The Journal of Open Source Software

autoStreamTree: Genomic variant data fitted to geospatial networks

Tyler K. Chafin [1¶], **Steven M. Mussmann** [2,3], **Marlis R. Douglas** [3], and **Michael E. Douglas** [3]

1 Biomathematics and Statistics Scotland, Edinburgh, United Kingdom **2** (current address) Abernathy Fish Technology Center, U.S. Fish & Wildlife Service, Longview, WA, United States of America **3** Department of Biological Sciences, University of Arkansas, Fayetteville, AR, United States of America ¶ Corresponding author

DOI: 10.21105/joss.06160

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @xin-huang
- @abhishektiwari

Submitted: 06 November 2023
Published: 27 March 2024

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Landscape genetics is a statistical framework that parses genetic variation within the context of spatial covariates, but current analytical methods typically fail to accommodate the unique topologies and autocorrelations inherent to network-configured habitats (e.g., streams or rivers). We developed `autoStreamTree` to analyze and visualize genome-wide variation across dendritic networks (i.e., riverscapes). `autoStreamTree` is an open source workflow (https://github.com/tkchafin/autostreamtree) that automatically extracts a minimal graph representation of a geospatial network from a provided shapefile, then 'fits' the components of genetic variation using a least-squares algorithm. To facilitate downstream population genomic analyses, genomic variation can be represented per-locus, per-SNP, or via microhaplotypes (i.e., phased data). We demonstrate the workflow by quantifying genetic variation in an empirical demonstration involving Speckled Dace (*Rhinichthys osculus*).

Statement of need

Network approaches, particularly those graph-theoretic in nature, are increasingly being used to capture functional ecological or evolutionary processes (e.g., dispersal, gene flow) within/among habitat patches (Peterson et al., 2013). In some cases (e.g., riverscapes) topological patterns are explicitly mirrored by the physical habitat, such that the network structure itself places constraints upon processes such as individual movement (Campbell Grant et al., 2007). It is no surprise then, that the importance of network properties such as topological complexity are increasingly implicated as driving evolutionary dynamics in dendritic habitats (Chiu et al., 2020; Thomaz et al., 2016).

Despite this, quantitative frameworks for modelling the relationships between evolutionary and ecological processes (e.g., through spatio-genetic associations) are predominantly focused on landscapes, and as such often involving mechanistic assumptions which translate poorly to networks. We address this limitation by providing a novel package, autoStreamTree, that facilitates network modeling of genome-scale data. It first computes a graph representation from spatial databases, then analyses individual or population-level genetic data to 'fit' distance components at the stream- or reach- level within the spatial network. Doing so within a network context allows the explicit coupling of genetic variation with other network characteristics (e.g., environmental covariates), in turn promoting a downstream statistical process which can be leveraged to understand how those features drive evolutionary processes (e.g., dispersal/gene flow). We demonstrate the utility of this approach with a case study in a small stream-dwelling fish in western North America.

Program Description

Workflow and user interface

autoStreamTree employs the Python networkx library (Hagberg et al., 2008) to parse geospatial input (i.e., large stream networks) into a graph structure with stream segments as edges, sampling locations as endpoints, and river junctions as nodes. Sample data comprise a tab-delimited table of geographic coordinates, genome-wide variant data in VCF format, and (optionally) a tab-delimited population map. The data structure 'graph' on which downstream computations are performed is built as follows: 1) Sample points are 'snapped' to nearest river network nodes (i.e., defining endpoints); 2) Shortest paths are identified between each set of endpoints (Dijkstra, 1959); and 3) A minimal network of original geometries, with contiguous edges derived by joining individual segments with junctions (nodes) retained that fulfill shortest paths.

Pairwise genetic distances from VCF-formatted genotypes (Danecek et al., 2011) are derived among individuals, sites, or populations (via a priori user-specifications). Options for sequence- and frequency-based statistics are provided (`-d/--dist`). Mantel tests are available to quantify correlations among genetic/hydrologic distance matrices. The primary workflow is a least-squares procedure analogous to that used to compute branch lengths within a neighbor-joining phylogenetic tree (Kalinowski et al., 2008). The procedure fits components of the genetic matrix to k-segments in a network, such that fitted distance values (r) for each segment separating two populations will sum to the observed pairwise matrix value. This provides a distance (r_k) for each of k-segments as the genetic distance 'explained' by that segment.

Workflow steps are controlled through the command-line interface (`-r/--run`), with results as plain text tables, and plots via the seaborn package (Waskom, 2021). Fitted distances are added as annotations to an exported geodatabase.

Features

Additional layers of control are provided to minimize pre-processing steps. Users may define individual/site aggregates: 1) Through a tab-delimited classification file; 2) By automatically deriving group membership geographically; or 3) Using an automated DBSCAN clustering method in scikit-learn (Pedregosa et al., 2011).

