Stream trees: a statistical method for mapping genetic differences between populations of freshwater organisms to the sections of streams that connect them

Steven T. Kalinowski, Michael H. Meeuwig, Shawn R. Narum, and Mark L. Taper

Abstract: Statistical approaches for studying the spatial distribution of genetic diversity that assume that organisms move through a two-dimensional landscape are not well suited to study populations of freshwater fish. We present a new statistical method for mapping genetic differences among populations of freshwater fish to the sections of streams that connect them. The method is useful for freshwater species that can only disperse through stream corridors and for other species that live in habitats for which there is one, and only one, corridor connecting each pair of populations (e.g., alpine organisms confined to ridge tops). The model is a simple extension of the least-squares method for constructing evolutionary trees. In this model, the genetic distances between populations are modeled as a sum of genetic distances mapped onto landscape features (e.g., stream sections). Analysis of simulated data shows that the method produces useful results with realistic amounts of data. The model was fit to empirical microsatellite data from four metapopulations of freshwater fish and showed an excellent fit in three out of four cases. Software to perform the necessary calculations is available from the authors at www.montana.edu/kalinowski.

Résumé : Les méthodologies statistiques pour l'étude la répartition spatiale de la diversité génétique qui présupposent que les organismes se déplacent dans un paysage bidimensionnel ne conviennent pas bien à l'étude démographique des poissons d'eau douce. Nous présentons une nouvelle méthode statistique pour cartographier les différences génétiques entre des populations de poissons d'eau douce en relation avec les cours d'eau qui les relient. Cette méthode est utile pour les espèces d'eau douce qui ne peuvent se disperser que par des corridors formés par des cours d'eau et pour les autres espèces pour lesquelles il existe un, et un seul, corridor reliant chaque paire de populations (par ex., les organismes alpins confinés aux sommets des crêtes). Le modèle est une simple extension de la méthode des moindres carrés utilisée pour construire des arbres phylogénétiques. Dans ce modèle, les distances génétiques entre les populations sont représentées comme la somme des distances génétiques cartographiées sur des éléments du paysage (par ex., des sections de cours d'eau). L'analyse de données simulées montre que la méthode fournit des résultats utiles avec un nombre réaliste de données. Nous avons ajusté le modèle à des données empiriques sur les microsatellites dans quatre métapopulations de poissons d'eau douce et obtenu un excellent ajustement dans trois des quatre cas. On peut obtenir des auteurs le logiciel pour réaliser les calculs nécessaires à www.montana.edu/kalinowski.

[Traduit par la Rédaction]

Introduction

The goal of landscape genetics is to understand how geography shapes the genetic composition of populations (e.g., Manel et al. 2003; Storfer et al. 2007). The first step in accomplishing this is usually to describe the spatial distribution of genetic variation. Once genetic relationships among populations have been mapped onto the physical landscape, the influence of landscape features upon genetic structure may be inferred. Increasingly abundant genetic data have recently heightened interest in such research, but investigations of genetic diversity and geography date back many decades (e.g., Wright 1943; Edwards and Cavalli-Sforza 1964).

Traditional methods for describing genetic differences between populations often have disadvantages that limit their usefulness for answering spatial questions. For example, the most commonly used strategy for inferring how geography has influenced population structure is to construct an evolutionary tree showing genetic relationships between populations and then use this tree to identify landscape features that are associated with genetic discontinuities. This approach can

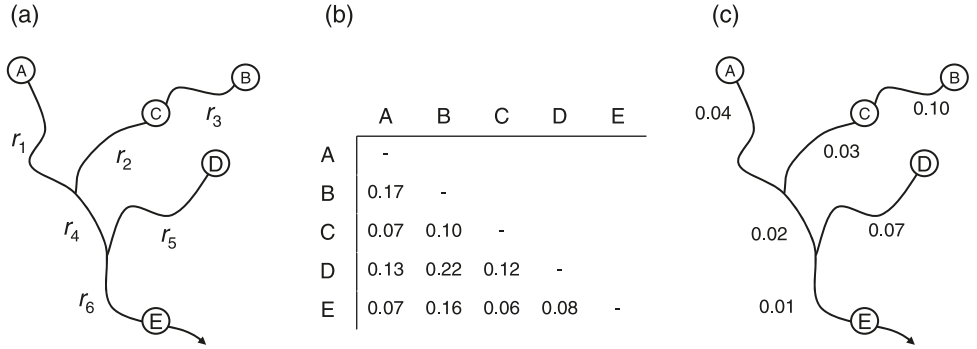
Received 29 November 2007. Accepted 9 September 2008. Published on the NRC Research Press Web site at cjfas.nrc.ca on 6 December 2008.
J20291

S.T. Kalinowski,[1] M.H. Meeuwig, and M.L. Taper. Department of Ecology, Montana State University, Bozeman, MT 59717, USA.
S.R. Narum. Columbia River Inter-Tribal Fish Commission, Hagerman Fish Culture Experiment Station, 3059-F National Fish Hatchery Road, Hagerman, ID 83332, USA.

[1]Corresponding author (e-mail: skalinowski@montana.edu).

 doi:10.1139/F08-171

Fig. 1. (*a*) A map showing the locations of five hypothetical populations of fish (A–E) connected by six stream sections. (*b*) A matrix of genetic distances for the five populations. The goal of the statistical method described in this manuscript is to estimate the parameters r_1 to r_6 so that the sum of the *r*s between each pair of populations is equal (or nearly equal) to the genetic distance observed between them. Such estimates are shown in the "stream tree" in *c*. In this example, there is a perfect fit of the model to the data (i.e., the sum of the genetic distances in the stream tree (*c*) are exactly equal to the genetic distances observed in the matrix (*b*).



be very successful. For example, a tree showing genetic relationships among populations of Atlantic salmon (*Salmo salar*) clearly showed that the largest genetic differences were associated with the Atlantic Ocean (King et al. 2001). Most trees, however, are not so easy to interpret because trees show only genetic similarity; relating the branching pattern of a tree to the physical landscape is often difficult.

