinder, and the results show that MDS plot can clearly distinguish outbreaks. In this study, we also improved on the weaknesses of existing MDS tools. On the one hand, we have improved the effect of clustering, while the other is to reduce the need for computing resources. In addition, we implemented AP in OutbreakFinder to facilitate cluster analysis. No need to specify the number of clusters is the advantage of using an AP. The source code for OutbreakFinder is available on GitHub.

ACKNOWLEDGEMENTS

We are grateful to NHRI, Taiwan for providing the hardware and software resources.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by grant from the Ministry of Science and Technology (MOST grant numbers 107-2311-B-110-001 and 108-2221-E-110-061-MY2), Taiwan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Ministry of Science and Technology: 107-2311-B-110-001 and 108-2221-E-110-061-MY2.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ming-Hsin Tsai conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper.
- Yen-Yi Liu conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper.
- Chih-Chieh Chen conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data is available at GitHub: https://github.com/skypes/Newton-method-MDS.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7600#supplemental-information.

REFERENCES

Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, Doualla-Bell F, Levac E, Gaulin C, Ramsay D, Huot C, Walker M, Sieffert C, Tremblay C. 2016. Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent *Salmonella enterica* serovar Heidelberg Clone in the context of outbreak investigations. *Journal of Clinical Microbiology* **54(2)**:289–295 DOI 10.1128/JCM.02200-15.

Den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z, Kinney A, Limberger R, Musser KA, Shudt M, Strain E, Wiedmann M, Wolfgang WJ. 2014. Rapid whole-genome sequencing for surveillance of *Salmonella enterica* serovar enteritidis. *Emerging Infectious Diseases* **20(8)**:1306–1314 DOI 10.3201/eid2008.131399.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* **315(5814)**:972–976 DOI 10.1126/science.1136800.

Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, DeSmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Moller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill F-X, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences of the United States of America* **109(8)**:3065–3070 DOI 10.1073/pnas.1121491109.

Hall BG, Barlow M. 2006. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. *Annals of Epidemiology* **16(3)**:157–169 DOI 10.1016/j.annepidem.2005.04.010.

Hunter SB, Vauterin P, Lambert-Fair MA, Van Duyne MS, Kubota K, Graves L, Wrigley D, Barrett T, Ribot E. 2005. Establishment of a universal size standard strain for use with the PulseNet standardized pulsed-field gel electrophoresis protocols: converting the national

databases to the new size standard. *Journal of Clinical Microbiology* **43(3)**:1045–1050 DOI 10.1128/JCM.43.3.1045-1050.2005.

Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Van Domselaar G, Deng X, Carleton HA. 2017. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Frontiers in Microbiology* **8**:375 DOI 10.3389/fmicb.2017.00375.

Leekitcharoenphon P, Hendriksen RS, Le Hello S, Weill FX, Baggesen DL, Jun SR, Ussery DW, Lund O, Crook DW, Wilson DJ, Aarestrup FM. 2016. Global genomic epidemiology of *Salmonella enterica* serovar typhimurium DT104. *Applied and Environmental Microbiology* **82(8)**:2516–2526 DOI 10.1128/AEM.03821-15.

Le NQK, Nguyen V-N. 2019. SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data. *PeerJ Computer Science* **5**:e177 DOI 10.7717/peerj-cs.177.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* **95(6)**:3140–3145 DOI 10.1073/pnas.95.6.3140.

Pybus OG, Fraser C, Rambaut A. 2013. Evolutionary epidemiology: preparing for an age of genomic plenty. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368(1614)**:20120193 DOI 10.1098/rstb.2012.0193.

Ragonnet-Cronin M, Hodcroft E, Hue S, Fearnhill E, Delpech V, Brown AJ, Lycett S, Database UHDR. 2013. Automated analysis of phylogenetic clusters. *BMC Bioinformatics* **14(1)**:317 DOI 10.1186/1471-2105-14-317.

Sievers F, Higgins DG. 2014. Clustal omega. *Current Protocols in Bioinformatics* **48(1)**:3.13.1–3.13.16 DOI 10.1002/0471250953.bi0313s48.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. 2015. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* **43(D1)**:D593–D598 DOI 10.1093/nar/gku1201.

Tettelin H, Davidson RM, Agrawal S, Aitken ML, Shallom S, Hasan NA, Strong M, De Moura VCN, De Groote MA, Duarte RS, Hine E, Parankush S, Su Q, Daugherty SC, Fraser CM, Brown-Elliott BA, Wallace RJ Jr, Holland SM, Sampaio EP, Olivier KN, Jackson M, Zelazny AM. 2014. High-level relatedness among *Mycobacterium abscessus* subsp. massiliense strains from widely separated outbreaks. *Emerging Infectious Diseases* **20(3)**:364–371 DOI 10.3201/eid2003.131106.

Timme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, Davis S, Tillman GE, Defibaugh-Chavez S, Carleton HA, Klimke WA, Katz LS. 2017. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **5**:e3893 DOI 10.7717/peerj.3893.

Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology* **15(11)**:524 DOI 10.1186/s13059-014-0524-x.

Vrbik I, Stephens DA, Roger M, Brenner BG. 2015. The Gap Procedure: for the identification of phylogenetic clusters in HIV-1 sequence data. *BMC Bioinformatics* **16**:355 DOI 10.1186/s12859-015-0791-x.