Users may also provide pre-computed genetic distance matrices either directly at individual or locus levels. Built-in options are provided to concatenate single-nucleotide polymorphisms (SNPs) either globally, or per contig. Individual-level statistics include uncorrected p-distances (i.e., proportion of nucleotide differences), aggregated by site- or at population-level (e.g., as median, arithmetic mean, or adjusted harmonic mean (Rossman, 1990)), or computed as distances via several frequency-based methods (e.g., Chord distance (Cavalli-Sforza & Edwards, 1967); F_{ST} (Weir & Cockerham, 1984)). autoStreamTree can also be computed per-locus by specifying `-r RUNLOCI`, and with `-c LOC` in the case of phased data to treat linked SNPs to microhaplotypes.

Demonstration

Empirical case study

To demonstrate autoStreamTree, we employed existing SNP data for Speckled Dace (*Rhinichthys osculus*)(Mussmann, 2018). Data represent 5,742 SNPs from 762 individuals across 78 localities in the Colorado River (USA), after removing those with >=50% missing data or minor allele frequency (MAF) < 0.1.

Stream networks were parsed directly as a minimal sub-graph from RiverATLAS, which contains various local-scale environmental/hydrological features as annotations (i.e., physiography, climate, land-cover, geology, anthropogenic effects) (Linke et al., 2019). Genetic distances were computed globally and per-locus among sites as linearized F_{ST} (Weir & Cockerham, 1984) (=F_{ST}/1-F_{ST}). To compare with Kalinowski et al. (2008), we used unweighted least-squares, iterative negative distance correction, and replicated analyses using linearized F_{ST} independently recalculated in Adegenet (Jombart, 2008).

We examined variation in per-locus fitted distances as a function of environmental and anthropogenic covariates, carried over as annotations to RiverATLAS. We reduced N=281 hydro-environmental RiverATLAS attributes using forward-selection following the implementation used in adeSpatial (Dray et al., 2018), after first removing variables which were invariant, containing missing values, exhibiting pairwise correlations (r) over 0.7, or having Variance Inflation Factor (VIF) >3. Remaining selected variables were used in redundancy analysis (RDA) to visualize variation in fitted distances as a function of environmental factors. Outliers were detected as those exhibiting z-scores >2.5 in any of the first 3 RDA axes.

Results and comparison

Runtimes are reported for a 2021 Macbook Pro, 16GB memory, 3.2GHz M1 CPU. Time required to calculate a minimal sub-graph containing 118 dissolved edges from RiverATLAS (North America shapefile totaling 986,463 original vertices) was 7m12s. Computing pairwise hydrologic distances required an additional 3s. Pairwise population genetic distances were computed in 16m28s (linearized F_{ST}), with Mantel test and distance fitting taking a total of 17s. Re-running the entire pipeline per-locus for 5,742 SNPs took 8h27m. Fitted-F_{ST} for autoStreamTree (Figure 1) matched that re-calculated using the Kalinowski et al. (2008) method (adjusted Rˆ2 $= 0.9955$; $p < 2.2$e-16). However, due to runtime constraints and manual pre-processing for the latter, per-locus distances were not attempted. The RDA selected 21 environmental variables, with 296 SNPs and 7 edges as outliers (Figure 1), with the dominant environmental driver being lake area (124 SNP outliers).

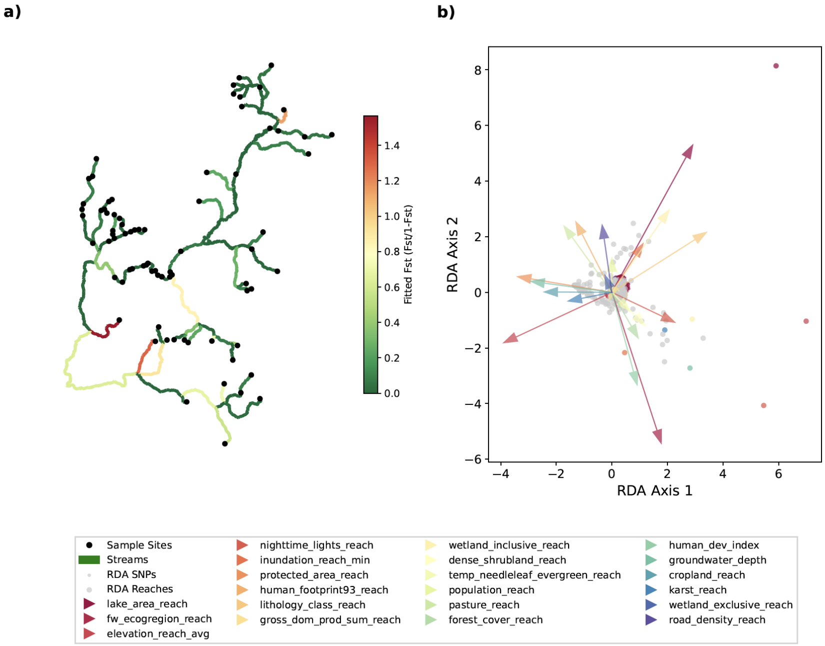

Figure 1: autoStreamTree output. Shown are F_{ST} distances fitted onto original stream network (A), variation in per-locus fitted-F_{ST} distances via pRDA (controlling for stream length) scaled by loci (B), and by stream segment (C). Outliers highlighted according to the most closely correlated environmental axis.

