Bioinformatics, 32(8), 2016, 1250–1252
doi: 10.1093/bioinformatics/btv717
Advance Access Publication Date: 9 December 2015
Applications Note

OXFORD

Genetics and population analysis

Fitchi: haplotype genealogy graphs based on the Fitch algorithm

Michael Matschiner[1,2]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway and [2]Zoological Institute, University of Basel, Vesalgasse 1, Basel 4051, Switzerland

Associate Editor: David Posada

Received on 10 June 2015; revised on 2 December 2015; accepted on 3 December 2015

Abstract

Summary: In population genetics and phylogeography, haplotype genealogy graphs are important tools for the visualization of population structure based on sequence data. In this type of graph, node sizes are often drawn in proportion to haplotype frequencies and edge lengths represent the minimum number of mutations separating adjacent nodes. I here present Fitchi, a new program that produces publication-ready haplotype genealogy graphs based on the Fitch algorithm.
Availability and implementation: http://www.evoinformatics.eu/fitchi.htm
Contact: michaelmatschiner@mac.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

For population-level sequence data of closely related individuals, haplotype genealogy graphs are commonly used to visualize population divergence, structure, and connectivity. As opposed to phylogenetic trees, these graphs focus on the relationships of haplotypes rather than of individuals. They are usually unrooted, and tips carrying identical haplotypes are collapsed into a single node. The haplotypes of internal nodes are estimated, and these nodes are also collapsed with terminal or other internal nodes wherever identical haplotypes are inferred. As a result, haplotype genealogies can be multifurcating even when sequences have evolved along a bifurcating tree. Haplotype genealogy graphs have been used to visualize processes as diverse as panmixia (Oomen *et al.*, 2011), population expansion (Zane *et al.*, 2006), parallel adaptation (Roesti *et al.*, 2014), sympatric speciation (Barluenga *et al.*, 2006) and introgression (Huerta-Sañchez *et al.*, 2014).

Multiple software tools have been developed to draw haplotype genealogy graphs from sequence data and have in their combination been used in over 10 000 publications. The most popular algorithms include statistical parsimony, implemented in TCS (Clement *et al.*, 2000), median-joining, implemented in the program Network (Bandelt *et al.*, 1999), and minimum-spanning, implemented in Arlequin (Excoffier and Lischer, 2010) and HapStar (Teacher and Griffiths, 2010). These three methods have also been implemented in the recent software PopART, together with a newly developed algorithm, integer neighbour-joining (Leigh and Bryant, 2015).

These algorithms have in common that they allow graphs to include reticulations which are visualized as loops in the genealogy. Due to this property, genealogy graphs produced with these methods are usually referred to as haplotype networks. However, it has been argued that reticulations in haplotype genealogies are often difficult to interpret as they can represent either ambiguous relationships or conflicting topologies due to recombination (Salzburger *et al.*, 2011). As an alternative, Salzburger *et al.* (2011) propose that haplotype genealogies should rather be based on phylogenetic trees without reticulations, which outperformed statistical parsimony in their analyses of simulated data sets. On the other hand, Mardulyn (2012) discusses arguments for and against the use of trees versus networks and concludes that networks appear more appropriate when alternative connections are equally parsimonious.

Here, I present Fitchi, a new program for the production of haplotype genealogy graphs. Following Salzburger *et al.* (2011), Fitchi uses reticulation-free phylogenetic trees and the Fitch algorithm (Fitch, 1970) to reconstruct haplotype genealogies.

2 Methods

Fitchi reads NEXUS format input files including an alignment of haplotypic sequences and a Newick format bifurcating tree of all sequences. While Salzburger *et al.* (2011) found maximum parsimony trees to perform better for highly similar sequence data than trees inferred with maximum likelihood methods, it is left to the user of Fitchi how the tree is obtained. Based on the alignment and the tree, sequences of internal nodes are reconstructed using the Fitch algorithm and Hamming distances (the number of mismatches) between all nodes are calculated as described in Salzburger *et al.* (2011) to produce a so-called Fitch tree (Fig. 1). If terminal sequences contain missing data, the Fitch algorithm is used to infer the sequence with the shortest Hamming distance to its parental node.

Subsequent to sequence reconstruction, branches with a Hamming distances of zero are collapsed, thus creating nodes with more than three connections. As each node is associated with a haplotypic sequence, the number of times this sequence occurs in the alignment is used to calculate the radius of the node in the graph visualization. If population names are provided by the user, nodes are drawn as pie charts indicating how often the haplotype represented by this node is found in each population. To infer the optimal layout of nodes in the graph visualization, Fitchi makes use of the neato algorithm implemented in Graphviz (Gansner and North, 2000). However, since this algorithm does not take node sizes into account, Fitchi only uses edge angles inferred by neato and recalculates edge lengths according to Hamming distances between adjacent nodes.

With longer and more divergent sequence alignments, the number of unique haplotypes can become large, leading to the reconstructed haplotype genealogy graphs being too cluttered for easy interpretation. Fitchi therefore includes several options to simplify haplotype genealogy graphs, highlighting only the most important patterns of population divergence: A minimum node size or Hamming distance can be specified to draw nodes only for haplotypes found with a given minimum frequency, or to collapse nodes separated by short Hamming distances. Alternatively, edge lengths can be calculated only on the basis of inferred transversions, which effectively collapses nodes separated by transitions only. These options are demonstrated with a worked example in Supplementary Fig. S1 in the Supplementary Data.

