


ORIGINAL RESEARCH

Machine-learning-based detection of adaptive divergence of the stream mayfly *Ephemera strigata* populations

Bin Li^{1,2}  | Sakiko Yaegashi^{2,3} | Thaddeus M. Carvajal² | Maribet Gamboa² | Ming-Chih Chiu² | Zongming Ren¹ | Kozo Watanabe²

¹Institute of Environmental and Ecology, Shandong Normal University, Jinan, China

²Department of Civil and Environmental Engineering, Ehime University, Matsuyama, Japan

³Department of Civil and Environmental Engineering, University of Yamanashi, Yamanashi, Japan

Correspondence

Kozo Watanabe, Department of Civil and Environmental Engineering, Ehime University, Bunkyo-cho 3, Matsuyama, 790-8577, Japan.
Email: watanabe.kozo.mj@ehime-u.ac.jp

Funding information

Japan Society for the Promotion of Science, Grant/Award Number: 16H04437, 16K18174 and 17H01666

Abstract

Adaptive divergence is a key mechanism shaping the genetic variation of natural populations. A central question linking ecology with evolutionary biology is how spatial environmental heterogeneity can lead to adaptive divergence among local populations within a species. In this study, using a genome scan approach to detect candidate loci under selection, we examined adaptive divergence of the stream mayfly *Ephemera strigata* in the Natori River Basin in northeastern Japan. We applied a new machine-learning method (i.e., random forest) besides traditional distance-based redundancy analysis (dbRDA) to examine relationships between environmental factors and adaptive divergence at non-neutral loci. Spatial autocorrelation analysis based on neutral loci was employed to examine the dispersal ability of this species. We conclude the following: (a) *E. strigata* show altitudinal adaptive divergence among the populations in the Natori River Basin; (b) random forest showed higher resolution for detecting adaptive divergence than traditional statistical analysis; and (c) separating all markers into neutral and non-neutral loci could provide full insight into parameters such as genetic diversity, local adaptation, and dispersal ability.

KEYWORDS

adaptive divergence, altitude, aquatic insect, local adaptation, random forest, STRUCTURE

1 | INTRODUCTION

A central question linking ecology with evolutionary biology is how spatial environmental heterogeneity can lead to adaptive divergence among local populations within a species. In stream ecosystems, adaptive divergence of aquatic insects is usually reported to be influenced by altitudinal gradient at the river corridor scale (Hughes, Schmidt, & Finn, 2009; Keller, Alexander, Holderegger, & Edwards, 2013; Polato et al., 2017). The underlying mechanism is that altitude is often strongly related to a number of environmental

factors such as temperature and oxygen availability which greatly influenced the life of organisms (Lytle & Poff, 2004; Halbritter, Billeter, Edwards, & Alexander, 2015; Keller & Seehausen, 2012). Thermal regimes directly regulate the growth of species, development, and mating behavior, and setting limits on distributions and abundances of species across landscapes (Li et al., 2013). Oxygen availability also restricts distributions by affecting respiratory metabolism of aquatic organisms (Rostgaard & Jacobsen, 2005).

Recently, there has been an increase in studies on the genetic basis of adaptive divergence in aquatic insects because of their

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

important role in freshwater ecosystem biomonitoring. Altitudinal genetic divergence has been reported in aquatic insects including caddis flies: *Plectrocnemia conspersa* and *Polycentropus flavomaculatus* (Wilcock, Bruford, Nichols, & Hildrew, 2007); *Stenopsyche maramorata* (Yaegashi, Watanabe, Monaghan, & Omura, 2014); stone flies: *Dinocras cephalotes* (Elbrecht et al., 2014); and mayflies: *Atalophlebia* (Baggiano, Schmidt, Sheldon, & Hughes, 2011). However, most of these studies were based on a given gene or a limited number of candidate genes.

The development of genome scanning approaches, such as amplified fragment length polymorphism (AFLP), allows the study of numerous anonymous markers (loci) rather than the study of a few candidate genes. Compared with neutral loci, loci influenced by directional selection (i.e., non-neutral loci) are expected to exhibit higher levels of genetic divergence (Kirk & Freeland, 2011). Therefore, by screening large numbers of candidate loci ("outlier" loci, reviewed by Nosil, Funk, & Ortiz-Barrientos, 2009), statistical methods can identify loci that are under direct selection or linked to loci under selection based on the level of genetic divergence. Selected non-neutral loci can be used to test hypotheses about the adaptive process. Also, neutral loci may be available for accurate tests of neutral processes, such as isolation by distance (IBD) (Oleksa, Chybicki, Gawroński, Svensson, & Burczyk, 2013) and gene flow patterns, avoiding the confounding effects of natural selection (Kirk & Freeland, 2011).

In the ordinary analysis of genome scanning, non-neutral loci are detected based on genetic variation among populations with different phenotypes or ecotypes (Bonin, Taberlet, Miaud, & Pompanon, 2006; Egan, Nosil, & Funk, 2008; Galindo & Rolán-Alvarez, 2009; Nosil, Egan, & Funk, 2008) or allopatric populations among different geographic localities (Gaggiotti et al., 2009; Medugorac et al., 2009; Renaut, Nolte, Rogers, Derome, & Bernatchez, 2011). Genome scanning can also be conducted using genetically defined populations with unknown phenotypes or ecotypes. For example, Bayesian clustering methods (Falush, Stephens, & Pritchard, 2003, 2007; Pritchard, Stephens, & Donnelly, 2000) can delineate genetic populations prior to any observable phenotypic divergence and, therefore, may provide insights into the early stages of adaptive divergence (Whiteley et al., 2011).

