Investigating the Genomic Distribution of Phylogenetic Signal with CloudForest

Reid Wagner
Minnesota Supercomputing Institute
The University of Minnesota
Minneapolis, Minnesota, USA
wagnerr@umn.edu

Benjamin S. Toups
Dept. of Biological Sciences
Louisiana State University
Baton Rouge, Louisiana, USA
btoup15@lsu.edu

Zhifeng Deng
Dept. of Mathematics
Florida State University
Tallahassee, Florida, USA
zd16d@my.fsu.edu

Kyle A. Gallivan
Dept. of Mathematics
Florida State University
Tallahassee, Florida, USA
kgallivan@fsu.edu

Jeremy M. Brown
Dept. of Biological Sciences and
Museum of Natural Science
Louisiana State University
Baton Rouge, Louisiana, USA
jembrown@lsu.edu

James C. Wilgenbusch
Minnesota Supercomputing Institute
The University of Minnesota
Minneapolis, Minnesota, USA
jwilgenb@umn.edu

ABSTRACT

A central focus of evolutionary biology is inferring the historical relationships among species and using this context to learn about how evolution has shaped diverse organisms. These historical relationships are represented by phylogenetic trees, and the methods used to infer these trees have been an active area of research for several decades. Despite this attention, phylogenetic workflows have changed little, even though extraordinary advances have occurred in the scale and pace at which genomic data have been collected in the past 20 years. Modern phylogenomic datasets have also raised fascinating new questions. Why do different parts of a genome often support different relationships among species? How are these different signals distributed across chromosomes? We developed a new computational framework, CloudForest, to tackle such questions. CloudForest is flexible, efficient, and tightly integrates a diverse set of tools. Here, we briefly describe the architecture of CloudForest, including the advantages it provides, and use it to investigate the distribution of phylogenetic signal along the entire X chromosome of 24 cat (Felidae) species.

CCS CONCEPTS

• **Software and its engineering** → **Software creation and management**; • **Computer systems organization** → *Architectures*; • **Applied computing** → **Computational biology**.

KEYWORDS

CloudForest, TreeScaper, containers, reproducibility, phylogenomics, HPC, Galaxy, framework, workflows, CIPRES

PEARC '21, July 18–22, 2021, Boston, MA, USA

ACM ISBN 978-1-4503-8292-2/21/07.
https://doi.org/10.1145/3437359.3465605

ACM Reference Format:
Reid Wagner, Benjamin S. Toups, Zhifeng Deng, Kyle A. Gallivan, Jeremy M. Brown, and James C. Wilgenbusch. 2021. Investigating the Genomic Distribution of Phylogenetic Signal with CloudForest. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465605

1 INTRODUCTION

Understanding the historical relationships among different species provides powerful context for learning about evolutionary processes and patterns of diversification. To infer these phylogenetic relationships, evolutionary biologists today typically use DNA sequences sampled from across the genomes of their species of interest. Modern phylogenomic datasets can contain sequences from dozens, hundreds, or thousands of different genomic regions. Some studies even use entire genomes to infer phylogenetic trees. Genomic data can provide enormous amounts of information about evolutionary history, but they are also complex and heterogeneous.

Many recent studies have shown that genes can vary extensively in the phylogenetic trees that they support, even when these genes are sampled from the same genomes. Some of this variation seems to be driven by biological processes (e.g., incomplete lineage sorting and horizontal gene transfer) [11], but some also seems to be caused by the analytical challenges of reconstructing ancient relationships [10, 13]. Given this striking variation, understanding how phylogenetic signal is distributed across genomes is an important goal of phylogenomics. What, if any, is the relationship between a gene's position in a genome (i.e., its genomic context) and its phylogenetic signal (i.e., the phylogenetic tree that it supports)?

To address this, and other pressing questions in phylogenomics, phylogenetic researchers need analytical frameworks that can deal with large, complicated data sets (of both sequences and trees), that run efficiently, that tightly integrate different tools, and that allow intuitive, visual exploration of interesting patterns. Currently, most phylogenetic analyses are run in a piecemeal fashion that requires users to set up their own workflows from scratch, often separately for different computing platforms, and to reformat output files from one program for input to another in long chains. These workflows

are easily disrupted as different software components change, are generally not optimized for the computing platforms on which they are run, and offer very few opportunities for visual exploration. Current approaches also do not permit iterative workflows where a "human-in-the-loop" can interact with the results generated at different stages of the analysis and switch course based on what is learned at the various intermediate steps.

