The Journal of Open Source Software

smot: a python package and CLI tool for contextual phylogenetic subsampling

Zebulun W. Arendsee [1], **Amy L. Vincent** [1], **and Tavis K. Anderson** [1¶]

1 Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA, USA ¶ Corresponding author

DOI: 10.21105/joss.04193

Software

- Review
- Repository
- Archive

Editor: Hugo Gruson

Reviewers:

- @Chjulian
- @marekborowiec

Submitted: 01 February 2022
Published: 20 December 2022

Summary

smot (Simple Manipulation Of Trees) is a command line tool and Python package with the pragmatic goal of distilling large-scale phylogenetic data to facilitate visualization without jeopardizing inference. This package offers subsampling algorithms that preserve reference taxa and tree topology, algorithms for classifying unlabeled tips given a subset of labeled reference tips, and functions for filtering phylogenetic trees. The smot tool has broad application in phylogenetic analysis and we demonstrate its utility using a genomic epidemiology study of influenza A virus in swine.

Statement of Need

Molecular phylogenetic analysis is initiated through the generation of a sequence dataset, followed by multiple sequence alignment, phylogenetic tree inference, and the identification of evolutionary relationships of interests (Baldauf, 2003). Given the rapid generation of large molecular sequence datasets, phylogenetic trees can become cumbersome and it may be necessary to subset data to address specific hypotheses or to facilitate visualization. For example, in a phylogenetic tree with thousands of taxa that are clustered into many monophyletic groups, the user may want to subsample the taxa while ensuring all groups and their evolutionary relationships are represented in the visualized tree. Alternatively, taxa on trees may be described and grouped by variables not defined by common ancestry such as geographical regions, phenotypes, sequence motifs, or host species and this information may be incomplete and require classification. In these cases, subsampling and classifying data on the phylogenetic tree can form the basis of hypotheses on how temporal, spatial, and other processes correlate with the evolutionary history of the studied population (Baum et al., 2005; Baum & Smith, 2013). Developing appropriate hypotheses can be facilitated through the subsampling, classifying, and filtering algorithms in smot.

The utility of smot is explained by the following three cases. For subsampling, smot may process a phylogenetic tree that has labeled clades: taxa from each clade may be subsampled while maintaining a set number of taxa from each clade and also retaining provided reference taxa. Alternatively, the number of taxa within a clade may be scaled as $n_{sampled} = n^{1/r}$, where n is the original number of taxa in the clade and r is a user provided scaling factor. For classification, smot can process a partially labeled tree by inferring missing labels and prepending them to the taxa names. For filtering, smot takes a tree with user-provided trait labels and then removes, subsamples or edits any monophyletic subtrees that meet specific criteria. The subsequent section will briefly introduce related tools currently applied in phylogenetic analyses and outline the role smot can play in a phylogeneticist's toolbox.

Related Work

The first purpose of `smot` is subsampling. Phylogenetic subsampling has been extensively researched (Mongiardino Koch, 2021). It may be used to remove sequences with quality problems or to reduce the dataset prior to computationally expensive operations. Programs such as Treemmer (Menardo et al., 2018), TreeTrimmer (Maruyama et al., 2013), and Treeshrink (Mai & Mirarab, 2018) approach the general problem of statistical subsampling while preserving specific diversity metrics of the original inferred phylogenetic tree. In contrast, `smot` is designed to subsample from groups within trees while preserving desired references.

The second purpose of `smot` is classification where unlabeled taxa in a tree are classified using a subset of labeled taxa. This is achieved by either patristic distance or monophyletic grouping. The method is based on submitted reference taxa rather than inferring clusters from the tree and then naming them. Inferring clusters de novo can be accomplished with tools such as phyCLIP (Han et al., 2019) or DYNAMITE (Magalis et al., 2021). Alternatively, taxa may be classified into clades using a fixed reference scaffold tree; this is done for influenza A virus in swine by `octoFLU` (Chang et al., 2019). `smot` depends on reference taxa, like `octoFLU`, but it extracts them from the taxa names or tables of attributes.

The third purpose of `smot` is filtering and coloring a tree based upon user queries. `smot` does not have more specialized phylogenetic and visualization utilities; these are be provided by phylommand (Ryberg, 2016) and ETE suite-toolkit (Huerta-Cepas et al., 2016) in Python and ggtree in R (Yu et al., 2017). However, `smot` may be easily integrated into analytical pipelines as a module and can be used to set leaf and branch coloring, which may then be visualized with a tree viewer such as FigTree.