Determining the link between non-neutral loci and environmental factors is one of the most difficult tasks in molecular ecology. Conventional statistical methods such as the partial Mantel test (Legendre & Fortin, 2010; Watanabe, Kazama, Omura, & Monaghan, 2014), distance-based redundancy analysis (dbRDA) (Watanabe & Monaghan, 2017), and multivariate analysis of variance (MANOVA) (Mccairns & Bernatchez, 2008) have been widely applied. However, these methods pose certain issues and limitations. One issue is the tendency of bias and high error rates that result from associating genetic variance and environmental distances (Guillot & Rousset, 2013; Legendre & Fortin, 2010; Legendre, Fortin, & Borcard, 2015). In addition, the Mantel test and dbRDA are limited to testing the linear independence

between genetic and environmental distances among local populations. This may due to the nonlinearity of these distances and possible information loss in the converting process. Additionally, there is often much difficulty in fulfilling underlying assumptions (e.g., normal distribution and homogeneity of variance) of conventional statistical methods such as MANOVA or multiple linear regression (Vittinghoff, Glidden, & McCulloch, 2012). Because of these concerns, modern statistical techniques, such as machine-learning methods, are now being developed as promising alternatives. Machine-learning methods are particularly effective in finding and describing structural patterns in data and providing the values of relative importance among variables (Biau & Scornet, 2016; Prasad, Iverson, & Liaw, 2006).

Among the variety of machine-learning methods available, random forest (RF) (Breiman, 2001) is one of the most widely used modeling techniques to generate high prediction accuracy and evaluate the relative importance of explanatory variables in the model (Biau & Scornet, 2016). RF is an ensemble tree-based method that constructs multiple decision trees from a data set and combines results from all the trees to create a final predictive model. In ecological studies, RF has been applied to community-level studies to predict the distributions of species and identify constrained environmental factors (Evans, Murphy, Holden, & Cushman, 2011; Pelletier, Carstens, Tank, Sullivan, & Espíndola, 2018; Smith & Carstens, 2019; Wedger, Topp, & Olsen, 2019). In most of these studies, environmental data have been used as independent variables to predict the presence or absence of species (dependent variables). The relative contributions of environmental variables to the distribution of species are quantified by their relative importance obtained from the RF model. It may therefore be possible to extend the use of RF to population genetic studies where environmental variables are used to predict the presence or absence of haplotypes or alleles at outlier loci. The relative importance of each environmental variable could be considered as its influence to outlier loci, which may strongly drive adaptive divergence.

In this study, we examined adaptive divergence using AFLP markers in populations of the stream mayfly *Ephemera strigata* from the Natori River Basin in northeastern Honshu Island, Japan. We have two main objectives: The first is to determine the extent of local adaptation at the genome level in natural populations and to quantify associations between environmental gradients and adaptive divergence, and the second objective is to apply a modified machine-learning method for determining the selection pressure on outlier loci. We first detected loci under selection (non-neutral loci) based on locus-specific genetic differentiation among populations. Rather than defining populations a priori using geographic or phenotypic information, we delineated populations based on the discontinuities in the AFLP variation among individuals using a hierarchical analysis of STRUCTURE (Falush, Stephens, & Pritchard, 2003, 2007; Pritchard et al., 2000; Vähä, Erkinaro, Niemelä, & Primmer, 2007). Secondly, focusing on non-neutral loci, we employed a machine-learning method (i.e., RF) to identify environmental variables

most likely to contribute to adaptive divergence. Additionally, we also conducted the ordinary distance-based redundancy analysis (dbRDA) to comparatively examine the feasibility of the method. Finally, focusing on the neutral loci, we examined dispersal pattern and dispersal distance of the species.

2 | METHODS

2.1 | Study site and sampling

Ephemera strigata is a well-studied mountain burrowing mayfly in Japan and Korea (Ban & Kawai, 1986; Lee, Hwang, & Bae, 2008). In this study, sampling was carried out in the Natori River catchment in the Miyagi Prefecture in northeastern Japan (Figure 1). Nymphal samples were collected at 11 sites from October 26 to November 12, 2010. At each site, we collected *E. strigata* individuals using a Surber net (30 × 30 cm quadrat with mesh size 250 μm) along 200–900 m stream reaches. All specimens were preserved in the field in 99.5% ethanol, transported to the laboratory, and identified to species level under a stereomicroscope (120×) using taxonomic keys (Kawai & Tanida, 2005).

We measured seven geographic parameters at each site using standard ecological methods in stream surveys (Hauer & Lamberti, 2007; Watanabe, Monaghan, & Omura, 2008). Stream order was determined using a 1:25,000 map. The width of the stream channel was measured as average value at 10 randomly selected cross sections using a tape measure. Longitude and latitude coordinates and altitude were recorded using a global positioning system on the riverside. Distance to river mouth was the distance between sampling site and river mouth, and riverine distance was the river course distance between each pair of two sites. Both parameters were measured on Google Maps using the ruler function. Because there is no correlation among selected variables based on collinearity analysis (variance inflation factor [VIF] < 10), we included all those variables in our further analysis.