In addition to the haplotype genealogy graph visualization, Fitchi reports basic alignment statistics as well as measures of population differentiation, including Weir and Cockerham's (1984) F_{st}, d_{XY} and d_f (Cruickshank and Hahn, 2014), and the genealogical sorting index as a measure of per-population monophyly (Cummings *et al.*, 2008).

Fitchi is written as a Python script and allows piping of input and output. Thus, it can easily be integrated into population genomics workflows, where haplotype genealogy graphs and measures of population differentiation are calculated, e.g. in sliding windows across chromosome-length alignments.

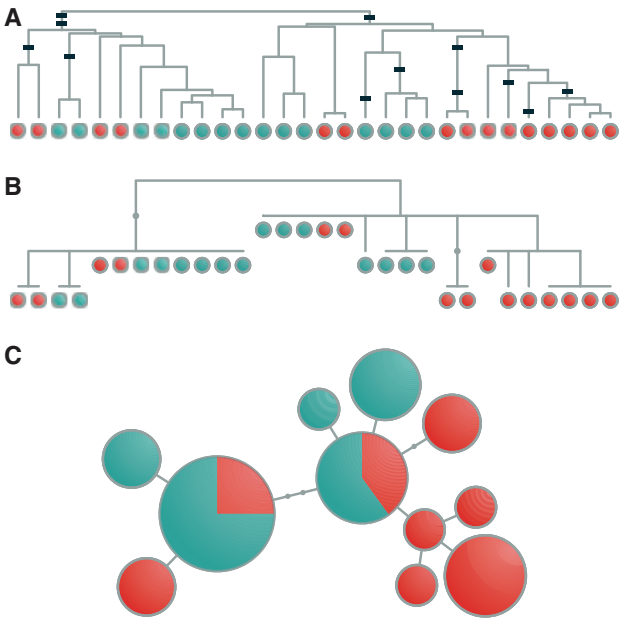

Fig. 1. Relationships between branching history, Fitch tree and haplotype genealogy graph. (**A**) The unobserved true phylogenetic history of sequences sampled from two populations with branch lengths proportional to time and mutations indicated by black bars. (**B**) One of the most parsimonious Fitch trees for the set of sampled sequences, with branch lengths according to Hamming distances. (**C**) Haplotype genealogy graph for the Fitch tree shown in (B), with node sizes proportional to haplotype frequencies. The data set used is provided as Supplementary File S1

3 Results and discussion

Fitchi has already been used for the construction of haplotype genealogy graphs in several publications (Damerau *et al.*, 2014; Roesti *et al.*, 2014, 2015). In Roesti *et al.* (2015), these haplotype genealogies visualize inverted regions on three chromosomes in divergent stickleback populations, where the main patterns of variation were effectively highlighted by showing only nodes with two or more sequences (option '-n 2').

In conclusion, haplotype genealogy graphs produced by Fitchi can be highly useful to visualize genetic variation. However, care must be taken when haplotype connections are ambiguous or may be affected by topological conflicts due to recombination. The robustness of results can be assessed by re-running Fitchi with several optimal parsimony trees, and by specifying different random number seeds (option '-s'). If these should produce different node connections, network approaches might be more suitable to illustrate the genetic variation (Mardulyn, 2012).

Acknowledgements

I thank M. Roesti, A. Runemark, D. Berner and W. Salzburger for test running Fitchi and for feature suggestions and valuable comments on the manuscript. The software and the manuscript also benefited greatly from comments by D. Posada and two reviewers.

Conflict of Interest: none declared.

References

Bandelt,H.J. *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, **16**, 37–48.

Barluenga,M. *et al.* (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, **439**, 719–723.

Clement,M. *et al.* (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1659.

Cruickshank,T.E. and Hahn,M.W. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.*, **23**, 3133–3157.

Cummings,M.P. *et al.* (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411–2422.

Damerau,M. *et al*. (2014) Population divergences despite long pelagic larval stages: lessons from crocodile icefishes (Channichthyidae). *Mol. Ecol*., **23**, 284–299.

Excoffier,L. and Lischer,H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour*., **10**, 564–567.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool*., **19**, 99–113.

Gansner,E.R. and North,S.C. (2000) An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper*., **30**, 1203–1233.

Huerta-Sañchez. *et al*. (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194–197.

Leigh,J.W. and Bryant,D. (2015) popart: full-feature software for haplotype network construction. *Method Ecol. Evol*., **6**, 1110–1116.

Mardulyn,P. (2012) Trees and/or networks to display intraspecific DNA sequence variation? *Mol. Ecol*., **21**, 3385–3390.

Oomen,R.A. *et al*. (2011) Mitochondrial evidence for panmixia despite perceived barriers to gene flow in a widely distributed waterbird. *J. Hered*., **102**, 584–592.

Roesti,M. *et al*. (2014) The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol*., **23**, 3944–3956.

Roesti,M. *et al*. (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun*., **6**, 8767.

Salzburger,W. *et al*. (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Mol. Ecol*., **20**, 1952–1963.

Teacher,A.G.F. and Griffiths,D.J. (2010) HapStar: automated haplotype network layout and visualization. *Mol. Ecol. Resour*., **11**, 151–153.

Weir,B.S. and Cockerham,C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Zane,L. *et al*. (2006) Demographic history and population structure of the Antarctic silverfish *Pleuragramma antarcticum*. *Mol. Ecol*., **15**, 4499–4511.