CloudForest was developed to address these challenges and limitations and to flexibly and robustly exploit the ever-changing set of computing resources available to the academic research community. CloudForest is comprised of an integrated and accessible suite of phylogenomic tools designed to make complex evolutionary analyses more accessible, flexible, intuitive, dynamic, and reproducible. Here, we describe the architecture of CloudForest and demonstrate some of the advantages it offers both in terms of computational speed and convenience. We demonstrate these advantages by using CloudForest to investigate the distribution of phylogenomic signal across >1,200 different regions of the X chromosome in 24 species of cats [6]. Our analysis reveals a striking relationship between physical location and phylogenetic signal across these regions.

2 ARCHITECTURE AND SCALING

CloudForest exists as a suite of programs all housed within a singular framework. When designing such a platform, we considered all options for its underlying framework. A main consideration when making this decision for CloudForest was how to best address the continuum of needs ranging from desktop workstations to HPC resources, in addition to designing a platform with the desired flexibility and ease of integration. Another key consideration was the long term sustainability of the proposed framework. For example, to what extent do existing tools have the widespread support of a community of developers and how well do potential solutions accommodate new and innovative functionality? After weighing different options, we decided on the Galaxy framework [1].

Galaxy represents a largely ready made system that only requires the implementation of individual programs in order to provide basic functionality. While other options may necessitate securing funding for general maintenance, the Galaxy framework has a large active user and developer base with several sources of funding for the upkeep and basic maintenance of the framework itself. This robust network of users and developers allows the CloudForest development team to focus their efforts on the implementation of more domain specific tools and features.

Galaxy possesses many ready-made features that lend themselves particularly well to the uses of CloudForest, such as saveable, customizable, and modular workflows that increase the ease of use and flexibility. Other tools, such as a built-in, extensible visualization framework [2] for sets of phylogenetic trees and nonlinear dimensionality reduction (NLDR) [5] plots for both high and low dimensional space, provide opportunities for more integrated and intuitive interactive analyses. Galaxy also allows easy implementation for customized programs made to visualize other aspects of gene tree variation analyses, such as covariance and affinity networks. All of these qualities and features make Galaxy well-suited for a tool such as CloudForest.

Phylogenomic analyses, such as those outlined here, often require large amounts of computing resources, and as such are often outsourced to HPC clusters for particularly demanding jobs. In addition to the benefits outlined previously, the Galaxy framework also allows for direct access to HPC compute resources as well as large RAM and CPU counts on normal workstations. Since CloudForest is contained within a containerized version of Galaxy using Docker [7], its tools have access to the virtual CPUs and RAM provided by Docker, as defined by the user, with thread and process alignment all handled by the framework.

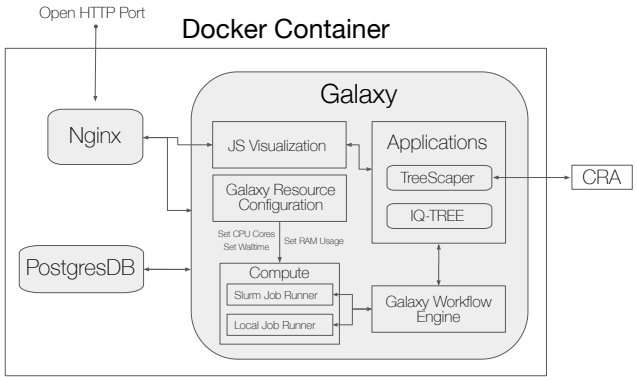

Figure 1: Basic overview of the CloudForest architecture as it relates to desktop workstations and HPC resources.