2.2 | DNA extraction and AFLP fingerprinting

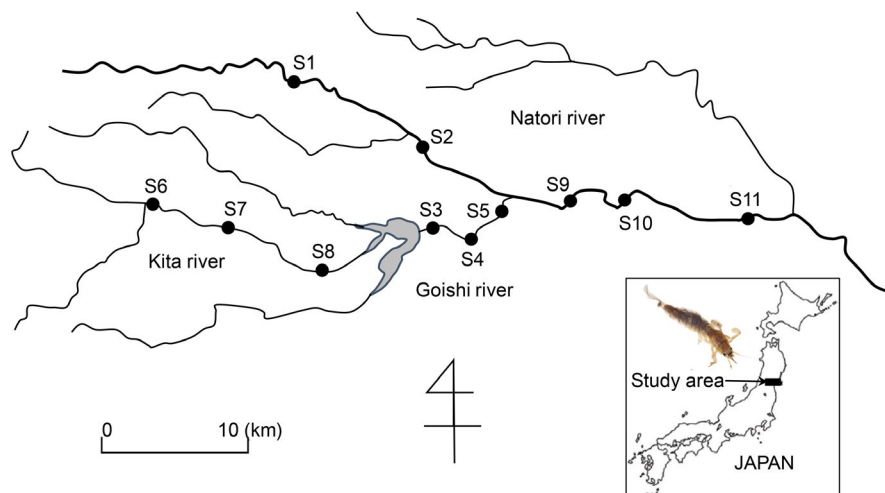
DNA extraction was performed on abdominal tissues after digestive tract removal. The DNA of each individual was extracted using the DNeasy 96 Blood & Tissue Kits (Qiagen). The concentration of extracted DNA was measured on a NanoDrop ND 1000 spectrometer (Thermo Fisher Scientific) and diluted to 50 ng/μl.

We genotyped 216 individuals from 11 sites with the AFLP method (Vos et al., 1995). The restriction step followed the protocol by Watanabe et al. (2014). The ligation step was performed by adding 1 U T4 DNA ligase (New England), 0.2 μl of 100 μM MseI adapter, 0.2 μM of EcoRI adapter, 2 μl T4 DNA ligase buffer (10×) (New England), and up to 20 μl dH₂O and incubating the solution at 16°C for 12 hr. The sequences of the MseI adapter and EcoRI adapters were extracted from Reisch (2007). The adapters were manually prepared as follows: (a) mixing equal molar amounts of adapter oligomer, (b) denaturing at 95°C for 5 min, and (c) incubating for 10 min at room temperature. Restricted or ligated products were then diluted at a 1:19 ratio with 0.1 × TE buffer. Preselective amplification was performed in a mixture of 0.06 μl of 100 μM MseI and EcoRI primers (Reisch, 2007), 15 μl of AFLP Amplification Core Mix (Applied Biosystems), 4 μl of each restricted/ligated product, and up to 29 μl dH₂O. Preselective polymerase chain reaction (PCR) parameters followed Reisch (2007). PCR products were diluted 20 times by 0.1 × TE buffer.

For selective amplifications, we employed three types of primer pairs (EcoRI-AGG & MseI-CAT, EcoRI-ACC & MseI-CAC, and EcoRI-AGG & MseI-CAC) that generate the most variable patterns in 64 types of selective primer pairs using three individuals. Each EcoRI primer was modified with Beckman Dye 2, 3, or 4 on the 5'-end. The mixture of selective PCR was 0.1 μl of 100 μM MseI and EcoRI primers, 15 μl of AFLP Amplification Core Mix (Applied Biosystems), and up to 20 μl dH₂O. We followed Reisch (2007) to set PCR parameters.

The selective PCR products were separated by capillary gel electrophoresis using CEQ8000 (Beckman Coulter). We extract all fragments which were ranged from 60 bp to 360 bp, using the

FIGURE 1 Map of 11 sampling sites and photograph of species *Ephemera strigata* in the Natori River Basin in northeastern Japan



default parameters. To adjust fluorescent intensity, each fluorescent PCR product was mixed with the following: EcoRI-AGG & MseI-CAT 4 μ l, EcoRI-ACC & MseI-CAC 2 μ l, and EcoRI-AGG & MseI-CAC 1 μ l. Peak sizes of PCR products were calibrated with DNA Size Standard 600 (Beckman Coulter) and calculated using the CEQ8000 software (Beckman Coulter) as per the instruction of manufacturer.

2.3 | Hierarchical STRUCTURE analysis

We defined populations based on discontinuities in AFLP variation using the individual-based Bayesian clustering method implemented in STRUCTURE v. 2.3 (Falush et al., 2003, 2007; Pritchard et al., 2000). We performed 20 runs of 50,000 iterations with a burn-in of 10,000 for each number of assumed populations (K) ranging from 1 to 15 using the admixture model and assuming correlated allele frequencies. We used a uniform prior for alpha (the parameter representing the degree of admixture) with a maximum of 10 and set AlphaPrpsd to 0.05. Lambda, the parameter representing the correlation in the parental allele frequencies, was estimated in a preliminary run using $K = 1$. The prior F_{ST} was set to the default value (mean = 0.01; standard deviation (SD) = 0.05).