The relationship between population genetic structure and geography is often explored by testing for an isolation-by-distance relationship. A significant correlation between genetic difference and geographic distance means that genetic differences are proportional to geographic distances and suggests that the rate of gene flow between populations is proportional to how far they are apart. If genetic differentiation shows a strong isolation-by-distance pattern, no further analysis may be required to explore how geography and genetic structure are related. However, the absence of correlation between genetic differentiation and geographic distance must be interpreted with caution, because there are two very different explanations for how such population structure could evolve. There may be no correlation between genetic structure and geography either because geography has not played an evolutionarily important role in shaping genetic population structure or because other geographic variables besides distance have been important. For example, barriers may prevent gene flow between populations that are close to each other, and corridors may facilitate gene flow between populations that are far apart. If a landscape contains barriers and corridors, populations on the landscape may not show an isolation-by-distance pattern, even though the genetic structure of the populations may have been shaped entirely by the physical landscape.

Many exciting analytic methods have recently been developed to map genetic differences among individuals and populations (e.g., Manni et al. 2004; Miller 2005; Foll and Gaggiotti 2006). The relative merits of these methods and related techniques have not been clearly identified, but most of these methods assume that individuals move through a two-dimensional landscape. For many freshwater organisms, there is only one potential corridor between each pair of populations, and this information should be taken into account when studying how genetic diversity is distributed throughout the landscape. Most of the current statistical approaches for studying landscape genetics cannot do this (the computer program BARRIER, described by Manni et al. (2004), is an exception) and are, therefore, poorly suited for describing genetic diversity for many freshwater fishes.

In this paper, we present a simple statistical method for mapping genetic differences between populations to the sections of streams that connect them. The method is a slight modification of the least-squares approach for making evolutionary trees and therefore has the advantage of using established statistical methods that have proven to be useful. We present this tool as a method for partitioning genetic differences among populations living in a watershed, but the method could be used for any populations for which there is only one possible path between each pair of populations (e.g., alpine populations connected by ridges, coastal organisms living on a shoreline).

A spatial model of genetic differentiation

Let us assume that a matrix of pairwise genetic distances (e.g., F_{ST}) has been estimated for a set of populations connected by streams in a watershed. The goal of our analysis is to map the genetic differences between these populations to the streams that connect them. The model that we propose does this by assigning each section of stream in the watershed a genetic distance that quantifies how much genetic differentiation occurs across that stream section. If, for example, a stream section contains a waterfall that is a barrier to gene flow, we would probably expect that stream section to be assigned a large genetic distance. Let us represent the genetic distance mapped to the *k*th stream section as r_k (Fig. 1). The algorithm that we describe below assigns values of *r* to each stream section in such a manner that if we sum up the *r*s for all of the stream sections between a pair of populations, the sum of these genetic distances will equal the observed genetic distances between the populations. For example, if the pairwise F_{ST} between populations A and B is equal to 0.17, and populations A and B are separated by stream sections 1, 2, and 3, we seek values of r_1, r_2, and r_3 so that $r_1 + r_2 + r_3 = 0.17$ (Fig. 1).

The main assumption in our model is that genetic distances between populations can be modeled as a sum of genetic distances for the stream sections that connect them. This may seem to be an overly simplistic model of genetic differentiation, but it makes the same assumption used to construct additive evolutionary trees (e.g., neighbor joining, unweighted pair group method with arithmetic mean (UPGMA); Felsenstein 2004). Additive evolutionary trees are constructed so that the sum of branch lengths connecting each pair of populations is approximately equal to the observed genetic distance between the populations. The main difference between the stream-based approach that we describe below and an evolutionary tree is that we use the network of streams in the watershed as the topology of the tree, with each stream section in the watershed corresponding to a branch in an evolutionary tree. For this reason, we call our model of genetic differentiation a stream tree.

Stream trees can be constructed from a matrix of genetic distances as follows. Let N represent the number of populations that have been sampled in a watershed, and let D_{ij} represent the genetic distance observed between populations i and j. Let S represent the total number of stream sections in the watershed (see below for a discussion of exactly how to define stream sections). Let r_k represent the genetic distance mapped to the kth stream section (Fig. 1), and let $\boldsymbol{r}$ represent the vector of all these distances, $\boldsymbol{r} = \{r_1, r_2, \ldots, r_S\}$. We seek values of $\boldsymbol{r}$ so that the sums of the genetic distances for stream sections connecting populations are close to the observed genetic distances between the populations. We use the indicator variable $x_{ij,k}$ to calculate these sums. $x_{ij,k}$ equals one if stream section k is between populations i and j and equals zero otherwise. The genetic distance predicted by our model between populations i and j, d_{ij}, is equal to

$$(1) \qquad d_{ij} = r_1 x_{ij,1} + r_2 x_{ij,2} + \ldots + r_S x_{ij,S}$$

We use least-squares estimation to estimate $\boldsymbol{r}$ (Cavalli-Sforza and Edwards 1967; for a review, see Felsenstein 2004, chapter 11). Let Q represent the sum of squared discrepancies between predicted genetic distances and observed genetic distances.

