ISDS 2019 Conference Abstracts

ZooPhy: A bioinformatics pipeline for virus phylogeography and surveillance

Matthew Scotch, Arjun Magge, Matteo Vaiente

Arizona State University, Tempe, Arizona, United States

Objective

We will describe the ZooPhy system for virus phylogeography and public health surveillance [1]. ZooPhy is designed for public health personnel that do not have expertise in bioinformatics or phylogeography. We will show its functionality by performing case studies of different viruses of public health concern including influenza and rabies virus. We will also provide its URL for user feedback by ISDS delegates.

Introduction

Sequence-informed surveillance is now recognized as an important extension to the monitoring of rapidly evolving pathogens [2]. This includes phylogeography, a field that studies the geographical lineages of species including viruses [3] by using sequence data (and relevant metadata such as sampling location). This work relies on bioinformatics knowledge. For example, the user first needs to find a relevant sequence database, navigate through it, and use proper search parameters to obtain the desired data. They also must ensure that there is sufficient metadata such as collection date and sampling location. They then need to align the sequences and integrate everything into specific software for phylogeography. For example, BEAST [4] is a popular tool for discrete phylogeography. For proper use, the software requires knowledge of phylogenetics and utilization of BEAUti, its XML processing software. The user then needs to use other software, like TreeAnnotator [4], to produce a single ("representative") maximum clade credibility (MCC) tree. Even then, the evolutionary spread of the virus can be difficult to interpret via a simple tree viewer. There is software (such as SpreaD3 [5]) for visualizing a tree within a geographic context, yet for novice users, it might not be easy to use. Currently, there are only a few systems designed to automate these types of tasks for virus surveillance and phylogeography.

Methods

We have developed ZooPhy, a pipeline for sequence-informed surveillance and phylogeography [1]. It is designed for health agency personnel that do not have expertise in bioinformatics or phylogeography. We created a large database of all virus sequences and metadata from GenBank [6] as well as a smaller database for selected viruses perceived to be of great interest for health agencies including: influenza (A, B, and C), Ebola, rabies, West Nile virus, and Zika virus.

In Figure 1A, we show our front-end architecture, created in the style of the influenza research database [7], that enables the user to search by: virus, gene name, host, time-frame, and geography. We also allow users to upload their own list of GenBank accessions or unpublished sequences. Hitting "Search" produces a Results tab which includes the metadata of the sequences. We provide a feature to randomly down-sample by a specified percentage or number. We also allow the user to download the metadata in CSV format or the unaligned sequences in FASTA format.

The final tab, "Run", includes a text box for specifying an email in order to send job updates and final results on virus spread. We also enable for the user to study the influence of predictors on virus spread (via a generalized linear model). Currently, we have predictors such as temperature, great circle distance, population, and sample size for selected countries. We also offer experts the ability to specify advanced modeling parameters including the molecular clock type (strict vs. relaxed), coalescent tree prior, and chain length and sampling frequency for the Markov-chain Monte Carlo. When the user selects "Start ZooPhy", a pre-processor eliminates incomplete or non-disjoint record locations and sends the rest for analysis.

Results

When initiated, the ZooPhy pipeline includes sequence alignment via Mafft [8] and creation of an XML template via BEASTGen for input into BEAST for discrete phylogeography. It then uses TreeAnnotator [3] to create an MCC tree from the posterior distribution of sampled trees. ZooPhy uses the MCC as input into SpreaD3 for a recreation of the time-estimated migration via a map. If the user selects the GLM option, the system runs an R script to calculate the Bayes factor of the inclusion probability for

ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

each predictor and draws a plot including the regression coefficient and its 95% Bayesian credible interval. We are currently working on new visualization techniques such as those demonstrated by Dudas et al. that combine time-oriented spread via a map and evolution on a phylogenetic tree annotated by discrete locations [9].

Conclusions

Recent advances in phylodynamics, bioinformatics, and visualization have demonstrated the potential of pipelines to support surveillance. One example is NextStrain which can perform real-time virus phylodynamics [10]. The system has recently been added as an app to the Global Initiative on Sharing Avian Influenza Data (GISAID) database for influenza tracking using DNA sequences [11]. This presentation will highlight a pipeline for virus phylogeography designed for epidemiologists who are not experts in bioinformatics but wish to leverage virus sequence data as part of routine surveillance. We will describe the development and implementation of our system, ZooPhy, and use real-world case studies to demonstrate its functionality. We invite ISDS delegates to use the system via our web portal, https://zodo.asu.edu/zoophy/ and provide feedback on system utilization.