`smot` is primarily designed to be used as a command line tool, but it can also be imported as a Python package. It complements the existing Python phylogenetics ecosystem through its subsampling, classifying, and filtering algorithms. Currently, Python packages for phylogenetics include the Phylo module of Biopython (Cock et al., 2009), the ETE-toolkit (Huerta-Cepas et al., 2016), DendroPy (Sukumaran & Holder, 2010), and TreeSwift, which offers a scalable foundation for building algorithms that work on large trees (Moshiri, 2020).

Core Algorithms

A common theme across `smot`'s algorithms is to group tips and then perform an action on each group. All grouping algorithms require labels on some or all of the tips. Labels may be assigned to tips using: entries provided in a table; input field index given a text separator in taxa names; or through application of regular expressions captures over taxa names. Given these initial labels, the tree can be grouped using a patristic, monophyletic, or paraphyletic algorithm. The patristic algorithm groups all tips together under the label of the nearest labeled tip by branch distance on the tree. The monophyletic algorithm descends from root to tip (trees are assumed to be rooted). When a subtree with one or more tips share a common label, and all other tips are unlabeled, the subtree is yielded as one monophyletic group. The paraphyletic algorithm also descends from root to tip, but rather than setting a monophyletic subtree to a group, it merges adjacent monophyletic groups with the same label down the tree. When a node is reached that is monophyletic for the two subtrees with different labels, each subtree is set as a group, ensuring that the branch nearest to a group border is sampled from. See **Figure 1** for an simple example of the difference between the monophyletic and paraphyletic grouping algorithms.