$$(2) \qquad Q = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (D_{ij} - d_{ij})^2$$

Least-squares estimates of $\boldsymbol{r}$ are the values of $\boldsymbol{r}$ that minimize Q. Several iterative and exact methods have been developed to find values of $\boldsymbol{r}$ for the mathematically identical problem of phylogeny estimation. Here we use a slight modification of Cavalli-Sforza and Edwards' (1967) approach. To estimate $\boldsymbol{r}$, we construct a matrix of genetic distances, $\mathbf{D}$, and a matrix of indicator variables, $\mathbf{X}$. The genetic distance matrix $\mathbf{D}$ contains the genetic distances for the N sampled populations arranged in a single column with $N(N-1)/2$ rows. We order the genetic distances so that the first row contains the genetic distance for populations 1 and 2, the second row contains the genetic distance for populations 1 and 3, and so forth. In the example shown in Fig. 1, $\mathbf{D}$ equals

$$(3) \qquad \mathbf{D} = \begin{bmatrix} D_{AB} \\ D_{AC} \\ D_{AD} \\ D_{AE} \\ D_{BC} \\ D_{BD} \\ D_{BE} \\ D_{CD} \\ D_{CE} \\ D_{DE} \end{bmatrix}$$

The matrix $\mathbf{X}$ contains the coefficients $x_{ij,k}$ in a matrix with $N(N-1)/2$ rows and S columns. Each row corresponds to a comparison between two populations (arranged in the same order as $\mathbf{D}$), and each column corresponds to a stream section. In the example shown (Fig. 1), $\mathbf{X}$ equals

$$(4) \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

As above, a one in the matrix indicates that a stream section connects two populations (or is at least part of the connection between populations). Given these matrices, least-squares estimates of $\boldsymbol{r}$ are obtained

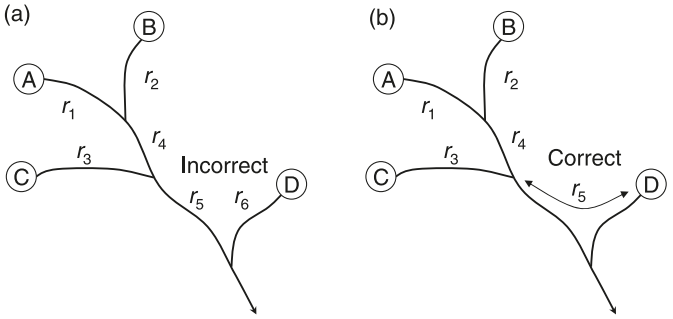
$$(5) \qquad \boldsymbol{r} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}$$

(Cavalli-Sforza and Edwards 1967). This method for estimating the genetic distances of each stream section in our model has been used for a long time to estimate branch lengths in evolutionary trees, e.g., by the computer program FITCH (part of the PHYLIP software package; Felsenstein 2005). The neighbor-joining algorithm for estimating trees (Saitou and Nei 1987) has probably been used more often than least-squares approaches, but both approaches are mathematically related (Felsenstein 2004).

A well-known problem with eq. 5 is that some of the estimated branch lengths can be negative (Felsenstein 1997), and negative genetic distances have no biological meaning. We have used the following iterative procedure when one or more estimates are negative. First, we identify the genetic distance that is most negative and constrain it to equal zero. This is done by removing its column from matrix $\mathbf{X}$. Then we re-estimate the genetic distances for the remaining stream sections. If some of resulting estimates are negative, we constrain the most negative distance to equal zero (in addition to the distance previously constrained to equal zero) and continue in this fashion until all estimates are nonnegative. Algorithms like this are commonly used in constrained optimizations (McCulloch and Searle 2001).

Once estimates of $\boldsymbol{r}$ have been obtained, it is natural to ask if the resulting model fits the data well. This can be

Fig. 2. A hypothetical stream network having four sampled populations (A–D) and stream sections defined either (*a*) incorrectly or (*b*) correctly for the stream network model.



done by calculating the coefficient of determination, R^2, for the stream tree:

$$(6) \qquad R^2 = 1 - \frac{\sum (d_{ij} - D_{ij})^2}{\sum (D_{ij} - \overline{D})^2}$$

where $\overline{D}$ is the average pairwise genetic distance between the populations and summation is taken over all pairs of populations. If R^2 is nearly equal to 1.0, the stream tree provides a good fit to the observed data.

We have modeled the genetic distance between each pair of populations as a linear function of the genetic distances for the stream sections connecting them (eq. 1) and have used least-squares estimation to estimate these stream-specific distances (eq. 5). This approach is mathematically equivalent to linear regression through the origin, which means that any linear regression software package can be used to estimate ***r***. If regression is used, the observed pairwise genetic distance between populations is the dependent variable. The indicator variables $x_{ij,k}$ are the independent variables, and the branch-specific genetic distances ***r*** are the unknown parameters in the regression model. This is a no-intercept model. Appendix A (Table A1) illustrates how the data might be organized to perform this analysis. There are two computational advantages to this approach. First, it can be implemented with ordinary statistical software. Second, some software packages offer methods for constraining regression coefficients to be positive, so the iterative approach described above may not need to be used. To prevent confusion, we note that some software packages calculate R^2 from uncorrected sum of squares for a no-intercept model instead of as we have in eq. 6.