CloudForest can access HPC resources in multiple ways. Two solutions for in-house HPC are the Slurm Workload Manager [3] or Condor [12]. For these, the user must have access to the HPC compute resource, including file system access and privileged administrative assistance on the target HPC system. The Docker container would then need to be started with all of the environment variables needed for HPC cluster processing. One scalable solution to address the need for computing beyond what is available in most desktop resources is via the CIPRES REST API (CRA) [8]. The CRA provides users with HPC compute access for a multitude of phylogenetic tools. The CloudForest – Galaxy application allows a user to provide their own third-party CRA credentials, and offload computation to the service. This communication is handled by TreeScaper, which parses a configuration file defining the tool and parameters. The longer term road map of this approach entails the expansion of HPC access for CloudForest to include full containerization of tools on HPC resources. This will involve packaging the tools, libraries, and input data needed to complete a computation within a Singularity container. This package will then be sent to a user-accessible HPC resource. For many tools, using CloudForest on a highly performant workstation can be sufficient; however, there is undoubtedly a need for such HPC integration, and this remains a priority of the development team.

3 EXAMPLE ANALYSIS - PHYLOGENETIC SIGNAL ACROSS THE X CHROMOSOME

To demonstrate the advantages provided by CloudForest for addressing pressing questions in phylogenomics, we analyzed a recently

published dataset that included dense sampling of different genomic regions from more than two dozen cat genomes [6]. We focused on the distribution of phylogenetic signal across the X chromosome, in particular, given evidence that recombination rate variation across different regions of the X is linked to differences in phylogenetic signal [6]. Li et al. [6], and other authors who have investigated such relationships in other species, had to create their own custom workflows to first infer phylogenetic trees for different regions, compare the resulting trees (guided by a priori knowledge of their group), and then create de novo graphics to summarize the results. With CloudForest, we were able to utilize a self-contained workflow, offload computation to external HPC resources, and quickly summarize and visualize the results.

To investigate phylogenetic signal across the X chromosome, we analyzed 1,273 100-kilobase windows of DNA that spanned the entire length of the X chromosome of 24 cat species. We divided these windows into a few groups, uploaded them into CloudForest, and created a separate Galaxy collection from each group of sequences. We then used each collection as input to the TreeScaper tool, which submitted IQ-TREE [9] jobs to CIPRES using the CRA for each sequence in the collection, with up to 50 jobs running in parallel. TreeScaper collected the resulting maximum-likelihood trees returned by the CRA, which were concatenated, sorted, and passed through text processing tools within CloudForest to filter failed results and extra information, producing a single list of trees. Twelve windows could not be analyzed due to missing sequences, leaving us with 1,261 trees in total.

From the resulting set of trees, we used TreeScaper [4] within CloudForest to calculate a matrix of weighted Robinson-Foulds (wRF) distances between trees, which provides information about the extent of variation in phylogenetic signal across different genomic regions. To better understand this variation, we used TreeScaper to perform non-linear dimensionality reduction (NLDR) [5, 14]. NLDR allows users to visualize sets of phylogenetic trees in Euclidean space by projecting trees as points in two or three dimensions, providing an intuitive summary of any structure present in the set. Visual inspection of the NLDR plot revealed an interesting J-shaped structure to the trees, including a region in which a group of trees was differentiated from the rest (see Fig. 2). Following NLDR, we calculated an affinity matrix by taking the reciprocal of the wRF distances and used these affinities to perform community detection (with the Configuration Null Model) in order to identify any signal of structure in the tree set [4].

Community detection analysis supported a two-community structure roughly matching the distribution of trees in the NLDR plot (see Fig. 2). After identifying this community structure, we mapped the trees in each community back to their location on the X chromosome and found that the trees had a striking relationship to a region's physical position on the X chromosome. More specifically, one community corresponded mostly to regions in the center of the X chromosome, while the other was mostly comprised of regions more distal on the X (i.e., near one end or the other) (see bottom of Fig. 2). Biologically, this result is important, as the center of the X chromosome is known to have a particularly low recombinaton rate. The relationship between recombination rate and phylogenetic signal sheds important light on the causes of gene