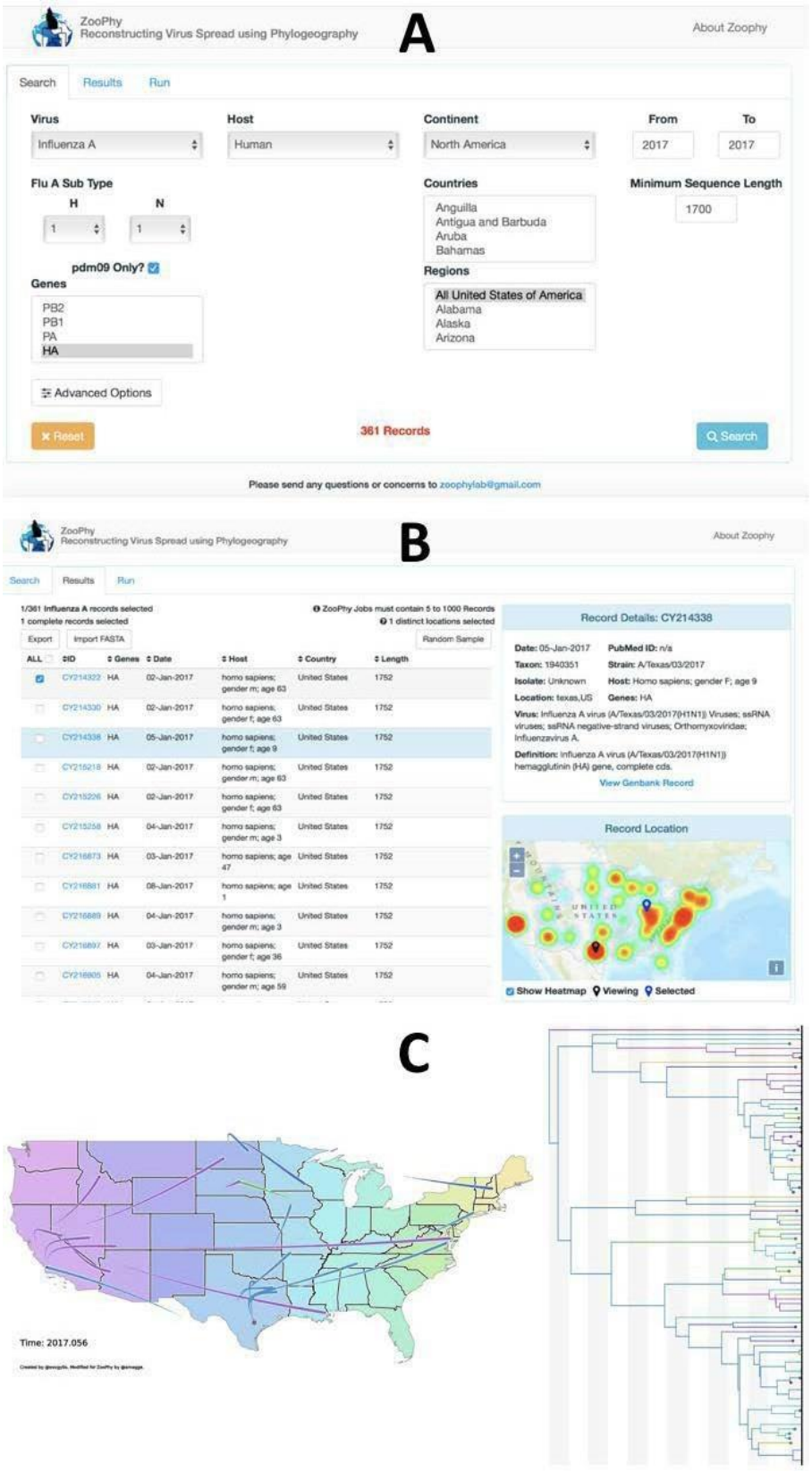
Acknowledgement

This work was supported by the National Library of Medicine of the NIH under award R01LM012080 (to MS).

References

1. Scotch M, Mei C, Brandt C, Sarkar IN, Cheung K. 2010. At the intersection of public-health informatics and bioinformatics: using advanced Web technologies for phylogeography. *Epidemiology*. 21(6), 764-68. PubMed https://doi.org/10.1097/EDE.0b013e3181f534dd
2. Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet*. 19, 9-20. PubMed https://doi.org/10.1038/nrg.2017.88
3. Avise JC. Phylogeography: the history and formation of species. 2000, Cambridge, Mass.: Harvard University Press.
4. Suchard MA, Lemey P, Baele G, et al. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. 4(1), vey016. PubMed https://doi.org/10.1093/ve/vey016
5. Bielejec F, Baele G, Vrancken B, et al. 2016. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol Biol Evol*. 33(8), 2167-69. PubMed https://doi.org/10.1093/molbev/msw082
6. Benson DA, Cavanaugh M, Clark K, et al. 2018. GenBank. *Nucleic Acids Res*. 46, D41-47. PubMed https://doi.org/10.1093/nar/gkx1094
7. Zhang Y, et al. 2017. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res*. 45, D466-74. PubMed https://doi.org/10.1093/nar/gkw857
8. Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. *Methods Mol Biol*. 1079, 131-46. PubMed https://doi.org/10.1007/978-1-62703-646-7_8
9. Dudas G, et al. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. 544(7650), 309-15. PubMed https://doi.org/10.1038/nature22040
10. Hadfield J, Megill C, Bell SM, Huddleston J, et al. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 34(23), 4121-23. PubMed https://doi.org/10.1093/bioinformatics/bty407
11. NextFlu. 2018; Available from: https://www.gisaid.org/epiflu-applications/nextflu-app/.

ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.



ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure1. ZooPhy search portal. A) A search for pdm09 H1N1 hemagglutinin (HA) sequences for 2017 in the U.S. B) The returned search result, including a U.S. heatmap of 361 pdm09 sequences. Before running the analytic pipeline, we show the geographic distribution of samples and enable the user to download the metadata and unaligned (FASTA) sequence file. C) We demonstrate geospatial results pertaining to spread and evolution that we will soon implement into ZooPhy.

ISDS Annual Conference Proceedings 2019. This is an Open Access article distributed under the terms of the Creative Commons AttributionNoncommercial 4.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.