Now that we have described the general structure of our geographic model of population differentiation and explained how the model parameters can be estimated, we will clarify exactly how to define stream sections. Each sampled population must first be labeled on a map. Next, the minimum number of stream sections connecting the populations must be identified. Each of these stream sections must end at either a sampled population or an intersection with *two* other stream sections. Stream sections may not be defined so that they are arranged in tandem fashion in which one section starts at the end of another (unless there is a sampled population at that point). A potentially confusing scenario occurs when a tributary enters a main-stem section downstream of all other populations (population D in Fig. 2). The tributary and the main-stem section immediately upstream of the tributary must be considered as a single stream section. This is because all fish passing through the tributary (e.g., coming from or going to population D in Fig. 2) must also pass through the adjacent section of the main stem if they are to travel to another population. There is no way to differentiate between an obstacle to movement in the tributary or the adjacent section of the main stem. In mathematical terms, if the main-stem section and the tributary were treated independently (e.g., Fig. 2*a*), there would not be a unique solution to eq. 2. Any pair of genetic distances for the tributary and immediately upstream main-stem section (r_5 and r_6 in Fig. 2) that had the same sum would fit the observed data equally well. Therefore, for the model to be applied correctly, the sections must be labeled as a single section (r_5 in Fig. 2*b*).

Materials and methods

Simulations

We used computer simulation to test whether the stream tree algorithm produced reasonable population structures using realistic amounts of data. These tests were not intended to be comprehensive; instead, we attempted to verify that the method worked well in circumstances that it should and that it clearly failed when applied to populations with inappropriate evolutionary histories. In each test, genotypes were simulated for samples from populations in a watershed, and the stream tree algorithm was used to estimate genetic distances for each stream section in the watershed. Three spatial models of gene flow were used in the simulations: a headwater model, a linear stepping stone model, and a semi-realistic model having a complex model of gene flow.

In the headwater model of gene flow (Fig. 3), eight populations were located at the headwaters of a watershed. We assumed that the rate of gene flow between populations was proportional to how closely populations were connected in the watershed. Populations that were connected through the upper part of the watershed (e.g., populations A and B) were given a migration rate of 10^{-1}. Populations that had connections in the middle and lower parts of the watershed were given pairwise migration rates of 10^{-3} and 10^{-5}, respectively. Each population was assumed to have a constant effective population size (N_e) of 1000. The infinite alleles mutation rate used to simulate genotypes was equal to 10^{-4}. Genotypes were simulated for 12 unlinked loci using standard coalescent methods (Hudson 1990). F_{ST} (Weir and Cockerham 1984) was used as a pairwise genetic distance.

Two variations of the headwater model of gene flow were also examined. In the "variable N_e" simulation, half of the populations (B, D, F, and H; Fig. 3) in the watershed were given a N_e of 500, and half the populations (A, C, E, and G) were given an N_e of 1000. Genetic drift is stronger in smaller populations, so in this scenario, the stream sections that connect these smaller populations (sections 2, 4, 6, and 8; Fig. 3) to the rest of the watershed should have a larger genetic distances than the stream sections connecting to the larger populations (sections 1, 3, 5, and 7). A second variation of the headwater model examined the impact of headwater exchange on the ability of a stream tree to describe genetic structure. In this model, genotypes were simulated using migration rates as above. However, before a stream

Fig. 3. The headwater model of gene flow used in simulations.

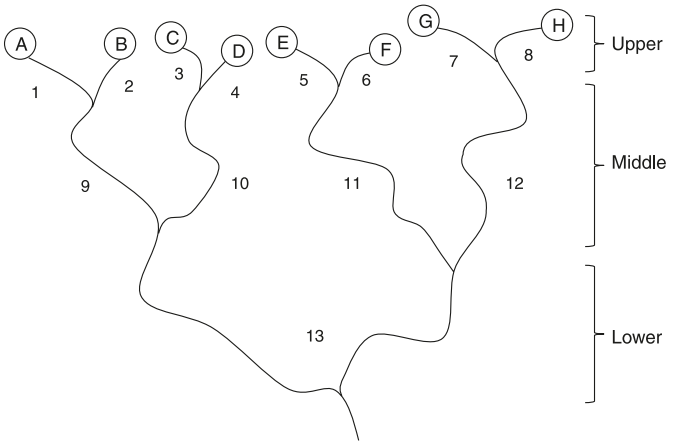
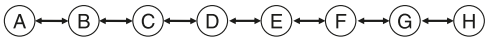


Fig. 4. The linear stepping stone model of gene flow used in simulations. Each population (A–H) was assumed to have the same effective population size and the same rate of gene flow into adjacent populations.

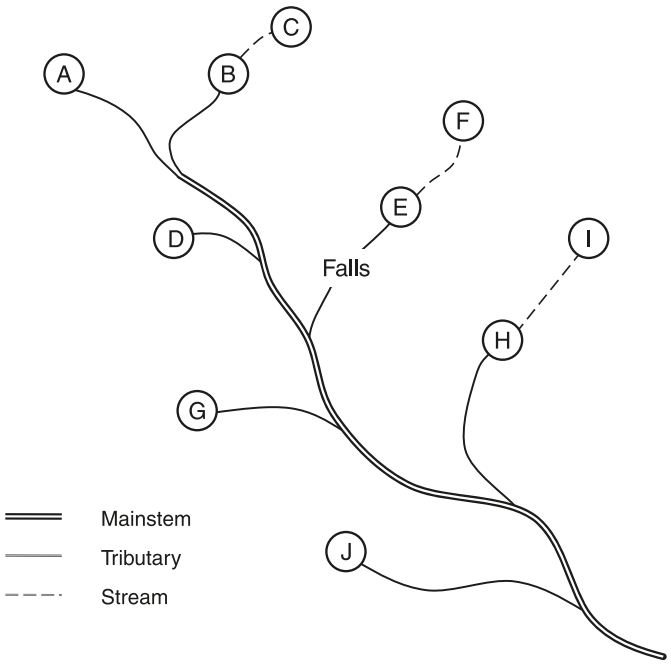


tree was constructed for the watershed, the samples from populations D and E were thoroughly admixed.

