

1 smot: a python package and CLI tool for contextual 2 phylogenetic subsampling

3 **Zebulun W. Arendsee¹, Amy L. Vincent¹, and Tavis K. Anderson^{1¶}**

4 **1** Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA, USA **true**

5 ¶ Corresponding author

DOI: [10.21105/joss.04193](https://doi.org/10.21105/joss.04193)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: ↗

Reviewers:

- [@chjulian](#)
- [@marekborowiec](#)

Submitted: 01 February 2022

Published: 30 November -001

License

Authors of papers retain
copyright and release the work
under a Creative Commons
Attribution 4.0 International
License ([CC BY 4.0](#)).

6 Summary

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

14 Statement of Need

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

30 The utility of smot is explained by the following three cases. For subsampling, smot may
31 process a phylogenetic tree that has labeled clades: taxa from each clade may be subsampled
32 while maintaining a set number of taxa from each clade and also retaining provided reference
33 strains. Alternatively, the number of strains within a clade may be scaled by sampling the r^{th}
34 root. For classification, smot can process a partially labeled tree by inferring missing labels and
35 prepending them to the taxa names. For filtering, smot may ingest a tree with user-provided
36 trait labels and then remove any monophyletic group that has no recent member or has too
37 few members. The subsequent section will briefly introduce related tools currently applied in
38 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 taxa in the tree while preserving objective metadata and annotations that are independent of tree topology.

The second purpose of smot is classification and annotation of unlabeled taxa and clades when provided representative strain classifications. Classification is achieved by either patristic distance or monophyletic purity, and 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), and for influenza A virus in swine by octoFLU (Chang et al., 2019). smot applies a hybrid approach where, rather than using a fixed reference scaffold tree, it extracts labels from the input tree based on annotated reference sequences and then infers the missing labels.

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 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 on taxa names to extract labels of interest. 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 cumulative 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.

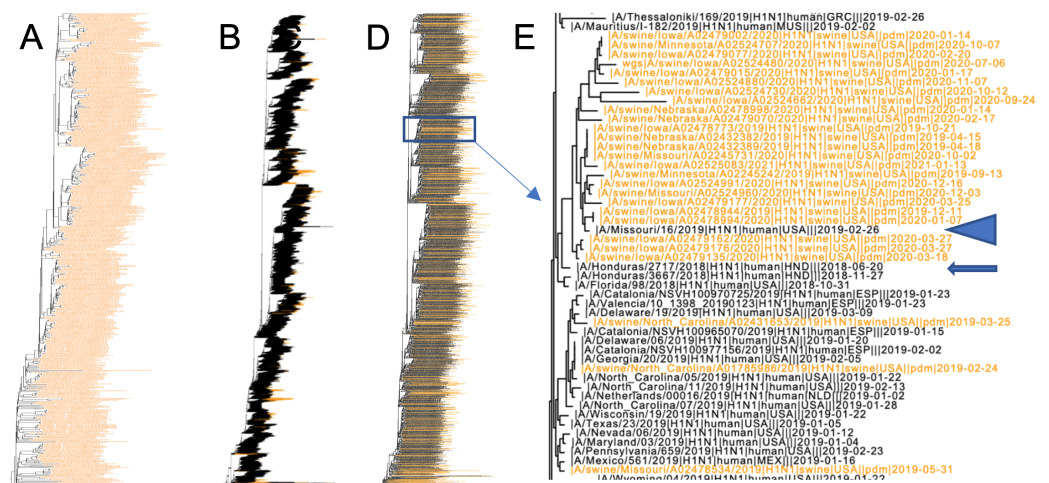


Figure 1: Figure 1. Interspecies transmission and evolution of the 2009 H1N1 influenza A virus pandemic (H1N1pdm09) lineage in swine and humans. (A) An inferred phylogenetic tree of influenza A virus (IAV) in swine hemagglutinin (HA) genes from the H1N1pdm09 lineage collected between 2015 and 2021. There 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.

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

91 Case Study: Inferring human-to-swine influenza A virus 92 transmission events

93 In 2009, an influenza A virus emerged in swine, was transmitted to humans and subsequently
94 caused the first pandemic of the 21st century (Smith et al., 2009). This H1N1 lineage
95 (H1N1pdm09) became endemic in humans and is regularly reintroduced to swine populations
96 globally (Nelson et al., 2015; Vijaykrishna et al., 2010). Phylogenetically, human-to-swine
97 introductions can be detected based upon tree topology: an isolated swine-derived HA gene
98 nested within a monophyletic group of human genes indicates interspecies transmission (Volz
99 et al., 2013). A similar tree structure can be used to infer zoonotic transmission from swine to

100 humans (Nelson et al., 2015). To illustrate the shared evolutionary history of the H1N1pdm09
101 lineage, we inferred phylogenetic trees based on the hemagglutinin (HA) gene collected from
102 only swine (**Figure 1A**) and from swine and human sequences together (**Figure 1B**). In both
103 cases, the scope of data required to infer host origin and interspecies transmission events
104 obscured visualization and the ability to infer the directionality of the transmission events.

