




# 1 autoStreamTree: Genomic variant data fitted to 2 geospatial networks

3 Tyler K. Chafin <sup>1¶</sup>, Steven M. Mussmann <sup>2,3</sup>, Marlis R. Douglas <sup>3</sup>, and  
4 Michael E. Douglas <sup>3</sup>

5 <sup>1</sup> Biomathematics and Statistics Scotland, Edinburgh, UK <sup>2</sup> (current address) Abernathy Fish  
6 Technology Center, U.S. Fish & Wildlife Service, Longview, WA, USA <sup>3</sup> Department of Biological  
7 Sciences, University of Arkansas, Fayetteville, AR, USA ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Chris Vernon](#) 

## Reviewers:

- [@xin-huang](#)
- [@abhishektiware](#)

Submitted: 06 November 2023

Published: unpublished

## License

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

## 8 Summary

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

## Statement of need

20 Network approaches, particularly those graph-theoretic in nature, are increasingly being used  
21 to capture functional ecological or evolutionary processes (e.g., dispersal, gene flow) within/  
22 among habitat patches ([Peterson et al., 2013](#)). In some cases (e.g., riverscapes) topological  
23 patterns are explicitly mirrored by the physical habitat, such that the network structure itself  
24 places constraints upon processes such as individual movement ([Campbell Grant et al., 2007](#)).  
25 It is no surprise then, that the importance of network properties such as topological complexity  
26 are increasingly implicated as driving evolutionary dynamics in dendritic habitats ([Chiu et al.,](#)  
27 [2020](#); [Thomaz et al., 2016](#)).

28 Despite this, quantitative frameworks for modelling the relationships between evolutionary  
29 and ecological processes (e.g., through spatio-genetic associations) are predominantly focused  
30 on landscapes, and as such often involving mechanistic assumptions which translate poorly  
31 to networks. We address this limitation by providing a novel package, autoStreamTree, that  
32 facilitates network modeling of genome-scale data. It first computes a graph representation  
33 from spatial databases, then analyses individual or population-level genetic data to ‘fit’ distance  
34 components at the stream- or reach- level within the spatial network. Doing so within a network  
35 context allows the explicit coupling of genetic variation with other network characteristics (e.g.,  
36 environmental covariates), in turn promoting a downstream statistical process which can be  
37 leveraged to understand how those features drive evolutionary processes (e.g., dispersal/ gene  
38 flow). We demonstrate the utility of this approach with a case study in a small stream-dwelling  
39 fish in western North America.  
40

## 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 latitude/longitude 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/scaffold. Individual-level statistics include uncorrected  $p$ -distances (i.e., proportion of nucleotide differences/alignment length), 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 ecosystem, after removing those with  $\geq 50\%$  missing data or minor allele frequency (MAF)  $< 0.1$ .

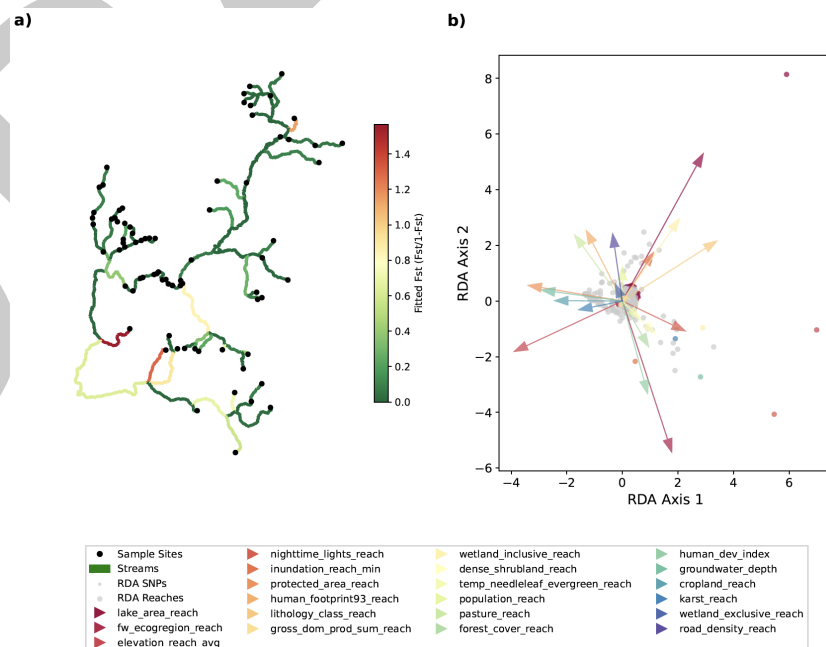
Stream networks were parsed directly as a minimal sub-graph from RiverATLAS, which contains various local-scale environmental/ hydrological features as annotations (i.e., physiography,

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 Adegnet (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/extract 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  $R^2 = 0.9955$ ;  $p < 2.2e-16$ ). 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 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])(<https://doi.org/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>

- 155 Kalinowski, Steven T, Meeuwig, Michael H, Narum, Shawn R, & Taper, M. L. (2008).  
156 Stream trees: A statistical method for mapping genetic differences between populations of  
157 freshwater organisms to the sections of streams that connect them. *Canadian Journal of*  
158 *Fisheries and Aquatic Sciences*, 65(12), 2752–2760. <https://doi.org/10.1139/F08-171>
- 159 Linke, S, Lehner, B, Ouellet Dallaire, C, Ariwi, J, Grill, G, Anand, M, Beames, P, Burchard-  
160 Levine, V, Maxwell, S, Moidu, H, & Tan, F. (2019). Global hydro-environmental sub-basin  
161 and river reach characteristics at high spatial resolution. *Scientific Data*, 6(1), 283.  
162 <https://doi.org/10.1038/s41597-019-0300-6>
- 163 Mussmann, S. M. (2018). Diversification across a dynamic landscape: Phylogeography and  
164 riverscape genetics of speckled dace (\**rhinichthys osculus*\*) in western north america. In  
165 *Ph.D. Dissertation*. University of Arkansas.
- 166 Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand,  
167 Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent,  
168 Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, & Duchesnay,  
169 M. P. E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine*  
170 *Learning Research*, 12, 2825–2830.
- 171 Peterson, Erin E., Ver Hoef, Jay M., Isaak, Dan J., Falke, Jeffrey A., Fortin, Marie-Josée, Jordan,  
172 Chris E., McNyset, Kristina, Monestiez, P., Ruesch, Aaron S., Sengupta, Aritra, Som,  
173 Nicholas, Theobald, David, Torgerson, E., Christian, & Wnger, S. J. (2013). Modelling  
174 dendritic ecological networks in space: An integrated network perspective. *Ecology Letters*,  
175 16(5), 707–719. <https://doi.org/10.1111/ele.12084>
- 176 Rossman, L. A. (1990). Design stream flows based on harmonic means. *Journal of Hydraulic*  
177 *Engineering*, 116(7), 946–950. [https://doi.org/10.1061/\(ASCE\)0733-9429\(1990\)116:](https://doi.org/10.1061/(ASCE)0733-9429(1990)116:7(946))  
178 [7\(946\)](https://doi.org/10.1061/(ASCE)0733-9429(1990)116:7(946))
- 179 Thomaz, A T, Christie, M R, & Knowles, L. L. (2016). The architecture of river networks  
180 can drive the evolutionary dynamics of aquatic populations. *Evolution*, 70(3), 731–739.  
181 <https://doi.org/10.1111/evo.12883>
- 182 Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source*  
183 *Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- 184 Weir, B. S., & Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population  
185 structure. *Evolution*, 1358–1370. <https://doi.org/10.2307/2408641>