In the linear stepping stone model of gene flow (Fig. 4), the rate of gene flow between adjacent populations was assumed to be equal and symmetric. Three different migration rates were examined: m = 0.01, 0.003, and 0.001. N_e for each population was 1000. As for all models, the infinite alleles mutation rate was 10^{-4}. Genotypes were simulated for 12 loci.

The third and last spatial model used to validate the stream tree algorithm was the semi-realistic model. This model assumed that rate of gene flow between populations was determined by the geography of the watershed. The model was inspired by an ongoing study of bull trout (*Salvelinus confluentus*) population genetic structure in Glacier National Park, Montana (M.H. Meeuwig, unpublished data). Four types of waterways were modeled: a main-stem river, tributary rivers, streams within tributaries, and a waterfall (see Fig. 5 for a map depicting the hypothetical watershed). Each of these four types of river sections was assigned a barrier coefficient, z, that quantified how much of a barrier to gene flow the section represented. The migration rate between each pair of populations was then modeled as $10^{-\Sigma z}$, where summation was taken over the barrier coefficients for each stream section between the populations. The barrier coefficient for main-stem river sections was given a value of 0.0, indicating that it was not a barrier to gene flow. Streams were assigned a barrier coefficient of 1.0, tributaries a coefficient of 2.0, and the waterfall a coefficient of 4.0 (indicating that the waterfall reduced the rate of gene flow by a factor of 10^{-4}). For simplicity, migration rates were assumed to be equal upstream and downstream. A few examples illustrate how migration rates were calculated (Fig. 5). The migration rate between populations A and B was $10^{-(1+1)}$ = 0.01. The migration rate between populations D and E was $10^{-(1+0+4)}$ = 0.00001. As above, N_e for each population was assumed to equal 1000, and the infinite alleles mutation rate

Fig. 5. The semi-realistic model of gene flow used in computer simulation. Each type of waterway in the figure (main stem, tributary, stream, and waterfall) was assumed to affect gene flow in a different manner.



was assumed to equal 10^{-4}. Genotypes were simulated for data sets with 12 loci.

Empirical examples

We constructed stream trees of four metapopulations of trout to illustrate the method and to assess whether stream trees fit empirical data well. The four data sets that we analyzed included (*i*) 19 populations of rainbow trout (*Oncorhynchus mykiss*) in the Klickitat River basin, Washington, genotyped at 13 microsatellite loci (Narum et al. 2008), (*ii*) 19 populations of bull trout in Glacier National Park, Montana, genotyped at 10 microsatellite loci (M.H. Meeuwig, unpublished data), (*iii*) 20 populations of bull trout in the Boise River, Idaho, genotyped at six microsatellite loci (Whiteley et al. 2006), and (*iv*) 16 populations of Lahontan cutthroat trout (*Oncorhynchus clarkii henshawi*) in the Marys River drainage, Nevada, genotyped at 11 microsatellite loci (Neville et al. 2006). We used F_{ST} (Weir and Cockerham 1984) as a pairwise genetic distance and used the statistical software package SAS (SAS version 9.1; SAS Institute Inc., Cary, North Carolina) to construct stream trees and calculate R^2 values (eq. 6). We were interested in how stream trees would compare with a traditional bifurcating evolutionary tree, so we made a neighbor-joining tree (Saitou and Nei 1987) for each of these four data sets and calculated a R^2 value for the tree in the same way as we did for stream trees.

Results

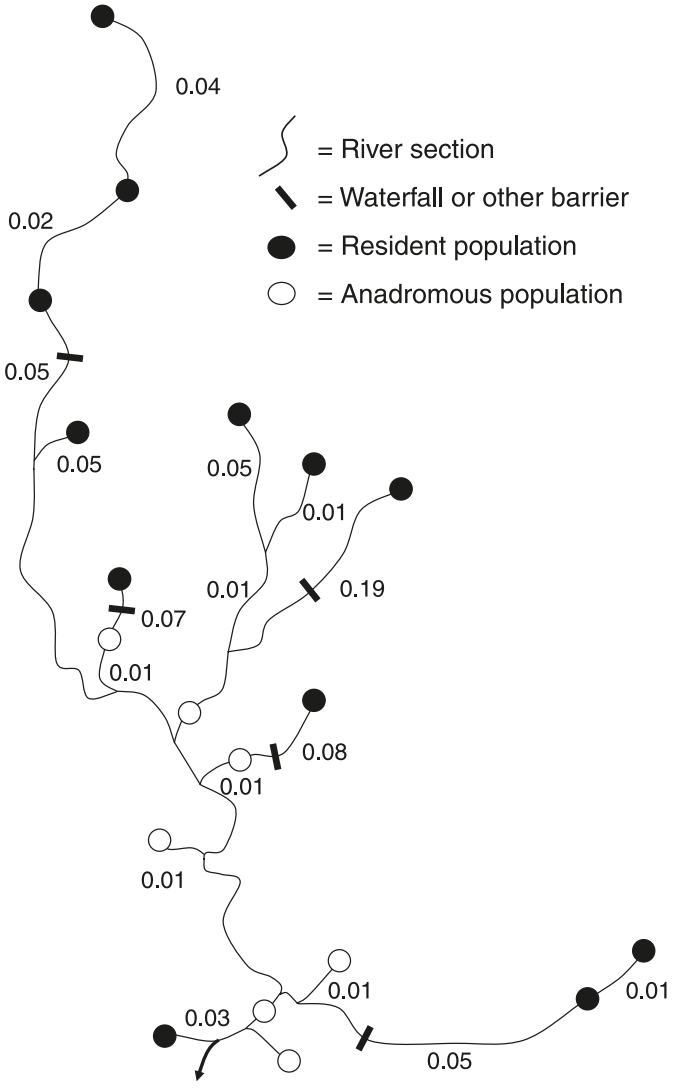
The stream tree algorithm did a very good job of describing genetic structure in the simulated models of gene flow. Results for each of the three models will be presented in turn.