105 The goal of this case study was to identify subsequent swine-to-swine transmission of
106 H1N1pdm09 that descended from unique human-to-swine spillovers. To achieve this, we
107 downloaded all swine and human H1N1pdm09 hemagglutinin (HA) genes sequences from
108 the Influenza Research Database (Zhang et al., 2017). Each gene sequence was labeled
109 through the database search interface by virus strain name, host (human or swine), collection
110 location, and date of collection. The HA genes were aligned with MAFFT (Katoh & Standley,
111 2013), and a maximum likelihood tree was inferred using a general time reversible model of
112 molecular evolution with gamma distributed rate variation in FastTree (Price et al., 2010). To
113 phylogenetically identify contemporary and sustained transmission of H1N1pdm09 in swine, we:
114 subsampled large human clades; removed swine taxa with no recently observed representatives;
115 and removed swine clades that had no evidence of swine-to-swine transmission (i.e., clades
116 with a single representative).

DRAFT

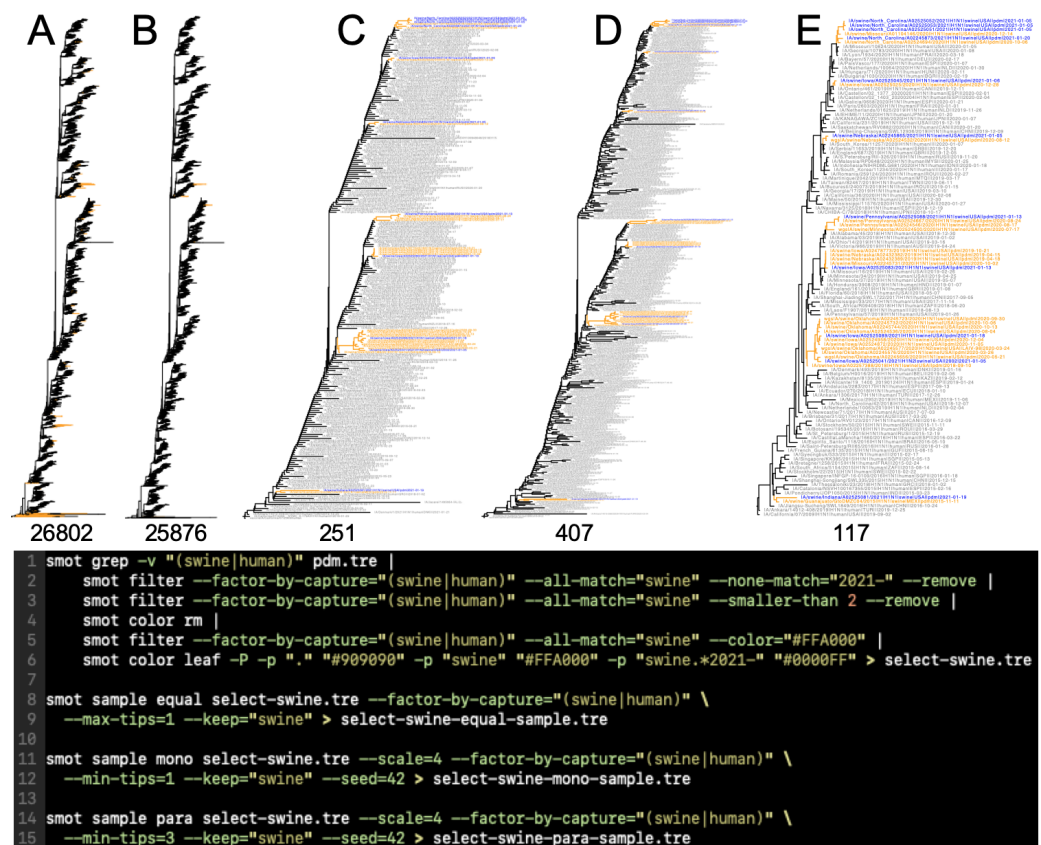


Figure 2: Figure 2. The smot tool and commands that demonstrate how a phylogenetic tree may be filtered and subsampled. The smot-processed phylogenetic tree can be used to identify human-to-swine spillover and sustained transmission of the 2009 H1N1 influenza A virus pandemic (H1N1pdm09) lineage in swine. (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; the clades annotated 1 through 4 in (E) represent four independent H1N1pdm09 clades circulating in US swine derived from unique human-to-swine spillover events. 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 2**). First, smot extracted clades where all taxa labels were 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 2B**); this tree was then subsampled with the three smot algorithms (**Figure 2C-E**). In **Figure 2C**, 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 2D** where the algorithm randomly selected $n_s = n^{1/4}$ tips from each monophyletic human clade, keeping a minimum of 1 tip. The third algorithm (**Figure 2E**) 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 2E**) 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.