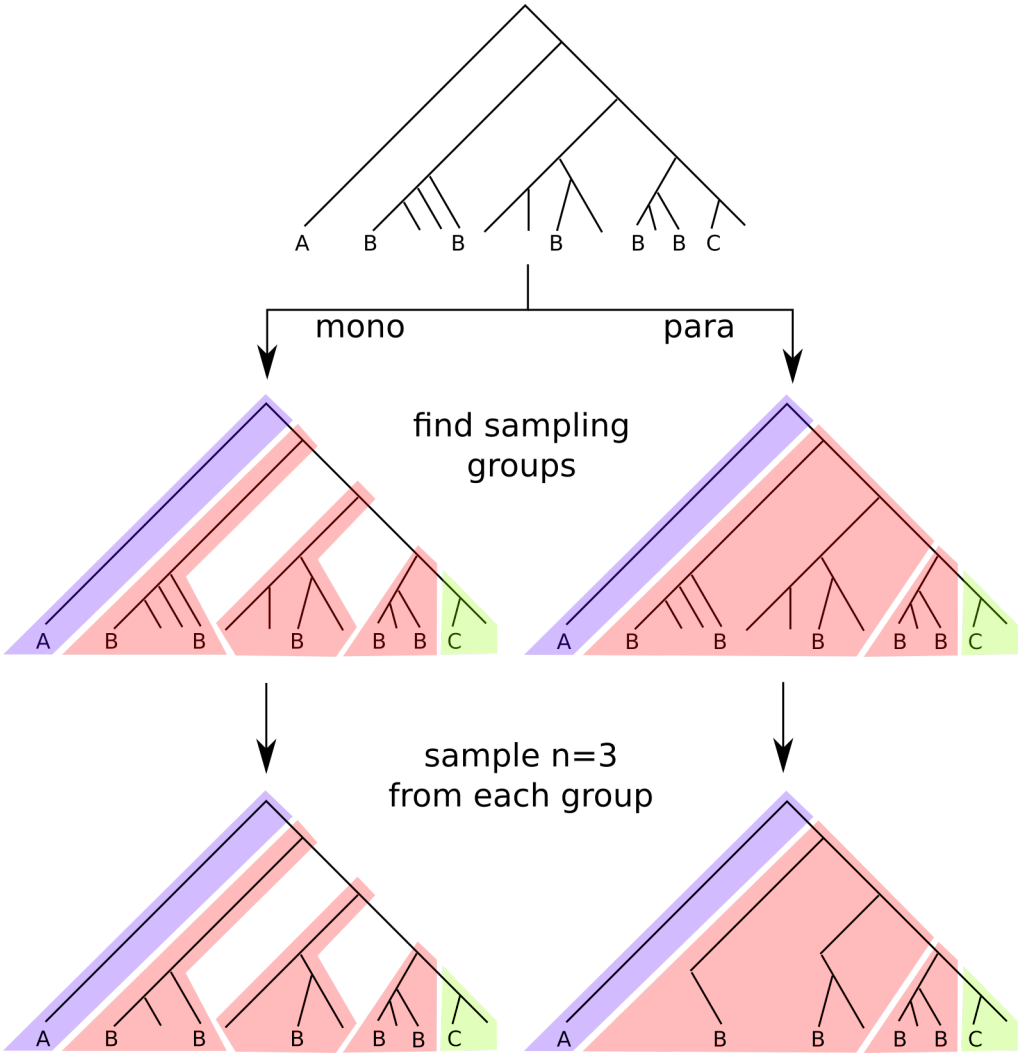


Figure 1: The monophyletic and paraphyletic algorithms differ in how they group the clades that will be downsampled (or otherwise acted upon). The paraphyletic algorithm groups adjacent monophyletic subtrees down the trunk but preserves the deepest monophyletic tree to guarantee that the nearest relative to a change in label is preserved.

Once a tree is partitioned into groups, it may be subsampled, classified, or filtered. Subsampling takes each partition and randomly selects either a set proportion of the tips (with an optional minimum tips) or a proportion that scales with group size. For scaled sampling, the number of sampled taxa equals $n^{1/r}$, where n is the number of tips in the group and r is the root (e.g., 2 for square root). Classifying either propagates the group label to all unlabeled members or assigns each unlabeled tip the label of the nearest labeled tip using a patristic classifier. Filtering performs an operation on each group under some condition, for example, it may delete all groups that have fewer than n members.