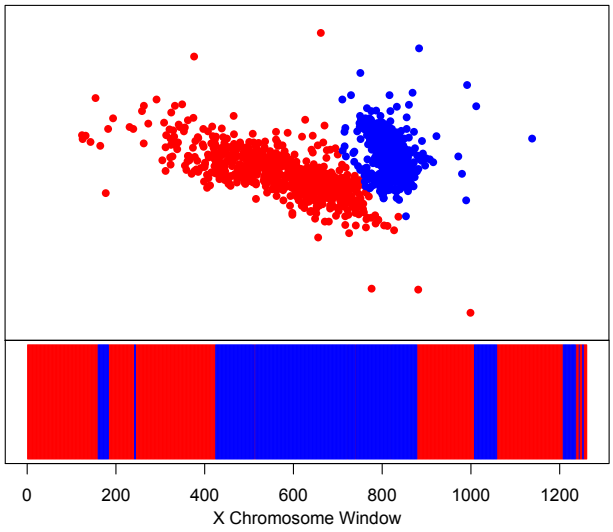

Figure 2: (Top) An NLDR plot of the maximum-likelihood trees from different regions of the X chromosome across 24 cat species, with individual trees colored according to the community in which they were placed. (Bottom) A representation of the ordering of the 1,261 windows along the X chromosome with the colors again depicting the communities to which a tree from each window was assigned.

tree variation in phylogenomic studies, suggesting that low recombination regions are resistant to introgression and more likely to reflect the true phylogenetic relationships among species [6].

Essentially, this entire analysis was conducted using a self-contained workflow in CloudForest that utilized built-in analytical and visualization tools. An analysis such as the one presented here shows only some of the types of interplay between programs, user interactivity, and novel visualization schemes that can be used to explore and better understand important patterns in phylogenomic datasets. Outside of CloudForest, the process of exploring similar datasets not only requires expertise in compiling, installing, and integrating various software packages, but it is also difficult to repeat these types of analyses. By packaging all the required components within a container and under a single, intuitive framework, CloudForest gives researchers a flexible means to connect programs into larger workflows, with intermediate outputs that can help to direct and guide more complex analyses. Importantly, these workflows can be saved and shared, so that others can explore and validate results and extend approaches to address novel questions that depend on complex evolutionary processes.

Although we did not perform a rigorous benchmark, the built-in offloading of parallel-running jobs to HPC resources via the CRA significantly reduced the time to perform the analysis. It took 8 hours and 43 minutes to run tree inference for a subsample of 319 sequences with the local IQ-TREE Galaxy tool, utilizing 4 cores of a 2GHz Quad-Core Intel Core i5, and 8GB RAM. The same analysis run with IQ-TREE via the CRA-integrated TreeScaper tool completed in less than one-fifth of the time with a duration of 1 hour and 31 minutes.

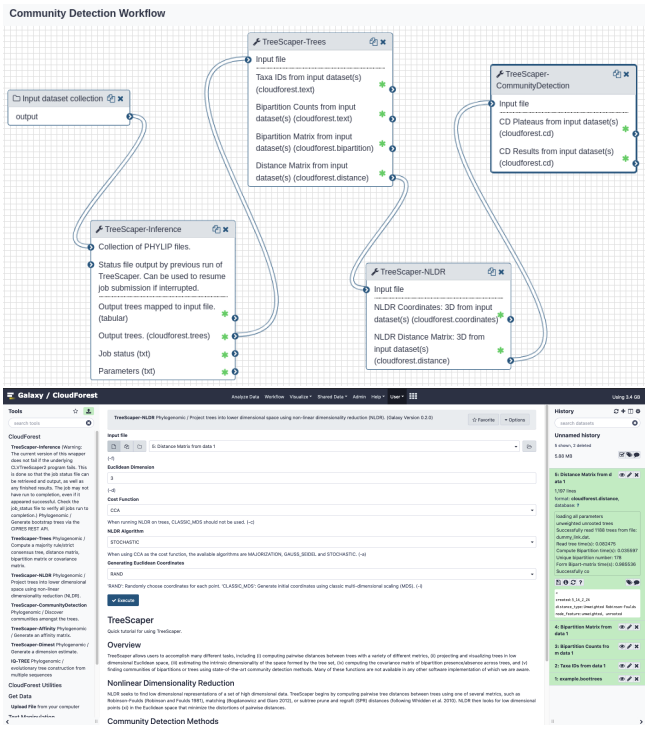

Figure 3: (Top) A screenshot of the native workflow editor, with a simplified version of the analysis in this paper. (Bottom) A screenshot of the configuration page to run a job with TreeScaper, displaying the options found in the native program.