To determine the optimal K , we computed the log-likelihood ($\ln P(K)$) for each K and selected K with the highest standardized second-order rate of change (ΔK) of $\ln P(K)$ (Evanno, Regnaut, & Goudet, 2005). Although this method helps to correctly identify K in most situations, it is known to have two limitations. First, it is useful only for the uppermost level of a hierarchical genetic structure. Second, it is unable to find the best K if $K = 1$ (i.e., if there is no population substructure) (Evanno et al., 2005). To address these limitations, we used a hierarchical approach for STRUCTURE analysis modified from Vähä et al. (2007), which repeats the analysis at lower hierarchical levels until no substructure can be uncovered. The advantage of our method was that we used the Wilcoxon two-sample test to control the round of repeated analysis instead of checking the pattern of individual membership. Specifically, we compared the mean value of $\ln P(K)$ from 20 runs with the optimal K (as determined using ΔK) with mean $\ln P(K = 1)$ using the Wilcoxon two-sample test (Rosenberg et al., 2001). If $\ln P(K = 1)$ was found to be significantly lower than $\ln P(K)$ at the optimal K , we repeated the analysis within each of the K populations. At each hierarchical level, individuals were assigned to subpopulations based on the individual membership coefficient (Pritchard et al., 2000).

2.4 | Outlier loci detection

We used two different statistical methods to identify outlier loci. Dfdist (adapted from Fdist; Beaumont & Nichols, 1996) uses coalescent simulations to generate thousands of loci evolving under a neutral model of symmetrical islands with a mean global F_{ST} close to the observed global F_{ST} . Mean F_{ST} was calculated using the default method by first excluding 30% of the highest and lowest observed

values. Empirical loci with F_{ST} values significantly greater ($p < .05$) than the simulated distribution (generated with 50,000 loci) were considered to be outliers. Dfdist can detect both divergent selection and balancing selection; however, we focused only on divergent selection in this study. BayeScan is a hierarchical Bayesian model-based method first described in Beaumont and Balding (2004) and modified by Foll and Gaggiotti (2008) for dominant markers (available at <http://cmpg.unibe.ch/software/bayescan/>). This Bayesian method is based on the concept that F_{ST} values reflect contributions from locus-specific effects, such as selection, and population-specific effects, such as local effective size and immigration rates. The main advantage of this approach is that it allows for different demographic scenarios and different amounts of genetic drift in each population (Foll & Gaggiotti, 2006, 2008). Using a reversible jump Markov chain Monte Carlo approach, the posterior probability of each locus being subjected to selection is estimated. A locus is deemed to be influenced by selection if its F_{ST} is significantly higher or lower than the expectation provided by the coalescent simulations. For all subsequent analyses, non-neutral loci were defined as outlier loci detected by the Dfdist and BayeScan methods at the 95% confidence level. Neutral loci were defined as loci detected by neither Dfdist nor BayeScan at the 95% thresholds. Loci detected as outliers by only one of the two methods were not considered in the further analyses. In order to check whether there have some loci being misidentified as outlier due to linkage disequilibrium, we further tested for pairwise linkage disequilibrium (LD) of the outlier loci detected by both methods, using 1,000 steps in the Markov chain and a dememorization of 1,000 steps in ARLEQUIN 3.5 (Excoffier & Lischer, 2010).

2.5 | Analysis of genetic diversity

F_{ST} was calculated with ARLEQUIN v. 3.5 using (a) all loci, (b) only neutral loci, and (c) only non-neutral loci. Global heterozygosity among all populations (H_t) and mean heterozygosity within populations (H_w) were estimated separately for neutral and non-neutral loci with AFLP-SURV v. 1.0 (Vekemans, Beauwens, Lemaire, & Roldán-Ruiz, 2002) using the Bayesian method with a uniform prior distribution of allele frequencies (Zhiotovskiy, 1999). Molecular variance analysis (AMOVA) was also conducted using ARLEQUIN to provide the estimates of genetic variations among and within sampling sites. For the test of IBD, we examined the correlations of pairwise F_{ST} with geographic distance and riverine distance (i.e., distance along the watercourse) between sites using GeneAIEx v. 6.5 (Peakall & Smouse, 2012). The genetic distance between each pair of sites was quantified using mean pairwise F_{ST} for neutral and non-neutral loci using the Bayesian-estimated allele frequencies generated by AFLP-SURV.

We conducted genetic spatial autocorrelation analysis using neutral loci for geographic distance. Eight geographic distance classes defined every 4 km (from 0–4 km to 28–32 km) were used in the analysis. Individuals within the same site were considered to be

separated by a distance of 0 km. We calculated Moran's I for each distance class using GeneAIEx, where I ranges from -1 to 1 and the positive values indicate that sites within a given distance class have similar genetic structure. We used jackknifing to estimate the 95% confidence intervals.

