The Journal of Open Source Software

Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens

John Huddleston[1, 2], James Hadfield[2], Thomas R. Sibley[2], Jover Lee[2], Kairsten Fay[2], Misja Ilcisin[2], Elias Harkins[2], Trevor Bedford[1, 2], Richard A. Neher[3, 4], and Emma B. Hodcroft[3, 4, 5]

1 Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA **2** Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA **3** Biozentrum, University of Basel, Basel, Switzerland **4** Swiss Institute of Bioinformatics, Basel, Switzerland **5** Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

DOI: 10.21105/joss.02906

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @dcnickle
- @Maghnuso

Submitted: 09 December 2020
Published: 07 January 2021

Summary and statement of need

The analysis of human pathogens requires a diverse collection of bioinformatics tools. These tools include standard genomic and phylogenetic software and custom software developed to handle the relatively numerous and short genomes of viruses and bacteria. Researchers increasingly depend on the outputs of these tools to infer transmission dynamics of human diseases and make actionable recommendations to public health officials (Black et al., 2020; Gardy et al., 2015). In order to enable real-time analyses of pathogen evolution, bioinformatics tools must scale rapidly with the number of samples and be flexible enough to adapt to a variety of questions and organisms. To meet these needs, we developed Augur, a bioinformatics toolkit designed for phylogenetic analyses of human pathogens.

Augur originally existed as an internal component of the nextflu (Neher & Bedford, 2015) and Nextstrain (Hadfield et al., 2018) applications. As a component of nextflu, Augur consisted of a single monolithic Python script that performed most operations in memory. This script prepared a subset of seasonal influenza sequences and metadata and then processed those data to produce an annotated phylogeny for visualization in the nextflu web application. When Nextstrain replaced nextflu and expanded to support multiple viral and bacterial pathogens, each pathogen received its own copy of the original script. The resulting redundancy of these large scripts complicated efforts to debug analyses, add new features for all pathogens, and add support for new pathogens. Critically, this software architecture led to long-lived, divergent branches of untested code in version control that Nextstrain team members could not confidently merge without potentially breaking existing analyses.

Implementation

To address these issues, we refactored the original Augur scripts into a toolkit of individual subcommands wrapped by a single command line executable, `augur`. With this approach, we followed the pattern established by samtools (Li et al., 2009) and bcftools (Li, 2011) where subcommands perform single, tightly-scoped tasks (e.g., "view," "sort," "merge," etc.) that can be chained together in bioinformatics pipelines. We migrated or rewrote the existing functionality of the original Augur scripts into appropriate corresponding Augur subcommands. To enable interoperability with existing bioinformatics tools, we designed subcommands to accept inputs and produce outputs in standard bioinformatics file formats wherever possible. For example, we represented all raw sequence data in FASTA format, alignments in either

FASTA or VCF format, and phylogenies in Newick format. To handle the common case where a standard file format could not represent some or all of the outputs produced by an Augur command, we implemented a lightweight JSON schema to store the remaining data. The "node data" JSON format represents one such Augur-specific file format that supports arbitrary annotations of phylogenies indexed by the name assigned to internal nodes or tips. To provide a standard interface for our own analyses, we also designed several Augur subcommands to wrap existing bioinformatics tools including `augur align` (mafft (Katoh et al., 2002)) and `augur tree` (FastTree (Price, 2010), RAxML (Stamatakis, 2014), and IQ-TREE (Nguyen et al., 2014)). Many commands including `augur refine`, `traits` and `ancestral` make extensive use of TreeTime (Sagulenko et al., 2018) to provide time-scaled phylogenetic trees or further annotate the phylogeny.

By implementing the core components of Augur as a command line tool, we were able to rewrite our existing pathogen analyses as straightforward bioinformatics workflows using existing workflow management software like Snakemake (Köster & Rahmann, 2012). Most pathogen workflows begin with user-curated sequences in a FASTA file (e.g., `sequences.fasta`) and metadata describing each sequence in a tab-delimited text file (e.g., `metadata.tsv`). Users can apply a series of Augur commands and other standard bioinformatics tools to these files to create annotated phylogenies that can be viewed in Auspice, the web application that serves Nextstrain (Figure 1). This approach allows users to leverage the distributed computing abilities of workflow managers to run multiple steps of the workflow in parallel and also run individual commands that support multiprocessing in parallel. Further, the Augur modules can be easily recombined both with each other and with user-generated scripts to flexibly address the differing questions and restrictions posed by a variety of human pathogens.