4 CONCLUDING REMARKS

Most modern phylogenetic inference requires an increasingly large set of specialized software tools to explore the variation present in large genomic-scale datasets. To date, integrating results of these independent analyses has not been intuitive, which has made it difficult or impossible to reproduce or verify the results of these complex analyses. Additionally, most programs currently lack an integrated and interactive approach to visualizing results at all stages of these complex workflow in a way that could help inform subsequent analysis approaches. CloudForest leverages the use of a popular analysis framework to orchestrate large-scale, complex phylogenomic analyses. The approach described in this paper makes otherwise complex analyses more accessible to a broader set of practitioners, while explicitly addressing growing concerns over the reproducibility of scientific research.

Development efforts for CloudForest are currently ongoing, with an open-source beta version available. With support from NSF and continued development efforts, a prototype version of CloudForest will be circulated to a group of beta testers to gather feedback and continue improving the platform prior to its full release.

ACKNOWLEDGMENTS

We would like to thank T. McGowan for his contributions to the original design and implementation of CloudForest, W.J. Murphy for providing the 24-cat species sequence data, and M. Miller, W. Pfeiffer, and M. Zhuang for their valuable guidance in using the CRA. The development of CloudForest is supported by National Science Foundation grants DBI-1934182 to JCW, DBI-1934157 to KAG, and DBI-1934156 to JMB.

REFERENCES

[1] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 Update. *Nucleic Acids Research* 46 (2018), W537–W544.

[2] Eberhard C. Too T. et al. Goecks, J. 2013. Web-based visual analysis for high-throughput genomics. *BMC Genomics* 14 (2013). https://doi.org/10.1186/1471-2164-14-397

[3] Björn Grüning. 2014. GitHub - bgruening/docker-galaxy-stable: Docker Images tracking the stable Galaxy releases. https://github.com/bgruening/docker-galaxy-stable.

[4] Wen Huang, Guifang Zhou, Melissa Marchand, Jeremy R. Ash, David Morris, Paul Van Dooren, Jeremy M. Brown, Kyle A. Gallivan, and Jim C. Wilgenbusch. 2016. TreeScaper: Visualizing and Extracting Phylogenetic Signal from Sets of Trees. *Molecular Biology and Evolution* 33, 1 (2016), 3314–16.

[5] John A. Lee and Michel Verleysen. 2007. *Nonlinear Dimensionality Reduction*. Springer-Verlag New York.

[6] Gang Li, Henrique V Figueiró, Eduardo Eizirik, and William J Murphy. 2019. Recombination-Aware Phylogenomics Reveals the Structured Genomic Landscape of Hybridizing Cat Species. *Molecular Biology and Evolution* 36, 10 (06 2019), 2111–2126. https://doi.org/10.1093/molbev/msz139

[7] Dirk Merkel. 2014. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal* 2014, 239 (2014), 2. https://doi.org/10.1097/01.NND.0000320699.47006.a3

[8] Mark A. Miller, Wayne Pfeiffer, and Terri Schwartz. 2010. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. *2010 Gateway Computing Environments Workshop (GCE)* (2010), 1–8.

[9] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32, 1 (2015), 268–74.

[10] Emilie J. Richards, Jeremy M. Brown, Anthony J. Barley, Rebecca A. Chong, and Robert C. Thomson. 2018. Variation Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much Gene Tree Variation Is Biological? *Systematic Biology* 67, 5 (2018), 847–60.

[11] Leonidas Salichos and Antonis Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497 (05 2013). https://doi.org/10.1038/nature12130

[12] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. 2001. *Condor – A Distributed Job Scheduler*. MIT Press.

[13] James C. Wilgenbusch and Kevin de Queiroz. 2000. Phylogenetic Relationships Among the Phrynosomatid Sand Lizards Inferred from Mitochondrial DNA Sequences Generated by Heterogeneous Evolutionary Processes. *Systematic Biology* 49, 3 (2000), 592–612.

[14] James C. Wilgenbusch, Wen Huang, and Kyle A. Gallivan. 2017. Visualizing Phylogenetic Tree Landscapes. *BMC Bioinformatics* 18, 85 (2017).