2.6 | Adaptive divergence modeling

We determined the environmental variables that drive adaptive divergence at non-neutral loci using the RF model (Blagus & Lusa, 2013; Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Maciejewski & Stefanowski, 2011). Stream order, width of the stream, longitude, latitude, altitude, and distance to river mouth were used to predict the band presence/absence patterns at each non-neutral locus for each individual. We assigned individuals from the same site to the same environmental conditions. The data set was imbalanced because the number of individuals with band presence was not equal to that with band absence. The individuals were thus classified into two classes (i.e., presence and absence). We solved the data imbalance problem by oversampling for the minority class through the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) using SMOTE function in DMwR package (Torgo, 2013). SMOTE creates synthetic minority class sample units by taking the difference between the feature vector (sample) under consideration and its nearest neighbor. It then multiplies this difference by a random number between 0 and 1 and adds it to the feature vector under consideration (Chawla et al., 2002). The RF model for each non-neutral locus was built using randomForest function in randomForest package (Liaw & Wiener, 2002) in the R program (R Development Core Team, 2015). Model performance was evaluated using the area under the receiver operating characteristic curve (AUC) (Janitza, Strobl, & Boulesteix, 2013). The AUC value typically ranged from 0.5 (random prediction) to a maximum value of 1, which represents the theoretical perfect model. As rules of thumb, an AUC value greater than 0.9 indicates very good model quality, a value smaller than 0.7 indicates poor model quality, and a value between 0.7 and 0.9 indicates good model quality (Baldwin, 2009).

We also conducted dbRDA as a comparative ordinary method. Among the seven environmental variables, we searched for the variables that best explain the most variation in F_{ST} at non-neutral loci. DbRDA was performed on the ordination solutions, rather than on the distance matrices (Legendre & Fortin, 2010). In this study, pairwise genetic distances at non-neutral loci among sites were used to screen environmental factors that most closely relate to genetic divergence (Watanabe & Monaghan, 2017). The best model, comprising significant predictors, was selected using forward selection with permutation tests and an inclusion threshold of $\alpha = 0.05$ using the ordistep function of the vegan package (Oksanen et al., 2018) in the R program (R Development Core Team, 2015). Significant differences were tested with the ANOVA.cca function in the vegan package.

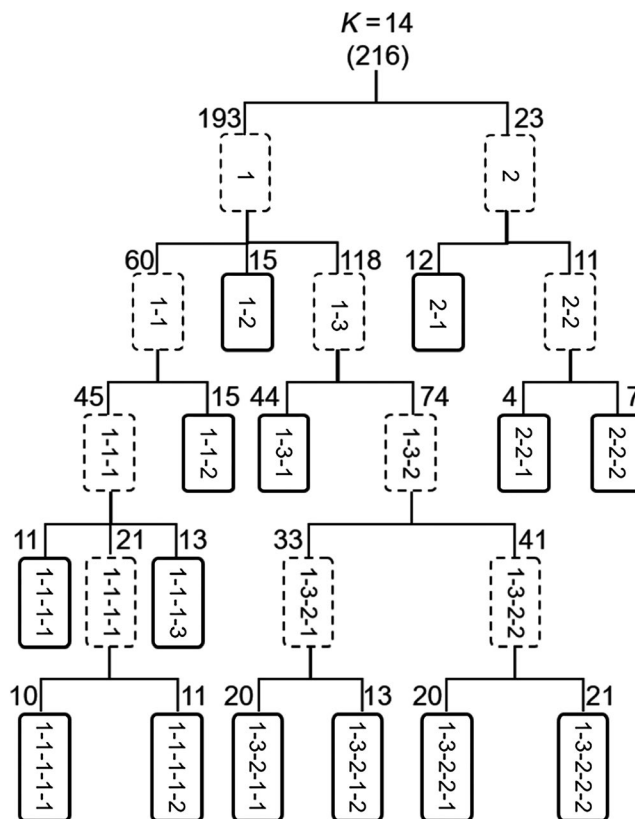


FIGURE 2 Subpopulation structure of *Ephemera strigata* as determined using STRUCTURE with hierarchical iterations. Dashed boxes indicate subpopulations, and solid boxes indicate final populations. Numbers at the top of boxes indicate the number of individuals assigned to the populations. A total of 14 groups (K) were defined from 216 individuals

3 | RESULTS

3.1 | Hierarchical STRUCTURE analysis

Hierarchical iterations by STRUCTURE detected significant substructure until the 4th iteration beyond the initial analysis (Figure 2). A total of 14 groups were defined for the 216 *E. strigata* individuals collected in 11 sites. The numbers of individuals assigned in each group were ranged from 4 to 44 (mean = 15.4; $SD = 9.5$). Most groups were widespread all over the sampling sites, whereas some groups were restricted to specific sites (data not shown). For example, the members of groups 2, 3, and 8 occurred only in upstream and middle-stream sites (Figure 1: upstream sites, S1 and S6-8; middle-stream sites, S2-5).