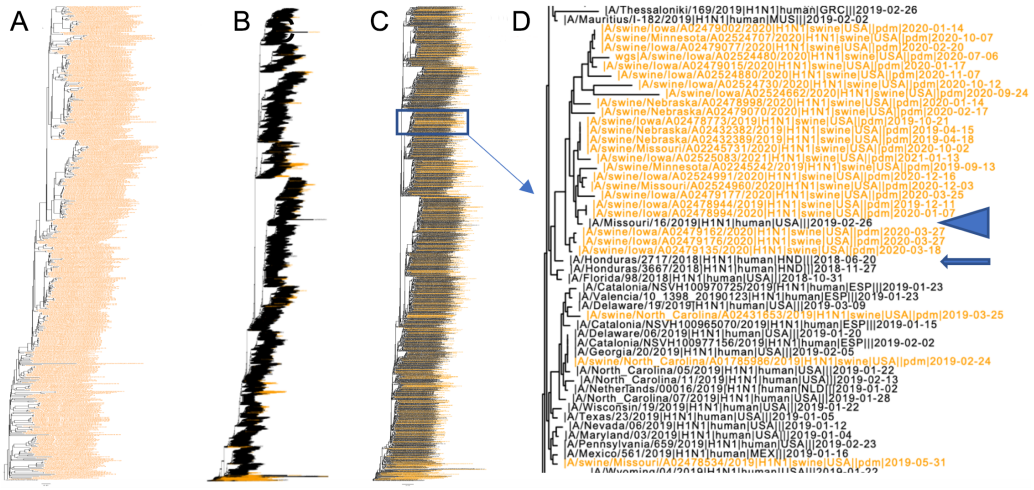


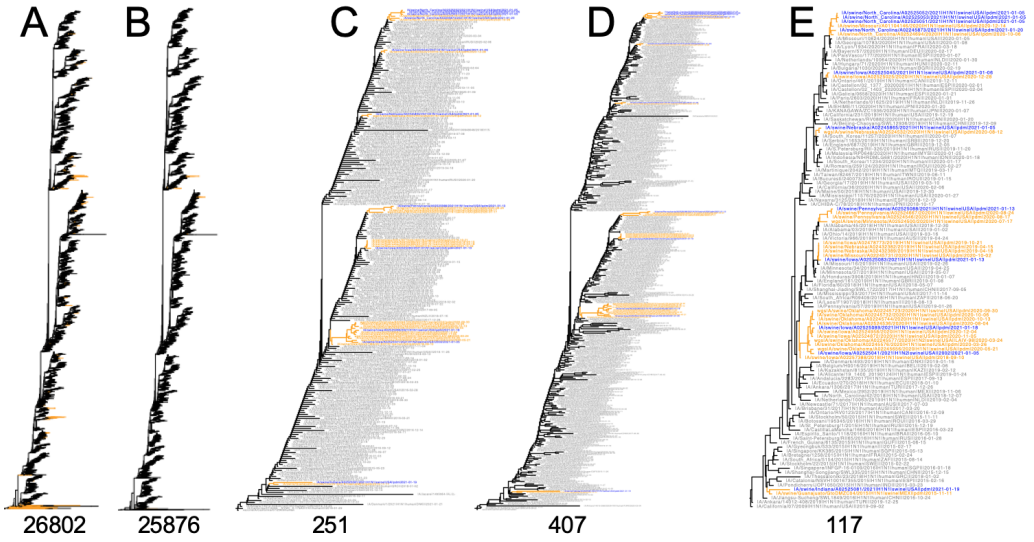
Figure 2: Interspecies transmission and evolution of the 2009 H1N1 influenza A virus pandemic (H1N1pdm09) lineage in swine and humans. (A) An inferred phylogenetic tree of influenza A virus (IAV) in swine hemagglutinin (HA) genes from the H1N1pdm09 lineage collected between 2015 and 2021. There are too many swine strains in the tree to read the labels even omitting the human influenza A virus H1N1 HA sequences necessary to capture the correct evolutionary context of the lineage. (B) An inferred phylogenetic tree of H1N1pdm09 lineage HA genes from humans and swine (26,802 genes, human in black, swine in orange). The tree is too large to see individual labels, and critical human-to-swine evolutionary linkage is obscured. To identify the evolutionary history of this IAV lineage, we include all swine HA genes to demonstrate onward transmission of the virus, and human HA genes to detect directionality of interspecies transmission. (C) An application of `smot`: human HA genes were down-sampled while keeping all swine genes. This ensured the context of human HA genes, allowing identification of human-to-swine spillovers and visualization of swine-to-swine transmission of the H1N1pdm09 lineage. All swine clades present in (B) are present in (C). (D) Using this approach, we identified a human-to-swine event (arrow) that seeded onward transmission in swine, followed by a single human HA gene nested within a monophyletic swine group (triangle) (blue rectangle in (C) and enlarged in the inset (D)). The human HA gene demonstrates a zoonotic (swine-to-human) transmission event. Subsampling human HA genes before building the tree or without considering context would likely obscure these two-way interspecies transmission events.

Case Study: Inferring human-to-swine influenza A virus transmission events

In 2009, an influenza A virus emerged in swine and was transmitted to humans causing the first pandemic of the 21st century (Smith et al., 2009). This H1N1 lineage (H1N1pdm09) became endemic in humans and is regularly reintroduced to swine populations globally (Nelson et al., 2015; Vijaykrishna et al., 2010). Phylogenetically, human-to-swine introductions can be detected based upon tree topology: an isolated swine-derived hemagglutinin (HA) gene nested within a monophyletic group of human genes indicates interspecies transmission (Volz et al., 2013). A similar tree structure can be used to infer zoonotic transmission from swine to humans (Nelson et al., 2015). To illustrate the shared evolutionary history of the H1N1pdm09 lineage, we inferred phylogenetic trees based on the HA genes collected from only swine (**Figure 2A**) and from swine and human sequences together (**Figure 2B**). In both cases, the size of the tree required to infer host origin and interspecies transmission events obscured visualization and the ability to infer the directionality of the transmission events.

The goal of this case study was to identify subsequent swine-to-swine transmission of H1N1pdm09 that descended from unique human-to-swine spillovers. To achieve this, we downloaded all swine and human H1N1pdm09 hemagglutinin (HA) genes sequences from the Influenza Research Database (Zhang et al., 2017). Each gene sequence was labeled

