Application Note

PhyloWGA: chromosome-aware phylogenetic interrogation of whole genome alignments

Richard H. Adams[1]*, Todd A. Castoe[2], and Michael DeGiorgio[1]*

[1]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL

[2]Department of Biology, University of Texas at Arlington, Arlington, TX

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Here we present *PhyloWGA*, an open source R package for conducting phylogenetic analysis and investigation of whole genome data

Availability: Available at Github (`https://github.com/radamsRHA/PhyloWGA`).

Contact: adamsr@fau.edu (R.H.A.), mdegiorg@fau.edu (M.D.)

1 Introduction

Whole genome data hold promise for refining species tree estimates because, at a fundamental level, such large datasets offer unprecedented opportunity to dramatically increase the sample size and resolution of phylogenetic analyses (White *et al.*, 2009; Jennings, 2016; Hobolth *et al.*, 2007). However, enthusiasm for conducting phylogenomic analyses across whole genome alignments (WGAs) is often curbed by the inherent challenges that arise when extracting and translating the deluge of information encoded within WGAs into meaningful estimates of evolutionary relationships.

An immediate challenge is to address the pervasive phylogenetic conflict observed in whole genome data. Both biology and methodology can generate discord among phylogenetic estimates across loci (Som, 2014), and understanding the causes and consequences of such conflict is paramount to mitigating its impacts (e.g., Reddy *et al.*, 2017). Many evolutionary processes are known to generate phylogenetic conflict (e.g., Adams *et al.*, 2018; Kutschera *et al.*, 2014), yet incomplete lineage sorting (ILS) is perhaps the most well-studied and biologically relevant (Degnan and Rosenberg, 2009; Edwards, 2009). Recombination decouples the phylogenetic histories of genomic regions, yielding unlinked genealogies that may (or may not) agree with the overall species tree due to ILS. Consequently, an entire chromosome cannot be reasonably modeled as a single non-recombining locus with only a single tree (Kubatko and Degnan, 2007), as it constitutes a mosaic of genealogies sequentially distributed along its length according to crossover events.

Even with whole genomes, tracts of contiguous well-aligned sequences are often confined to a few kb, which may be insufficient for accurate gene tree estimation. Thus, while it is possible to accommodate ILS by simply inferring separate gene trees across loci, these trees may be unreliable due to gene tree estimation error (Roch and Warnow, 2015). An ideal phylogenomic dataset would effectively accommodate both ILS and gene tree error by using sufficiently-long yet recombination-free loci. However, constructing such a dataset is difficult, and constructing "supergenes" (i.e., concatenated loci consistent with a single gene tree) can be challenging (White *et al.*, 2009; Jennings, 2016; Adams and Castoe, 2019).

Consideration of ILS, recombination, and other factors contributing to gene tree variation is therefore important for any attempt to accurately reconstruct evolutionary relationships from multilocus data for which the degree and distribution of phylogenetic conflict is unknown (Springer and Gatesy, 2016). While many methods exist for estimating individual gene trees (e.g., Stamatakis, 2014), sampling and extracting phylogenetic loci (e.g., Bravo *et al.*, 2019; Costa *et al.*, 2016), or inferring species trees (e.g., Boussau *et al.*, 2013), few bioinformatics platforms are specifically designed for conducting detailed and efficient phylogenomic interrogation of WGAs. To address this, we introduce *PhyloWGA*, an open-source R package developed as a user-friendly suite for conducting streamlined phylogenetic analysis and investigation of WGAs. At its core are two bioinformatics pipelines, *Chromo.Phylome* and *Chromo.Crawl*, designed to reconstruct a chromosome-specific set of gene trees and to apply a series of model-based phylogenetic congruency tests along the length of a WGA, respectively. *PhyloWGA* is flexible and extensible for the easy incorporation of future phylogenetic investigations of genome-scale datasets. Importantly, while other approaches have attempted to cluster loci into supergenes (e.g., Mirarab *et al*. 2014; Bayzid *et al*. 2015), recent studies have shown that such approaches may be plagued with high rates of error (Liu and Edwards 2015; Roch *et al.*, 2019; Adams and Castoe, 2019). Unlike

previous 'statistical binning' methods (Mirarab *et al*. 2014; Bayzid *et al*. 2015), which ignore linkage information (i.e., proximity of adjacent loci in assembled genomes), here we develop an approach that inherently incorporates the spatial organization of loci and uses a model-based framework to explicitly test whether such loci share a common tree (see *Supplementary Note* for further discussion).

