Bioinformatics, 34(16), 2018, 2848–2850
doi: 10.1093/bioinformatics/bty173
Advance Access Publication Date: 19 March 2018
Applications Note

OXFORD

Phylogenetics

ALPHA: a toolkit for Automated Local PHylogenomic Analyses

R. A. Leo Elworth[1,*], Chabrielle Allen[1], Travis Benedict[1], Peter Dulworth[1] and Luay Nakhleh[1,2,*]

[1]Computer Science and [2]BioSciences, Rice University, Houston, TX 77004, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on October 12, 2017; revised on February 23, 2018; editorial decision on March 14, 2018; accepted on March 16, 2018

Abstract

Summary: The evolutionary histories of individual regions across a genomic alignment—called 'local genealogies'—can differ from each other, due to processes such as recombination. Elucidating and analyzing these local genealogies are important for a large number of inference tasks, including those pertaining to species phylogenies, evolutionary processes and trait mapping. In this paper, we present a toolkit for **a**utomated **l**ocal **ph**ylogenomic **a**nalyses, or ALPHA. The purpose of this toolkit is to provide a wide array of functionalities for automated inference of local genealogies as well as analyses based on these local genealogies. The toolkit uses sliding windows to construct local genealogies and can compute a wide array of local phylogeny based statistics, such as the D-statistic. The toolkit comes with a graphical user interface and several import/export functionalities. Over the last few decades, much emphasis in phylogenomics has been put on developing tools for inferring species phylogenies. This toolkit complements those efforts by emphasizing the 'local' aspect of phylogenomics.

Availability and implementation: ALPHA is freely available for installation and use, including source code, at https://github.com/chilleo/ALPHA.

Contact: chilleo@gmail.com or nakhleh@rice.edu

1 Introduction

The relationship between the evolutionary history of a set of genomes and the evolutionary histories of individual regions, or loci, in those genomes can be complex due to processes such as recombination and selection (Hein *et al.*, 2004). In the past three or four decades, a wide array of algorithms and software tools were developed for inferring evolutionary histories of populations or species, most of which make use of genealogies of individual loci as the input data (Nakhleh, 2013; Szöllősi *et al.*, 2014).

Local genealogies—evolutionary trees on individual genomic regions—serve as the input data to most species phylogeny inference methods. They are also useful for elucidating evolutionary processes, inferring protein function and mapping traits. For example, in Fontaine *et al.* (2015), the authors used local genealogies to postulate hypotheses about introgression between medically relevant mosquito vectors.

We report here on ALPHA, a toolkit for automated local phylogenomic analyses, whose purpose is to provide implementations for automated analyses in two categories: inference of and analyses based on local genealogies in phylogenomic studies. The toolkit currently has several functionalities for both categories of analyses, allows for multiple import/export features, and is equipped with a graphical user interface that allows for various visualizations of the analyses.

2 Materials and methods

The ALPHA toolkit is a graphical user interface (GUI) application that is built in Python and uses several popular bioinformatics libraries and tools (Cock *et al.*, 2009; Huerta-Cepas *et al.*, 2016; Hunter, 2007; Jones *et al.*, 2014; Sukumaran and Holder, 2010; Than *et al.*, 2008) to perform its inferences and analyses. ALPHA's main

strengths, aside from its quality of life features for bioinformaticians, would be its ease of use for performing analyses that require integration of several standard phylogenetic tools and techniques into a single pipeline, and for visualizing a wide array of the corresponding results, which we believe most users will find to be publication quality visualizations. For a more detailed description of all the features of ALPHA (see Section 2.3). The ALPHA Toolkit can be installed on both Windows and Mac computers.

2.1 Inference of local genealogies

The toolkit uses one of the most common techniques for inferring local genealogies: sliding a window of a fixed width across the genomic alignment and inferring a tree on the sites within the window. The user specifies a genomic alignment as input and the window size and offset to control how the window is slid across the alignment. In this context, the window offset is defined to be the number of bases to skip from the start of one window to the start of the next window. For instance, if the window size and offset are equal the windows will be adjacent and non-overlapping. Decreasing this offset would create overlap between windows. An installation of RAxML (Stamatakis, 2006) is used for building the local genealogical trees. RAxML is a ubiquitous tree building software that returns a maximum likelihood tree for a multiple sequence alignment. RAxML is extremely fast and its speed does not compromise its ability to find trees with good likelihood scores, which has led to its widespread adoption. The alignment and genealogy for each window is saved and can be used for a number of analyses.

Figure 1 demonstrates a few of the analyses and accompanying visualizations that can be performed by the toolkit.