through the database search interface by virus strain name, host (human or swine), collection location, and date of collection. The HA genes were aligned with MAFFT (Katoh & Standley, 2013), and a maximum likelihood tree was inferred using a general time reversible model of molecular evolution with gamma distributed rate variation in FastTree (Price et al., 2010). To phylogenetically identify contemporary and sustained transmission of H1N1pdm09 in swine, we: subsampled large human clades; removed swine taxa with no recently observed representatives; and removed swine clades that had no evidence of swine-to-swine transmission (i.e., clades with a single representative).




```
smot grep -v "(swine|human)" pdm.tre |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --none-match="2021-" --remove |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --smaller-than 2 --remove |
    smot color rm |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --color="#FFA000" |
    smot color leaf -P -p "." "#909090" -p "swine" "#FFA000" -p "swine.*2021-" "#0000FF" > select-swine.tre

smot sample equal select-swine.tre --factor-by-capture="(swine|human)" \
  --max-tips=1 --keep="swine" > select-swine-equal-sample.tre

smot sample mono select-swine.tre --scale=4 --factor-by-capture="(swine|human)" \
  --min-tips=1 --keep="swine" --seed=42 > select-swine-mono-sample.tre

smot sample para select-swine.tre --scale=4 --factor-by-capture="(swine|human)" \
  --min-tips=3 --keep="swine" --seed=42 > select-swine-para-sample.tre
```

Figure 3: Cleaning and subsampling to extract the . The `smot`-processed phylogenetic tree can be used to identify human-to-swine spillover and sustained transmission of the 2009 H1N1 influenza A virus pandemic (H1N1pdm09) lineage in swine. (A) The original phylogenetic tree with human and swine H1N1pdm09 HA genes (n=26802) collected between 2009 and 2021. (B) The tree after filtering to keep only the swine clades that had more than one member and at least one 2021 representative. (C-E) The trees after subsampling with the (C) equal, (D) mono, and (E) para algorithms, respectively. Tip labels colored in orange represent swine hosts and orange branch coloring represents clades where all hosts are swine; blue tip labels are swine HA genes collected in 2021; each swine subtree represents an independent H1N1pdm09 clade circulating in US swine derived from a unique human-to-swine spillover event. The `smot` pipeline that produced the trees (C-E) was written in Bash and documentation and explanation of the code is provided in the GitHub README (https://github.com/flu-crew/smot) or the Flu Crew documentation page (https://flu-crew.github.io/docs/).

This process was achieved with a series of `smot` commands (**Figure 3**). First, `smot` extracted clades where all taxa labels where annotated with either the term "human" or "swine". We then removed all monophyletic swine clades without a detection in 2021 and all swine clades with a single member (i.e., isolated spillovers without evidence of sustained transmission). The resultant tree contained all human HA genes and swine HA genes for strains with evidence of contemporary circulation (**Figure 3B**): this tree was then subsampled with the three `smot`

algorithms (**Figure 3C-E**). In **Figure 3C**, we sampled 1 tip from each monophyletic human clade and generated a tree that demonstrated unique human and swine monophyletic clades. A similar presentation was generated in **Figure 3D** where the algorithm randomly selected $n^{1/4}$ tips from each monophyletic human clade, where n is the number of tips in the original clade, keeping a minimum of 1 tip. The third algorithm (**Figure 3E**) sampled paraphyletically, allowing human branches across the backbone to be jointly subsampled, allowing greater compression of the tree and more tractable visualization. The final tree (**Figure 3E**) demonstrated seven independent H1N1pdm09 human-to-swine spillover events with evidence of persistent swine-to-swine transmission Importantly, the tree was sufficiently compressed for labels to be readable on a single page while still providing the human context HA genes needed to resolve the seven unique human-to-swine spillover events.

```
smot grep -v "(swine|human)" pdm.tre |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --none-match="2021-" --remove |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --smaller-than 2 --remove |
    smot color rm |
    smot filter --factor-by-capture="(swine|human)" --all-match="swine" --color="#FFA000" |
    smot color leaf -P -p "." "#909090" -p "swine" "#FFA000" -p "swine.*2021-" "#0000FF" > select-swine.tre

smot sample equal select-swine.tre --factor-by-capture="(swine|human)" \
  --max-tips=1 --keep="swine" > select-swine-equal-sample.tre

smot sample mono select-swine.tre --scale=4 --factor-by-capture="(swine|human)" \
  --min-tips=1 --keep="swine" --seed=42 > select-swine-mono-sample.tre

smot sample para select-swine.tre --scale=4 --factor-by-capture="(swine|human)" \
  --min-tips=3 --keep="swine" --seed=42 > select-swine-para-sample.tre
```

Availability

`smot` is available on PyPi and the source is hosted on GitHub at https://github.com/flu-crew/smot. Additional documentation is available in the Flu Crew documentation page (https://flu-crew.github.io/docs/).