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](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.

- 175 <https://doi.org/10.1093/molbev/msw046>
- 176 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version
177 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4),
178 772–780. <https://doi.org/10.1093/molbev/mst010>
- 179 Magalis, B. R., Marini, S., Salemi, M., & Prosperi, M. (2021). DYNAMITE: A phylogenetic
180 tool for identification of dynamic transmission epicenters. *bioRxiv*. <https://doi.org/10.1101/2021.01.21.427647>
181
- 182 Mai, U., & Mirarab, S. (2018). TreeShrink: Fast and accurate detection of outlier long
183 branches in collections of phylogenetic trees. *BMC Genomics*, 19(5), 23–40. <https://doi.org/10.1186/s12864-018-4620-2>
184
- 185 Maruyama, S., Eveleigh, R. J., & Archibald, J. M. (2013). Treetrimmer: A method for
186 phylogenetic dataset size reduction. *BMC Research Notes*, 6(1), 1–6. <https://doi.org/10.1186/1756-0500-6-145>
187
- 188 Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihua, L. K., Trauner,
189 A., Beisel, C., Borrell, S., & Gagneux, S. (2018). Treemmer: A tool to reduce large
190 phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, 19(1), 1–8.
191 <https://doi.org/10.1186/s12859-018-2164-8>
- 192 Mongiardino Koch, N. (2021). Phylogenomic subsampling and the search for phylogenetically
193 reliable loci. *Molecular Biology and Evolution*, 38(9), 4025–4038. <https://doi.org/10.1101/2021.02.13.431075>
194
- 195 Moshiri, N. (2020). TreeSwift: A massively scalable python tree package. *SoftwareX*, 11,
196 100436. <https://doi.org/10.1016/j.softx.2020.100436>
- 197 Nelson, M. I., Stratton, J., Killian, M. L., Janas-Martindale, A., & Vincent, A. L. (2015).
198 Continual reintroduction of human pandemic H1N1 influenza A viruses into swine in the
199 united states, 2009 to 2014. *Journal of Virology*, 89(12), 6218–6226. <https://doi.org/10.1128/JVI.00459-15>
200
- 201 Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—approximately maximum-
202 likelihood trees for large alignments. *PloS One*, 5(3), e9490. <https://doi.org/10.1371/journal.pone.0009490>
203
- 204 Ryberg, M. (2016). Phylommand - a command line software package for phylogenetics.
205 *F1000Research*, 5, 2903. <https://doi.org/10.12688/f1000research.10446.1>
- 206 Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K.,
207 Cheung, C. L., Raghwani, J., Bhatt, S., others. (2009). Origins and evolutionary genomics
208 of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459(7250), 1122–1125.
209 <https://doi.org/10.1038/nature08182>
- 210 Sukumaran, J., & Holder, M. T. (2010). DendroPy: A python library for phylogenetic
211 computing. *Bioinformatics*, 26(12), 1569–1571. <https://doi.org/10.1093/bioinformatics/btq228>
212
- 213 Vijaykrishna, D., Poon, L., Zhu, H., Ma, S., Li, O., Cheung, C., Smith, G., Peiris, J., & Guan,
214 Y. (2010). Reassortment of pandemic H1N1/2009 influenza A virus in swine. *Science*,
215 328(5985), 1529–1529. <https://doi.org/10.1126/science.1189132>
- 216 Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral phylodynamics. *PLoS Computational
217 Biology*, 9(3), e1002947. <https://doi.org/10.1371/journal.pcbi.1002947>
- 218 Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An R package
219 for visualization and annotation of phylogenetic trees with their covariates and other
220 associated data. *Methods in Ecology and Evolution*, 8(1), 28–36. <https://doi.org/10.1111/2041-210X.12628>
221

- 222 Zhang, Y., Aevertmann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S.,
223 Kumar, S., Larsen, C. N., Lee, A. J.others. (2017). Influenza research database: An
224 integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*,
225 45(D1), D466–D474. <https://doi.org/10.1093/nar/gkw857>

DRAFT