3.2 | Outlier detection and genetic diversity

Using our criterion of 95% significance with both Dfdist and BayeScan, 10 non-neutral loci and 346 neutral loci were detected from the 372 polymorphic AFLP loci. Dfdist alone detected 10 outlier loci under divergent selection and 11 outlier loci under balancing

selection. Outlier loci under balancing selection were not investigated in this study. All 10 outlier loci under divergent selection were consistently identified by BayeScan, which alone identified 26 outliers. LD analysis found that 8%–15% of possible pairwise combinations of outlier loci were statistically linked (Randomization test, $p < .01$). The proportions of significant pairs were higher than expected by chance; however, there was no locus pair that was consistently in disequilibrium in multiple populations. Total genetic variation (H_t) was lower at neutral loci than at non-neutral loci and the same trend occurred in mean genetic variation within sites (H_w ; Table 2). Mean global F_{ST} among all sites for all AFLP loci was 0.029 ($p < .01$; AMOVA). When measured using neutral or non-neutral loci, we found global F_{ST} values of 0.021 ($p < .01$) and 0.039 ($p < .01$), respectively (Table 2).

3.3 | Detection of adaptive divergence

We separately built one RF model for each of the 10 non-neutral loci (Table 1). Of the 10 non-neutral loci, loci 56, 89, and 254 were well-predicted (i.e., $AUC > 0.7$) with altitude being the most important environmental variable (Figure 3), suggesting that the genetic divergence of these loci was mainly driven by altitude. Based on dbRDA, only genetic divergence at locus 254 was significantly predicted ($p < .05$) (Figure 4). Altitude explained 54% of the genetic divergence at this locus. However, for the other non-neutral loci, no significant relationship with environmental factors was found with dbRDA ($p > .05$).

IBD was not significant for either geographic ($r = .11$, $p = .33$) or riverine distance ($r = .06$, $p = .49$) (Figure 5). The results of the spatial autocorrelation analysis based on neutral loci showed significant positive autocorrelation coefficients at the shortest range of 0–4 km (Figure 6).

TABLE 1 Sample size, AUC, OOB error rates, and key factors defined by random forest for each non-neutral locus (sample size was shown with abundant category/rare category to show data imbalance)

Locus	Sample size (n = 216)	AUC	OOB	Key factor
56	202/14	0.85	5.12%	Altitude
254	175/41	0.79	12.54%	Altitude
89	199/17	0.74	11.48%	Altitude
247	204/12	0.67	7.72%	River width
36	182/34	0.52	13.43%	Stream order
90	152/64	0.51	35.22%	Latitude
98	174/42	0.51	22.78%	Latitude
97	130/86	0.51	33.09%	Distance to river mouth
260	185/31	0.50	15.41%	River width
289	200/16	0.50	12.98%	River width

4 | DISCUSSION

In this study, we newly employed a modified RF model to examine the relationship between environmental factors and adaptive divergence at non-neutral loci. An oversampling process was added using SMOTE in DMwR package in R to balance the data set before RF model building. Ordinary statistical tests of multiple linear regression method require assumptions that data are normally distributed with homogeneity of variance and independent from one another

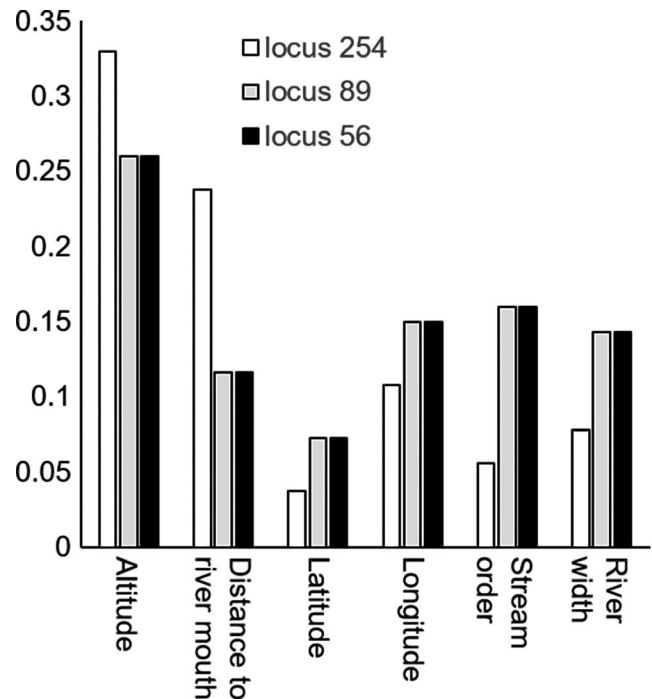


FIGURE 3 Relative importance of environmental variables based on the random forest model for three non-neutral loci (56, 89, and 254)

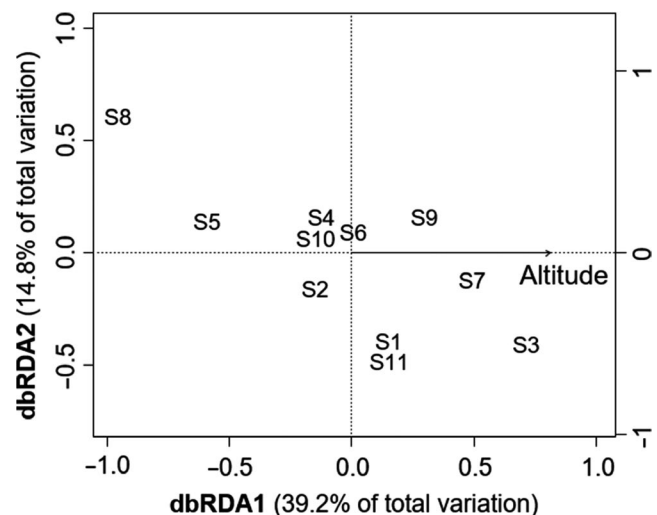


FIGURE 4 Distance-based redundancy analysis (dbRDA) describing the influence of environmental heterogeneity on genetic variation at a non-neutral locus (254)