Table 1. The proportion of variation explained by the stream tree model of genetic differentiation described in this paper and neighbor-joining (NJ) trees for four metapopulations of freshwater fish.

					R^2	
Study	Species	N	S	F_{ST}	Stream tree	NJ tree
Narum et al. 2008	Rainbow trout	19	28	0.10	0.974	0.987
M.H. Meeuwig, unpublished data	Bull trout	19	29	0.27	0.971	0.952
Neville et al. 2006	Cutthroat trout	16	22	0.12	0.994	0.934
Whiteley et al. 2006	Bull trout	20	36	0.06	0.681	0.742

Note: N, number of populations sampled; S, number of stream sections in the watershed; F_{ST}, global F_{ST} for all populations sampled. Rainbow trout, *Oncorhynchus mykiss*; bull trout, *Salvelinus confluentus*; cutthroat trout, *Oncorhynchus clarkii henshawi*.

Fig. 6. The stream tree for rainbow trout (*Oncorhynchus mykiss*) populations in the Klickitat River. All sections of river not labeled have a genetic distance of zero.



The average R^2 value for the headwaters model was greater than 0.99, showing that the stream tree did an excellent job of fitting the model to the data. The results also made evolutionary sense. The average genetic distance assigned to the stream sections in the headwaters, which were connected by high rates of gene flow, was 0.0022; the average genetic distance assigned to middle reaches, which had a lower rate of gene flow, was 0.012; and the average genetic distance assigned to the lower reaches, which had the lowest rate of gene flow, was 0.13.

Results were similar for the variable N_e version of the headwaters model. The average R^2 was greater than 0.99, and the average genetic distance for the headwater streams connected to populations with an N_e of 500 (stream sections 2, 4, 6, and 8) was nearly twice as large as the genetic distance mapped to headwater streams connected to populations with an N_e of 1000 (stream sections 1, 3, 5, and 7) (0.0046 vs. 0.0025). This shows that the stream tree algorithm was successfully able to identify the genetic differentiation associated with the stronger genetic drift in the small populations.

As expected, the stream tree algorithm did not do a good job of describing population structure for the headwater exchange model. In this model, populations D and E were genetically identical, despite being far apart on the network of streams. The average R^2 for this model was 0.72 (down from 0.99 for the standard headwater model). This shows that violations of the assumption of the stream tree model of genetic differentiation can be detected by checking R^2 values.

The stream tree algorithm generally did a good job of describing population structure in the linear stepping stone model, but the simulations revealed an important limitation of the stream tree approach. The R^2 value for stream trees constructed for the linear stepping stone model was 0.94 when the migration rate was 0.01 between adjacent populations. This R^2 is lower than values obtained for the headwater model reported above. The decrease is apparently caused by sampling error, for if the number of loci sampled were increased, R^2 would climb towards 1.0 (results not shown). Not only did stream trees constructed from linear stepping stones have relatively high R^2 values, but the genetic distances assigned to the stream sections faithfully depicted the symmetric pattern of gene flow between populations. For the case of a migration rate of 0.01, each stream section was assigned a genetic distance of approximately 0.011. When the migration rate was decreased to 0.003, the fitted genetic distance between populations increased to approximately 0.031, and the R^2 value stayed high (average = 0.93).

Something interesting happened, however, when stream trees were made for the stepping stone model having a lower rate of gene flow. When the migration rate was decreased to 0.001, the average R^2 value dropped to 0.74 (and even lower values were obtained when lower rates of gene

flow were tested). This relatively poor fit was not caused by sampling error but by an inability of the stream tree to accurately depict genetic relationships among populations that are highly differentiated. A close look at the stream tree illustrates the problem. When the rate of gene flow between adjacent populations was 0.001, the average genetic distance between adjacent populations was approximately 0.10. The stream tree algorithm might be expected, therefore, to assign a genetic distance of 0.10 to each stream section in the linear stepping stone model. However, if this was done, the genetic distance between the first and last of the eight populations would be equal to 0.7. This would be a problem, because the maximum value of F_{ST} is approximately equal to the average homozygosity of the loci examined (Kalinowski 2002), and in this case, this was 0.36. Assigning a value of 0.10 to each stream section would, therefore, dramatically overestimate the observed genetic distance between the first and last population (which averaged 0.36, as expected). The best that the stream tree algorithm could do was to assign each section of the river a genetic distance of approximately 0.06. This underestimated the genetic distance between adjacent populations but did not drastically overestimate the genetic distance between distant populations.