First, the different topologies built can be visualized alongside the frequency at which they occur. The user can filter this set of topologies to only show a desired set of topologies, where the user can set a threshold for the number of the most frequently observed topologies to be visualized. These visualizations are shown in Figure 1A and B. Figure 1C shows the local genealogies by their physical locations across the genomic alignment. Additional features include statistics on informative sites within windows (Fig. 1D and E) and comparisons of local genealogies to a user-specified species phylogeny using a variety of measures and statistics such as in Figure 1F. For convenience to users, a graphical user interface is also included for converting most genome alignment formats to other common formats.

2.2 Analyses based on local genealogies

With the advent of many new approaches for obtaining local genealogies in addition to window-sliding approaches, we have also included a suite of features for quantifying the differences between competing approaches for local genealogy inference. Given the ubiquity of *ms* (Hudson, 2002) for simulating genome alignments with recombination, the user can input two files containing local genealogies for each site of an alignment, formatted in the *ms* style, and the toolkit will calculate and plot various comparison statistics. One such analysis, which compares times to most recent common ancestors of the trees for each local region, is shown in Figure 1G.

The D statistic (Durand *et al.*, 2011), also known as the 'ABBA-BABA test,' calculates how far the site patterns of mutations in local regions of a genome alignment deviate from the distribution under a null model of no introgression. This statistic is commonly used to highlight local regions that are candidates for having been gained through introgression. Though there exist many tools for calculating the D statistic across a genomic alignment, to our knowledge this

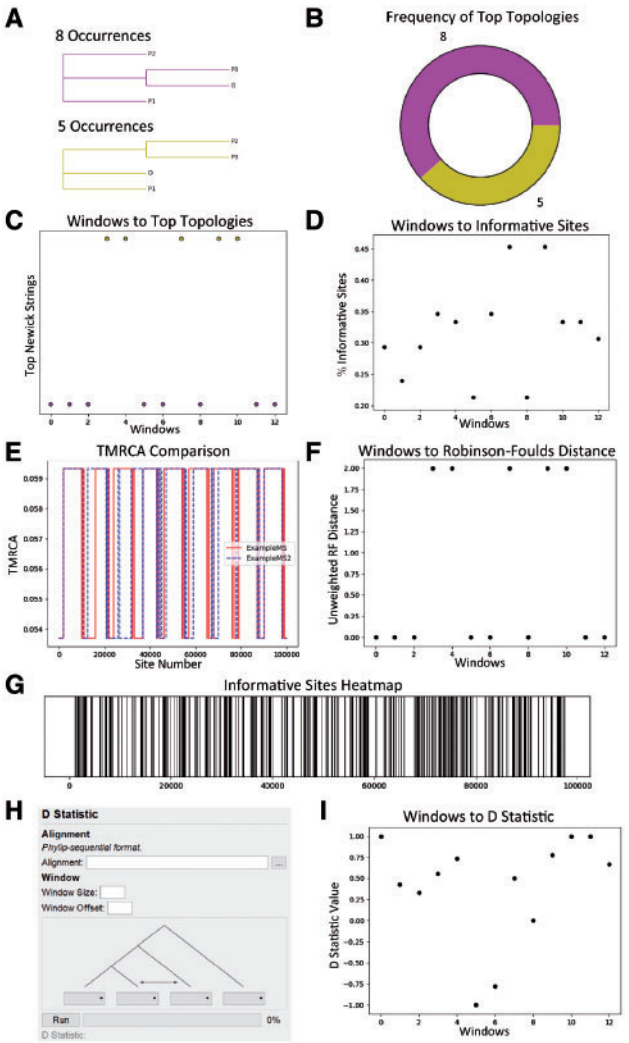

Fig. 1. A subset of the graphical capabilities of the ALPHA toolkit. (**A**) The top two most frequent topologies for local genealogies in an example sequence alignment. (**B**) A donut plot of the frequency of the two most frequent topologies for local genealogies in an example alignment. (**C**) The locations in the example alignment of the windows where the top two topologies were found. (**D**) The amount of informative sites in the alignment in each window across the alignment. (**E**) A comparison of the most recent common ancestor times between two local genealogy predictions for an alignment. (**F**) The weighted Robinson Foulds distance between each window's calculated local genealogy versus an assumed species phylogeny. (**G**) A heat map of the informative sites in a multiple sequence alignment. (**H**) The graphical user interface (GUI) for calculating the D statistic, or 'ABBA-BABA' test. (**I**) The computed D statistic values for each window across an example alignment

toolkit includes the first graphical user interface for D statistic calculations. The user inputs a four taxon genome alignment and graphically specifies the assumed species phylogeny structure with an outgroup. The toolkit then calculates and plots the D statistic values.

2.3 Full list of features

The following is a full list of the specific features of the ALPHA toolkit. This list includes both the analyses that can be performed as well as corresponding visualizations.