FIGURE 5 Isolation by distance calculated using geographic (a) and riverine (b) distance. Solid lines indicate correlations between Wright's fixation index (F_{ST}) and geographic ($r = .11$, $p = .33$) or riverine distance ($r = .06$, $p = .49$) calculated with the Mantel tests

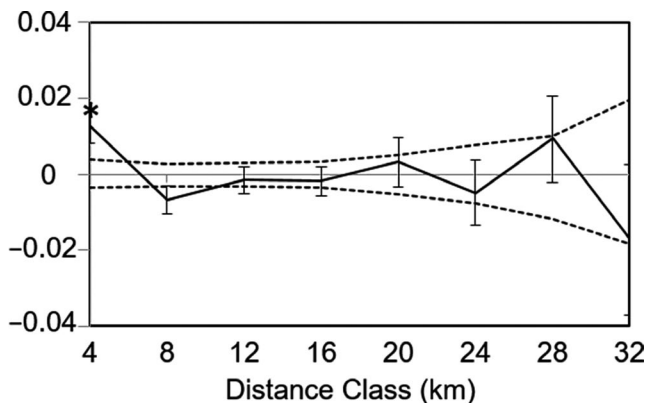
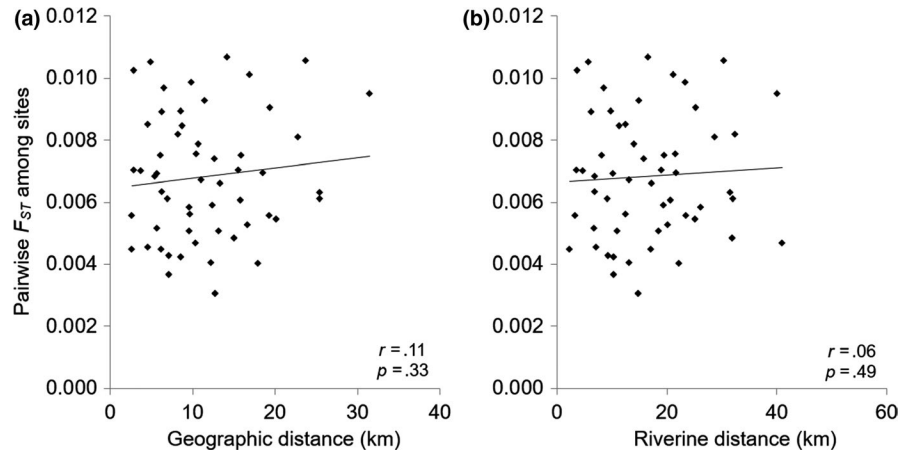


FIGURE 6 Spatial autocorrelation at 4-km distance classes based on geographic distance for neutral loci. Dashed lines indicate permuted 95% confidence intervals, and error bars indicate jackknifed 95% confidence intervals. * indicates significant spatial autocorrelation ($p < .05$)

TABLE 2 Genetic diversity and divergence measured using the following: (1) all loci, (2) only neutral loci, and (3) only non-neutral loci

	H_t	H_w	F_{ST}
All loci	0.1358	0.1357	0.029
Neutral loci	0.1173	0.1155	0.021
Non-neutral loci	0.4379	0.3523	0.039

Note: H_t = total expected heterozygosity; H_w = mean expected heterozygosity within sites; and F_{ST} = Wright's fixation index among sites.

(Vittinghoff Glidden, & McCulloch, 2012), which are often difficult to fulfill. The environmental factors investigated in this study did not show strong independency among variables (data not shown). However, the utilization of RF could overcome such difficulty, accommodating pronounced nonlinearities in the exploration of gene–environment relationships in large genomic data sets (Biau & Scornet, 2016; Breiman, 2001; Fitzpatrick & Keller, 2015).

We developed 10 RF models for each of the 10 non-neutral loci detected by both BayeScan and Dfdist. As a result, 3 out of the 10 non-neutral loci (loci 56, 89, and 254) showed good model prediction

performance ($AUC > 0.7$), whereas the other 7 non-neutral loci could not be well-modeled. The reason why we could not build a good model for the 7 non-neutral loci is probably because the natural selection on these loci was driven by the other environmental factors which were reported in other studies but not included in our analysis (e.g., velocity, chl-a) (Brouwer, Bessee-Lototskaya, ter Braak, Kraak, & Verdonschot, 2017; Li et al., 2016; Watanabe et al., 2014). Because RF can effectively perform well with a large number of variables (Genuer, Poggi, & Tuleau-Malot, 2010), it is recommended in future studies to include as many environmental variables as possible to gain a deeper insight into the role of these factors to adaptive divergence.