Conclusion

The utility of autoStreamTree was demonstrated with a population genomic dataset as a demonstrative case study. The benefits of the automated approach are underscored by locus-wise microhaplotype versus SNP analysis, which in turn feeds into a quantitative framework that allows 'outlier' loci exhibiting environmental/spatial associations within the autocorrelative structure of the network to be detected. This may potentially imply adaptive variation (although not evaluated herein). In addition, the approach is portable to other data types – indeed, any distance matrix that can be appropriately modeled additively can be supplied, and the process is generalizable to any manner of spatial network.

Acknowledgements

Links to non-Service websites do not imply any official U.S. Fish & Wildlife Service endorsement of the opinions or ideas expressed therein or guarantee the validity of the information provided. The findings, conclusions, and opinions expressed in this article represent those of the authors, and do not necessarily represent the views of the U.S. Fish & Wildlife Service.

This work was supported by University of Arkansas via Doctoral Fellowships (TKC/SMM) and endowments (MED: 21st Chair Century Global Change Biology; MRD: Bruker Life Sciences Professorship). TKC is currently supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). Data and scripts associated with this work can be found on the Open Science Framework (doi: 10.17605/OSF.IO/9BKGR).

References

Campbell Grant, E. H., Lowe, W. H., & Fagan, W. F. (2007). Living in the branches: Population dynamics and ecological processes in dendritic networks. *Ecology Letters*, *10*(2), 165–175. https://doi.org/10.1111/j.1461-0248.2006.01007.x

Cavalli-Sforza, L L, & Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, *19*(3), 550–570. https://doi.org/10.1111/j.1558-5646.1967.tb03411.x

Chiu, M C, Li, B, Nukazawa, K, Resh, V H, Carvajal, T, & Watanabe, K. (2020). Branching networks can have opposing influences on genetic variation in riverine metapopulations. *Diversity and Distributions*, *26*(12), 1813–1824. https://doi.org/10.1111/ddi.13160

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000. G. A. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271. https://doi.org/10.1007/BF01386390

Dray, S., Blanchet, G., Borcard, D., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H. H., & Dray, M. S. (2018). Package "adespatial." In *R package*.

Hagberg, A., Swart, P., & Chult, D. S. (2008). Exploring network structure, dynamics, and function using NetworkX. *Los Alamos National Lab.(LANL), Los Alamos, NM (United States), LA-UR-08-05495; LA-UR-08-5495*.

Jombart, T. (2008). Adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kalinowski, Steven T, Meeuwig, Michael H, Narum, Shawn R, & Taper, M. L. (2008). Stream trees: A statistical method for mapping genetic differences between populations of freshwater organisms to the sections of streams that connect them. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(12), 2752–2760. https://doi.org/10.1139/F08-171

Linke, S, Lehner, B, Ouellet Dallaire, C, Ariwi, J, Grill, G, Anand, M, Beames, P, Burchard-Levine, V, Maxwell, S, Moidu, H, & Tan, F. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, *6*(1), 283. https://doi.org/10.1038/s41597-019-0300-6

Mussmann, S. M. (2018). Diversification across a dynamic landscape: Phylogeography and riverscape genetics of speckled dace (*Rhinichthys osculus*) in Western North America. In *Ph.D. Dissertation*. University of Arkansas.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, & Duchesnay, M. P. E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peterson, Erin E., Ver Hoef, Jay M., Isaak, Dan J., Falke, Jeffrey A., Fortin, Marie-Josée, Jordan, Chris E., McNyset, Kristina, Monestiez, P., Ruesch, Aaron S., Sengupta, Aritra, Som, Nicholas, Theobald, David, Torgerson, E., Christian, & Wnger, S. J. (2013). Modelling dendritic ecological networks in space: An integrated network perspective. *Ecology Letters*, *16*(5), 707–719. https://doi.org/10.1111/ele.12084

Rossman, L. A. (1990). Design stream flows based on harmonic means. *Journal of Hydraulic Engineering*, *116*(7), 946–950. https://doi.org/10.1061/(ASCE)0733-9429(1990)116:7(946)

Thomaz, A T, Christie, M R, & Knowles, L. L. (2016). The architecture of river networks can drive the evolutionary dynamics of aquatic populations. *Evolution*, *70*(3), 731–739. https://doi.org/10.1111/evo.12883

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Weir, B. S., & Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 1358–1370. https://doi.org/10.2307/2408641