Acknowledgements

We gratefully acknowledge pork producers, swine veterinarians, and laboratories for participating in the USDA Influenza A Virus in Swine Surveillance System and publicly sharing sequences. This work was supported in part by: the U.S. Department of Agriculture (USDA) Agricultural Research Service [ARS project number 5030-32000-231-000-D]; the U.S. Department of Agriculture (USDA) Animal and Plant Health Inspection Service [ARS project number 5030-32000-231-080-I]; the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services [contract number 75N93021C00015]; the USDA Agricultural Research Service Research Participation Program of the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the USDA Agricultural Research Service [contract number DE-AC05- 06OR23100]; and the SCINet project of the USDA Agricultural Research Service [ARS project number 0500-00093-001-00-D]. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA, DOE, or ORISE. USDA is an equal opportunity provider and employer.

References

Baldauf, S. L. (2003). Phylogeny for the faint of heart: A tutorial. *TRENDS in Genetics*, *19*(6), 345–351. https://doi.org/10.1016/s0168-9525(03)00112-4

Baum, D. A., & Smith, S. D. (2013). Tree thinking. *An Introduction to Phylogenetic Biology. Roberts and Company Publishers*. https://doi.org/10.22269/210921

Baum, D. A., Smith, S. D., & Donovan, S. S. (2005). The tree-thinking challenge. *Science*, *310*(5750), 979–980. https://doi.org/10.1126/science.1117727

Chang, J., Anderson, T. K., Zeller, M. A., Gauger, P. C., & Vincent, A. L. (2019). octoFLU: Automated classification for the evolutionary origin of influenza a virus gene sequences detected in US swine. *Microbiology Resource Announcements*, *8*(32), e00673–19. https://doi.org/10.1128/mra.00673-19

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & others. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Han, A. X., Parker, E., Scholer, F., Maurer-Stroh, S., & Russell, C. A. (2019). Phylogenetic Clustering by Linear Integer Programming (PhyCLIP). *Molecular Biology and Evolution*, *36*(7), 1580–1595. https://doi.org/10.1093/molbev/msz053

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Magalis, B. R., Marini, S., Salemi, M., & Prosperi, M. (2021). DYNAMITE: A phylogenetic tool for identification of dynamic transmission epicenters. *bioRxiv*. https://doi.org/10.1101/2021.01.21.427647

Mai, U., & Mirarab, S. (2018). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*(5), 23–40. https://doi.org/10.1186/s12864-018-4620-2

Maruyama, S., Eveleigh, R. J., & Archibald, J. M. (2013). Treetrimmer: A method for phylogenetic dataset size reduction. *BMC Research Notes*, *6*(1), 1–6. https://doi.org/10.1186/1756-0500-6-145

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., Trauner, A., Beisel, C., Borrell, S., & Gagneux, S. (2018). Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, *19*(1), 1–8. https://doi.org/10.1186/s12859-018-2164-8

Mongiardino Koch, N. (2021). Phylogenomic subsampling and the search for phylogenetically reliable loci. *Molecular Biology and Evolution*, *38*(9), 4025–4038. https://doi.org/10.1101/2021.02.13.431075

Moshiri, N. (2020). TreeSwift: A massively scalable python tree package. *SoftwareX*, *11*, 100436. https://doi.org/10.1016/j.softx.2020.100436

Nelson, M. I., Stratton, J., Killian, M. L., Janas-Martindale, A., & Vincent, A. L. (2015). Continual reintroduction of human pandemic H1N1 influenza a viruses into swine in the united states, 2009 to 2014. *Journal of Virology*, *89*(12), 6218–6226. https://doi.org/10.1128/JVI.00459-15

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Ryberg, M. (2016). Phylommand - a command line software package for phylogenetics. *F1000Research*, *5*, 2903. https://doi.org/10.12688/f1000research.10446.1

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., & others. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature*, *459*(7250), 1122–1125. https://doi.org/10.1038/nature08182

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Vijaykrishna, D., Poon, L., Zhu, H., Ma, S., Li, O., Cheung, C., Smith, G., Peiris, J., & Guan, Y. (2010). Reassortment of pandemic H1N1/2009 influenza a virus in swine. *Science*, *328*(5985), 1529–1529. https://doi.org/10.1126/science.1189132

Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral phylodynamics. *PLoS Computational Biology*, *9*(3), e1002947. https://doi.org/10.1371/journal.pcbi.1002947

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., & others. (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, *45*(D1), D466–D474. https://doi.org/10.1093/nar/gkw857