This inability of the stream tree algorithm to accurately depict genetic relationships among highly differentiated populations should not be viewed as a critical shortcoming of the algorithm. The source of the problem is that F_{ST} asymptotically approaches a maximum value as populations become highly diverged. Switching to a genetic distance that has a range of 0 to ∞ may alleviate this problem if genetic differentiation is not too extreme. For the case of the stepping stone model with a migration rate of 0.001, switching to Nei's (1978) distance raised the average R^2 to over 0.99. If, however, genetic differentiation is so extreme that populations have diverged to the point that they share no alleles, this will not solve the problem. The important lesson here is that stream trees (and any summary of population structure) should be interpreted with caution when genetic distances approach their maximum value.

The semi-realistic model of gene flow yielded no surprises. In all of the simulated data sets, the genetic distance assigned to the stream section having the waterfall was the highest value in the stream tree (average = 0.33). The tributary sections had an average genetic distance of 0.12, which was appropriate because they had the second lowest rate of gene flow between them. The stream sections having a migration rate of 0.1 had an average genetic distance of 0.0038, and the main-stem sections had an average of 0.0031. This shows that there is some bias in the main-stem estimates (the correct estimate would be 0.0). However, such bias is inevitable because we constrained genetic distances in stream trees to be nonnegative (the median estimate of the genetic distance for main-stem sections was 0.0, as we would wish). The similarity of the estimates for the main-stem sections and the stream sections (0.0031 vs. 0.0038) shows that there is a limit to how accurately subtle amounts of genetic differentiation can be detected. This is not surprising; from a genetic perspective, a migration rate of 0.1 is very high. In these simulations, this corresponded to 100 migrants per generation. Detecting the subtle amount of genetic differentiation that would persist in the presence of such gene flow, and differentiating it from the case of no differentiation, is possible but requires substantially more than 12 loci (results not shown).

Empirical examples

In three out of four empirical tests that we performed, our stream tree model of genetic differentiation had an R^2 value greater than 0.97 (Table 1). This indicates that the genetic differentiation in three of the four species can be successfully mapped to the stream sections connecting the populations. In one metapopulation of bull trout (Whiteley et al. 2006), the stream tree model of population structure had a modest R^2 of 0.68. We do not have a satisfactory explanation for why the stream tree model was not successful for this metapopulation, but Whiteley et al. (2006) did find an unexpected genetic discontinuity in one of the drainages that they could not explain and suggested that there may have been headwater exchange in another drainage.

The stream tree constructed for rainbow trout in Klickitat watershed was consistent with the geography of the basin (Fig. 6). The largest genetic distances on the stream tree correspond to stream section having waterfalls or beaver dams that are thought to prevent upstream fish passage. The smallest genetic distances on the map generally corresponded to sections of the main-stem Klickitat and to stream sections connecting anadromous populations. Note in particular that each section of the lower main-stem Klickitat River was assigned a genetic distance of zero. This suggests that these stream sections are not even a subtle barrier to gene flow among steelhead.

The four empirical examples showed that stream trees usually did as good of a job of summarizing a matrix of genetic distances as neighbor-joining trees (Table 1). Neighbor-joining trees had higher R^2 values for two out of the four cases. The neighbor-joining tree had a poor fit to the data of Whiteley et al. (2006), as did the stream tree. Other than this, no obvious trends were observed.

Software

We have written a computer program, *StreamTree*, to perform the calculations described in this paper. *StreamTree* runs on the Windows operating system and is available for free download at www.montana.edu/kalinowski. Instructions and a sample input file are available at that Web site.

Discussion

We have presented a novel statistical method to map genetic differences among populations to stream sections of a watershed. The method is conceptually and computationally straightforward. It worked well with simulated data and with three out of the four empirical data sets that we analyzed. In the one empirical example where our model did not provide a good fit for the data, neither did a neighbor-joining tree.

The goal of landscape genetics is to understand how geography affects genetic diversity. Such research often has two steps: the first descriptive and the second explanatory. Stream trees are primarily useful for the first step, describing the spatial distribution of genetic diversity. The stream tree algorithm assigns genetic distances to stream sections

but does not try to explain how these differences evolved, or even to correlate them with physical characteristics of the river. The only geographic information used by the algorithm is the topology of the watershed. Because of this, stream trees will be equally useful for mapping genetic diversity to stream sections whether genetic diversity has been shaped by geography or by historical processes unrelated to geography.

The stream tree model of population structure could be modified to include geographical features other than stream sections. For example, if two watersheds were separated by a mountain range, the mountain range could be included as a geographic feature "connecting" the populations on each side of the range (provided that this was done in such a way that only one path connected each pair of populations).

The stream tree model assumes that genetic differences between populations accumulate in an additive manner through a watershed. This is not the same as assuming an isolation-by-distance relationship; the stream tree algorithm can assign short stream sections a large genetic distance and long stream section a small genetic distance. The algorithm, however, does assume that genetic differentiation does not decrease with distance. This means that if populations A, B, and C are arranged in a watershed and that population B is between populations A and C, the stream tree algorithm assumes the genetic distance between A and B, D_{AB}, to be less than or equal to the genetic distance between A and C, D_{AC} (the model also assumes that $D_{BC} \leq D_{AC}$). If this assumption is not true, the stream model will not fit the genetic data well.

There are at least a few evolutionary processes that can violate this assumption. First, as shown in the "headwater exchange" simulation, if gene flow between populations A and C occurs outside of the recognized water channels (i.e., headwater exchange or transplant of fish by fisheries managers), populations A and C may be more similar to each other than to population B. Alternatively, the three populations may have historically been connected by high rates of gene flow, but then became isolated from each other. If population B experiences a population bottleneck that causes rapid genetic drift, and populations A and C retain a large effective population size, populations A and C might be the most genetically similar. Lastly, populations A and C might have a different life history than population B and may have more gene flow between them than to population B. Other evolutionary scenarios that create patterns are possible, but populations having such distributions of genetic diversity are probably less common than populations having spatially additive relationships.