2 Implementation

PhyloWGA is written in R 3.6.1 (R Core Team, 2018) and includes a suite of functions for WGA-scale phylogenomic analyses. It is built upon several maximum likelihood (ML) phylogenetic frameworks, including IQ-TREE (Nguyen *et al.*, 2014) and CONCATEPILLAR (Leigh *et al.*, 2008). The primary input for *PhyloWGA* is a WGA in fasta format partitioned by chromosome, and *PhyloWGA* includes a number of functions for processing WGAs for this purpose. *PhyloWGA* allows users to specify an array of experimental parameters that define the scope of a *PhyloWGA* analysis, including the size and distribution of chromosomal windows, as well as the particular type of analysis.

The function *Chromo.Phylome* uses IQ-TREE to estimate a "chromosome-specific phylome"—a set of locus-specific phylogenetic trees for a given chromosomal alignment. On completion, *Chromo.Phylome* outputs a set of phylogenetic tree models fitted to each respective genomic window. The IQ-TREE commands are customizable, such that *Chromo.Phylome* can accommodate diverse analyses, including tree inference, model selection, and model adequacy tests across a WGA. Finally, replicate analyses can be run to explore different experimental settings.

The module *Chromo.Crawl* "crawls" across a WGA, testing for congruence among contiguous windows, and concatenates them when there is insufficient evidence of gene tree disagreement. It employs the program CONCATEPILLAR, which implements a likelihood-based model test to assess whether a set of loci likely share the same phylogeny. *Chromo.Crawl* sequentially applies this model test to contiguous windows, and then concatenates these windows into a single supergene if there is evidence of congruency (see *Supplementary Note* for details). *Chromo.Crawl* will then move to the next adjacent window to assess whether it can be combined with the previous windows. This merging of contiguous windows into a supergene continues until the model test reveals discordance, with the process restarting at the current window.

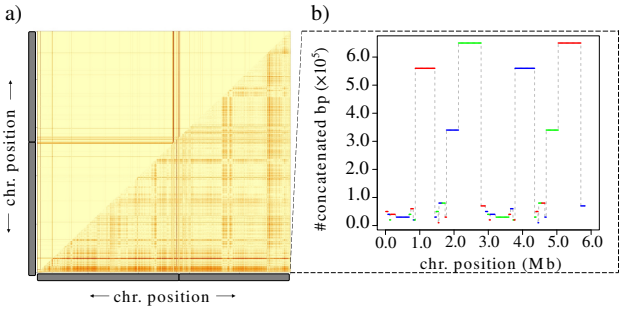

Fig. 1. (a) Heatmap of pairwise Robinson-Foulds (Robinson and Foulds, 1981; upper triangle) and geodesic (Billera *et al*., 2001; lower triangle) distances along windows of human chromosome 1 (larger distances as darker shades). (b) Number of concatenated bases indicated for each of the 38 supergenes (locations defined as alternating green, red, and blue) by applying *Chromo.Crawl* to the first six Mb of the chromosome.

3 Biological Application

We applied *PhyloWGA* on a WGA of 12 primates (Moorjani *et al.*, 2016), to characterize the phylogenetic landscape of human chromosome 1. We coarsely (100 kb windows) estimated locus phylogenetic trees by setting *Chromo.Phylome* to conduct ML-based phylogenetic analysis at each window with a GTR+Γ substitution model, and to then compute tree distances between windows. We next investigated phylogenomic conflict at a fine scale (10 kb windows) by applying *Chromo.Crawl* to track concatenated window sizes (i.e., supergene lengths) across the chromosome.