The modular Augur interface has enabled phylogenetic and genomic epidemiological analyses by academic researchers, public health laboratories, and private companies. Most recently, these tools have supported the real-time tracking of SARS-CoV-2 evolution at global and local scales (Alm et al., 2020; Bedford et al., 2020; The Nextstrain Team, 2020). This success has attracted contributions from the open source community that have allowed us to improve Augur's functionality, documentation, and test coverage. To facilitate Augur's continued use as part of wider bioinformatics pipelines in public health, we have committed to work with and contribute to open data standards such as PHA4GE (Griffiths et al., 2020) and follow recommendations for open pathogen genomic analyses (Black et al., 2020). Augur can be installed from PyPI (nextstrain-augur) and Bioconda (augur). See the full documentation for more details about how to use or contribute to development of Augur.

Figures

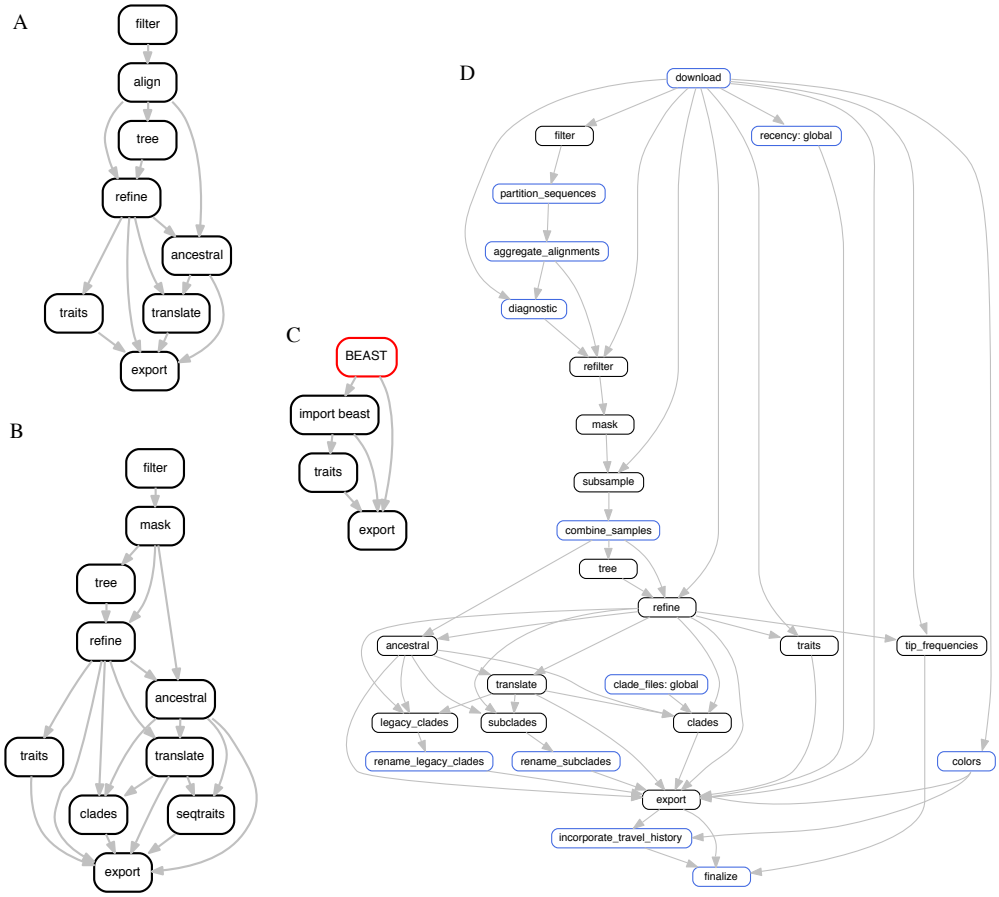

Figure 1: Example workflows composed with Snakemake from Augur commands for A) Zika virus, B) tuberculosis, C) a BEAST analysis, and D) the Nextstrain SARS-CoV-2 pipeline as of 2020-11-27. Each node in the workflow graph represents a command that performs a specific part of the analysis (e.g., aligning sequences, building a tree, etc.) with Augur commands in black, external software in red, and custom scripts in blue. A typical workflow starts by filtering sequences and metadata to a desired subset for analysis followed by inference of a phylogeny, annotation of that phylogeny, and export of the annotated phylogeny to a JSON that can be viewed on Nextstrain. Workflows for viral (A) and bacterial (B) pathogens follow a similar structure but also support custom pathogen-specific steps. Augur's modularity enables workflows that build on outputs from other tools in the field like BEAST (C) as well as more complicated analyses such as that behind Nextstrain's daily SARS-CoV-2 builds (D) which often require custom scripts to perform analysis-specific steps. Multiple outgoing edges from a single node represent opportunities to run the workflow in parallel. See the full workflows behind A, B, and D at https://github.com/nextstrain/zika-tutorial, https://github.com/nextstrain/tb, and https://github.com/nextstrain/ncov.

Acknowledgments

Thank you to all of the open source community members who have contributed to Augur. Specifically, we thank Eric Danielson, Eddie Lebow, Barney Potter, Ryan Grout, Sai Kiran Kollapudi, Mingye Wang, Carol Willing, Louise Moncla, Thomas Caswell, Sidney Bell, Terry Jones, Christian Clauss, Julien Bordellier, Gytis Dudas, Cameron Devine, Samuel Zhang, Akshay Subramanian, Christopher Tomkins-Tinch, Danielle Kain, Pierre Barrat-Charlaix, Rhys