1. A multiple sequence alignment can be broken into windows, with RAxML run on each individual window. The user can

specify the size of the windows as well as how far to offset the amount of basepairs until the next window. An assumed species tree can be automatically generated by running RAxML on the full alignment in addition to the individual windows. If desired, both the exact RAxML command and assumed species phylogeny can be entered manually.

- The user can choose to do bootstrapping when generating a tree for a genomic window, and can set a confidence threshold for individual tree nodes based on how many bootstrap trees shared that split.
- The user can specify to run all window based analyses in a rooted fashion, where the user specifies a specific taxa to be the root in their alignment.
- The following values can be calculated and visualized:
    - The user can view the trees of the most frequent topologies of the local genealogies. Of the most frequently occurring trees, the user specifies how many they would like to view.
    - The user can visualize where the most frequent topologies occur in the sequence alignment. Again the user can specify how many top topologies they are interested in for the visualization.
    - The user can generate a donut plot of the most frequently occurring topologies.
    - The amount of informative sites within each window can be calculated and visualized.
    - A heatmap can be generated to view the density of informative sites along the genome alignment.
    - Both the weighted and unweighted Robinson-Foulds distance for each window between the local genealogy and an assumed species phylogeny can be calculated and visualized.
    - At each window, the probability of the local genealogy given an assumed species tree can be calculated and visualized.

2. A File Converter GUI is provided as a quality of life feature for converting multiple sequence alignments between common genome file formats.
3. Two files, or the previously obtained results from ALPHA and another file, containing the local genealogy at every site of a genome alignment in the *ms* style, can be compared against each other.
    - When one of the files represents the true local genealogies from a simulation performed by *ms*, the Robinson-Foulds distance from the inferred local genealogies to the true local genealogies can be calculated and visualized.
    - The percentage of matches between local genealogies for each site can be displayed.
    - A line graph can be generated comparing the time until the most recent common ancestor for each local genealogy at each site.
4. The ABBA-BABA test can be performed on individual windows of a four taxon genome sequence alignment. Again, the size of the windows and the offset between them can be specified by the user. The results can be visualized.

2.4 Future directions

There is a wide range of ways that ALPHA could be extended in the future. For example, we chose to use RAxML for building the local genealogies for each window. In principle, any tree building method could be used instead. Currently, a user can modify the source of ALPHA to adapt it to their chosen tree building software, though in the future ALPHA could be modified to work with more software packages by default. The same could be said about using the *ms* style for storing local genealogies for each site of a genomic alignment. There are many ways of doing this, and the software could be changed to accommodate other formats.

More involved future directions would include moving away from the assumption of fixed-width window-based analyses. For instance, a user could enter a map of recombination locations as a list of site locations to be used as the boundaries between windows for all current ALPHA analyses. Even more sophisticated local phylogenomic analyses, such as HMM-based (Hobolth *et al.*, 2007) or fully Bayesian (Rasmussen *et al.*, 2014) approaches, could be incorporated alongside the window based approach. Finally, additional statistics such as the split score (Allman *et al.*, 2017) can be added alongside the currently calculated statistics of the ALPHA toolkit.

Funding

This research was funded in part by NSF grants CCF-1541979 and DMS-1547433. Co-authors CA, TB and PD conducted this research as part of the Rice Undergraduate Data Science Summer Program funded generously by NSF grant DMS-1547433, Two Sigma and the Office of the Provost of Rice University.

Conflict of Interest: none declared.

References

Allman,E.S. *et al.* (2017) Split scores: a tool to quantify phylogenetic signal in genome-scale data. *Syst. Biol.*, **66**, 620–636.

Cock,P.J. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Durand,E.Y. *et al.* (2011) Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, **28**, 2239–2252.

Fontaine,M.C. *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524.

Hein,J. *et al.* (2004). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA.

Hobolth,A. *et al.* (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla from a coalescent hidden Markov model. *PLoS Genetics*, **3**, e7.

Hudson,R. (2002) Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huerta-Cepas,J. *et al.* (2016) Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.

Hunter,J.D. (2007) Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

Jones,E. *et al.* (2014) {SciPy}: open source scientific tools for {Python}. http://www.scipy.org (23 March 2018, date last accessed).

Nakhleh,L. (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.*, **28**, 719–728.

Rasmussen,M.D. *et al.* (2014) Genome-wide inference of ancestral recombination graphs. *PLoS Genet.*, **10**, e1004342.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Sukumaran,J. and Holder,M.T. (2010) Dendropy: a python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Szöllősi,G.J. *et al.* (2014) The inference of gene trees with species trees. *System. Biol.*, **64**, e42–e62.

Than,C. *et al.* (2008) Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.