As we have noted above, the stream tree algorithm only provides a map of genetic differences among populations. We anticipate that more geographically explicit models will be constructed in which the genetic distance between populations is explicitly related to stream width, gradient, or other physical variables. This will be more helpful in identifying how landscape characteristics shape genetic variation among populations, which, of course, is the ultimate goal of landscape genetics.

Acknowledgements

We thank H. Neville and A. Whiteley for making their data accessible for this analysis. We thank P. Hedrick, H. Neville, M. Sawaya, P. Spruell, A. Whiteley, and two anonymous reviewers for comments that improved this manuscript. We thank the National Science Foundation and Montana Fish, Wildlife and Parks (contract No. 060327 to M.L.T.) for funding.

References

Cavalli-Sforza, L.L., and Edwards, A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. **19**: 233–257. PMID:6026583.

Edwards, A.W.F., and Cavalli-Sforza, L.L. 1964. Reconstruction of evolutionary trees. *In* Phenetic and phylogenetic classification. *Edited by* V.H. Heywood and J. McNeill. Systematics Association Publ. 6, London, UK. pp. 67–76.

Felsenstein, J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst. Biol. **46**: 101–111. doi:10.2307/2413638. PMID:11975348.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates Inc., Sunderland, Massachusetts.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, Washington.

Foll, M., and Gaggiotti, O. 2006. Identifying the environmental factors that determine the genetic structure of populations. Genetics, **174**: 875–891. doi:10.1534/genetics.106.059451. PMID:16951078.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7**: 1–44.

Kalinowski, S.T. 2002. Evolutionary and statistical properties of genetic distances. Mol. Ecol. **11**: 1263–1273. doi:10.1046/j.1365-294X.2002.01520.x. PMID:12144649.

King, T.L., Kalinowski, S.T., Schill, W.B., Spidle, A.P., and Lubinski, B.A. 2001. Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation. Mol. Ecol. **10**: 807–821. doi:10.1046/j.1365-294X.2001.01231.x. PMID:11348491.

Manel, S., Schwartz, M.K., Luikart, G., and Taberlet, P. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. **18**: 189–197. doi:10.1016/S0169-5347(03)00008-9.

Manni, F., Guerard, E., and Heyer, E. 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum. Biol. **76**: 173–190. doi:10.1353/hub.2004.0034. PMID:15359530.

McCulloch, C.E., and Searle, S.R. 2001. Generalized, linear, and mixed models. John Wiley and Sons, New York.

Miller, M.P. 2005. Alleles in space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. J. Hered. **96**: 722–724. doi:10.1093/jhered/esi119. PMID:16251514.

Narum, S.R., Zendt, J.S., Graves, D., and Sharp, W.R. 2008. Influence of landscape on resident and anadromous life history types of *Oncorhynchus mykiss*. Can. J. Fish. Aquat. Sci. **65**: 1013–1023. doi:10.1139/F08-025.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics, **89**: 583–590. PMID:17248844.

Neville, H.M., Dunham, J.B., and Peacock, M.M. 2006. Landscape attributes and life history variability shape genetic structure of trout populations in a stream network. Landsc. Ecol. **21**: 901–916. doi:10.1007/s10980-005-5221-4.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new

method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406–425. PMID:3447015.

Storfer, A., Murphy, M.A., Evans, J.S., Goldberg, C.S., Robinson, S., Spear, S.F., Dezzani, R., Delmelle, E., Vierling, L., and Waits, L. 2007. Putting the 'landscape' in landscape genetics. Heredity, **98**: 128–142. doi:10.1038/sj.hdy.6800917. PMID:17080024.

Weir, B.S., and Cockerham, C.C. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution, **38**: 1358–1370. doi:10.2307/2408641.

Whiteley, A.R., Spruell, P., Rieman, B.E., and Allendorf, F.W. 2006. Fine-scale genetic structure of bull trout at the southern limit of their distribution. Trans. Am. Fish. Soc. **135**: 1238–1253. doi:10.1577/T05-166.1.

Wright, S. 1943. Isolation by distance. Genetics, **28**: 139–156. PMID:17247075.

Appendix A

Table A1. An example data file that could be used to construct a stream tree using linear regression for the populations depicted in Fig. 1.

COMP	DOBS	SEC_1	SEC_2	SEC_3	SEC_4	SEC_5	SEC_6
AB	0.17	1	1	1	0	0	0
AC	0.07	1	1	0	0	0	0
AD	0.13	1	0	0	1	1	0
AE	0.07	1	0	0	1	0	1
BC	0.10	0	0	1	0	0	0
BD	0.22	0	1	1	1	1	0
BE	0.16	0	1	1	1	0	1
CD	0.12	0	1	0	1	1	0
CE	0.06	0	1	0	1	0	1
DE	0.08	0	0	0	0	1	1

Note: The column at far left, COMP, lists the populations being compared. This is useful to keep the data organized but is not used in the regression. The next column, DOBS, is the observed genetic distance for each pair of populations. The columns SEC_1 to SEC_6 contain 0s and 1s that indicate whether each section of stream (1 to 6) is between the populations listed in the COMP column. When conducting the regression, DOBS is the dependent variable and SEC_1 through SEC_6 are the independent variables. There is no intercept. The output from the regression will be a set of six coefficients (0.04, 0.03, 0.10, 0.02, 0.05, 0.01). These are the genetic distances to assign to each section of stream in the watershed.