Kidd, Chris Woszczak, Tony Tung, Mathias Walter, and Zachary Sailer. Thank you to Dan Fornika from BCCDC Public Health Laboratory for creating the first conda recipe for Augur in Bioconda. JHu is a Graduate Research Fellow and is supported by the NIH grant NIAID F31AI140714. TB is a Pew Biomedical Scholar. RAN and EBH are supported by University of Basel core funding. This work is supported by NIH awards NIGMS R35 GM119774-01, NIAID U19 AI117891-01 and NIAID R01 AI127893-01.

References

Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., Melidou, A., Neher, R. A., O'Toole, Á., Pereyaslov, D., & The WHO European Region Sequencing Laboratories and GISAID EpiCoV Group. (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance*, *25*(32), 2001410. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410

Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., Shrestha, L., Nguyen, T. N., Adler, A., Brandstetter, E., Cho, S., Giroux, D., Han, P. D., Fay, K., Frazar, C. D., … Jerome, K. R. (2020). Cryptic transmission of SARS-CoV-2 in Washington state. *Science*. https://doi.org/10.1126/science.abc0523

Black, A., MacCannell, D. R., Sibley, T. R., & Bedford, T. (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, *26*(6), 832–841. https://doi.org/10.1038/s41591-020-0935-z

Gardy, J., Loman, N. J., & Rambaut, A. (2015). Real-time digital pathogen surveillance — the time is now. *Genome Biology*, *16*, 155. https://doi.org/10.1186/s13059-015-0726-x

Griffiths, E. J., Timme, R. E., Page, A. J., Alikhan, N.-F., Fornika, D., Maguire, F., Mendes, C. I., Tausch, S. H., Black, A., Connor, T. R., Tyson, G. H., Aanensen, D. M., Alcock, B., Campos, J., Christoffels, A., Silva, A. G. da, Hodcroft, E., Hsiao, W. W. L., Katz, L. S., … MacCannell, D. R. (2020). *The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology*. https://doi.org/10.20944/preprints202008.0220.v1

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, bty407. https://doi.org/10.1093/bioinformatics/bty407

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000. G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Neher, R. A., & Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, *31*(21), 3546–3548. https://doi.org/10.1093/bioinformatics/btv381

Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von, & Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Price, P. S. A. A., Morgan N. AND Dehal. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, *5*(3), 1–10. https://doi.org/10.1371/journal.pone.0009490

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vex042

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

The Nextstrain Team. (2020). *Nextstrain/ncov*. https://github.com/nextstrain/ncov