We used *PhyloWGA* to 'paint a picture' of the chromosome-wide landscape of phylogenomic conflict for this WGA. Visualizing pairwise distances among genomic regions provides a detailed depiction of conflict (Fig. 1a), indicating higher levels of conflict (darker shades) nearby telomeres and the centromere, as well as intermittent bands of high and low discordance throughout. We recovered 38 supergenes for the first six Mb of the chromosome, with a mean of ~116 kb per supergene (Fig. 1b). Though the underlying reasons for such widespread discordance in these data remain unknown, both biology and methodology likely play a role. For example, high levels of discordance observed in particular regions (i.e., darker bands of Fig. 1a and shorter supergenes in Fig. 1b) may represent poor alignment quality, regions of high recombination, or other processes unique to such regions. Both the accuracy and computational efficiency of *PhyloWGA* are influenced by the experimental design (e.g., window size and distribution) and evolutionary conditions (e.g., recombination rate and nucleotide model complexity) of the particular WGA analysis (see *Supplementary Note* for details).

Funding

This work was supported by National Science Foundation grants to M.D. (DEB-1753489 and BCS-2001063) and T.A.C. (DEB-1655571) and by a National Institutes of Health grant to M.D. (R35M128590).

Conflict of Interest: none declared.

References

Adams,R.H. *et al.* (2018) Assessing the impacts of positive selection on coalescent-based species tree estimation and species delimitation. *Syst. Biol.*, **67**, 1076–1090.

Adams,R.H. and Castoe,T.A. (2019) Statistical binning leads to profound model violation due to gene tree error incurred by trying to avoid gene tree error. *Mol. Phylogenet. Evol.,* **134**, 164–171.

Bayzid,M.S. *et al.* (2015) Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS One*, **10**, e0129183.

Billera,L.J. *et al.* (2001) Geometry of the space of phylogenetic trees. *Adv. Appl. Math.*, **27**, 733–767.

Boussau,B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.

Bravo,G.A. *et al.* (2019) Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ*, **7**, e6399.

Costa,I.R. *et al.* (2016) In silico phylogenomics using complete genomes: a case study on the evolution of hominoids. *Genome Res.*, **26**, 1257–1267.

Degnan,J.H. and Rosenberg,N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340.

Edwards,S. V (2009) Is a new and general theory of molecular systematics emerging? *Evol. Int. J. Org. Evol.*, **63**, 1–19.

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

Jennings,W.B. (2016) Phylogenomic data acquisition: principles and practice. Boca Raton, CRC Press/Taylor & Francis.

Kubatko,L.S. and Degnan,J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, **56**, 17–24.

Kutschera,V.E. *et al.* (2014) Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Mol. Biol. Evol.*, **31**, 2004–2017.

Leigh,J.W. *et al.* (2008) Testing congruence in phylogenomic analysis. *Syst. Biol.*, **57**, 104–115.

Liu,L. and Edwards,S. V (2015) Comment on “Statistical binning enables an accurate coalescent-based estimation of the avian tree”. *Science*, **350**, 171.

Mirarab, Siavash, *et al.* (2014) "Statistical binning enables an accurate coalescent-based estimation of the avian tree." *Science* **346**, 1250463.

Moorjani,P. *et al.* (2016) Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci.*, **113**, 10607–10612.

Nguyen,L.-T. *et al.* (2014) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

R Foundation for Statistical Computing. (2018) R: a Language and Environment for Statistical Computing.

Reddy,S. *et al.* (2017) Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst. Biol.,* **66**, 857–879.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Roch,S. and Warnow,T. (2015) On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst. Biol.*, **64**, 663–676.

Roch,S. *et al.* (2019) Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Syst. Biol.*, **68**, 281–297.

Som,A. (2014) Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.*, **16**, 536–548.

Springer,M.S. and Gatesy,J. (2016) The gene tree delusion. *Mol. Phylogenet. Evol.*, **94**, 1–33.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

White,M.A. *et al.* (2009) Fine-scale phylogenetic discordance across the house mouse genome. *PLoS Genet.*, **5**. 11.