To compare the performance of RF with ordinary statistical analysis, we also conducted dbRDA on each of the 10 non-neutral loci. One locus (locus 254) was well-modeled by dbRDA. This locus was one of the 3 loci modeled by RF, and the selected environmental factor (i.e., altitude) was consistent with RF. The low number of loci modeled in dbRDA may be because of its limited usage for testing only linear independency and its low independency. The ranking of variable importance in RF stems from the idea that if the variable is not important, then rearranging its values should not affect the prediction accuracy of the model (Breiman, 2001). This algorithm could reduce the influence of variable dependency as compared with dbRDA (Archer & Kimes, 2008; Genuer et al., 2010).

In this study, we used populations delineated by a hierarchical STRUCTURE analysis for the identification of non-neutral loci as an alternative to geographic or phenotypic populations which are usually used in the ordinal analysis of genome scanning. The STRUCTURE analysis successfully delineated populations with significant difference in genetic terms, which is difficult to detect using visible characters such as phenotypes, ecotypes, or geographic localities (Pritchard et al., 2000). The STRUCTURE analysis can delineate genetic populations among individuals prior to any observable phenotypic divergence, and hence, may provide a means to look at early stages of adaptive divergence prior to any phenotypic divergence in the population delineation and detection of non-neutral loci (Whiteley et al., 2011).

The hierarchical approach which we newly introduced to the STRUCTURE analysis enabled us to study the finer population

structure (i.e., higher K) than ordinal STRUCTURE analysis, which stops the analysis once the uppermost hierarchical level is found. The number of populations (K) is an important determinant in the outlier detection (Foll & Gaggiotti, 2008). We also conducted outlier loci detection based on the geographic populations and the uppermost hierarchical level of STRUCTURE analysis that delineated only two populations; however, we could not detect any outlier loci. Nevertheless, fine hierarchical level (e.g., the 4th iteration in our hierarchical STRUCTURE analysis) will define weak population structure based on very subtle differences, which may introduce the risk of overfitting.

By employing a genome scan approach in this study, we comparatively used neutral and non-neutral loci in examining genetic diversity and genetic distance. We found an interesting pattern of greater genetic divergence at non-neutral loci than that at neutral loci. This pattern is consistent with three caddis flies species and one mayfly species studied in the same catchment system (Watanabe et al., 2014). Moreover, there are several other supporting studies which compared levels of genetic divergence between morphological traits as analogous to non-neutral markers and neutral DNA markers in other macroinvertebrate species such as snails (Cook, 1992), spiders (Gillespie & Oxford, 1998), and damselflies (Wong, Smith, & Forbes, 2003). Based on the results of D_{fdist} , all the 10 non-neutral loci were under divergent selection rather than stabilizing selection; therefore, they presented greater genetic divergence than neutral loci (Table 2).

One of the most interesting findings of this study is that mountain burrowing mayfly *E. strigata* present adaptive divergence along an altitude gradient. Altitude is often reported to be closely related to a number of environmental factors that greatly influence the life cycle or development of organisms (Lytle & Poff, 2004; Múrria, Bonada, Arnedo, Prat, & Vogler, 2013; Halbritter et al., 2015). For example, altitude influences the phenology of insects, restricting the mating period to only a few days, leading to asynchronous emergence that may act as a reproductive barrier between populations (e.g., Watanabe & Monaghan, 2017; Yaegashi et al., 2014) or as a regulation of their metabolism (Gamboa, Tsuchiya, Matsumoto, Iwata, & Watanabe, 2017). This variable also influences air density, and in addition to its significance for respiration, this implies more energy is required for flight. The hemoglobin gene and other genes with a potential role for adaptation to low O_2 may show divergence between the populations along an altitude gradient (Keller et al., 2013).

In principle, unlike non-neutral markers, neutral markers are suitable for examining neutral process occurring under the drift-migration balance. Former population genetic studies that inferred dispersal pattern of stream insects usually used all DNA markers without classification of neutral and non-neutral loci (Mila, Carranza, Guillaume, & Clobert, 2010; Miller, Blinn, & Keim, 2002). This may potentially cause overestimation of genetic drift because non-neutral loci under divergent selection will increase the genetic divergence which had not occurred from genetic drift (Kirk & Freeland, 2011). Therefore, we used only neutral markers in inferring the dispersal pattern.

In the result of IBD analysis based on neutral loci, we did not observe any significant IBD for either geographic or riverine distances, suggesting that populations are not in a genetic drift-migration equilibrium at the geographic scale (Figure 5). The results of spatial autocorrelation analysis based on neutral loci showed significant positive autocorrelation coefficients at the shortest distance range (i.e., 0–4 km, Figure 6a), indicating low-dispersal ability in this species. Such observation is understandable because mayflies are generally considered as having low-dispersal ability in mountain streams (Barber-James, Gattolliat, Sartori, & Hubbard, 2007). The limited dispersal distances were also observed in stoneflies due to their poor dispersal abilities (Briers, Cariss, & Gee, 2003; Briers, Gee, Cariss, & Geoghegan, 2004). In contrast, caddis flies were frequently reported to show strong dispersal ability. Yaegashi et al. (2014) reported species *Stenopsyche marmorata* exhibited dispersal ability along stream corridors up to 12 km.

In conclusion, the modified RF approach applied in this study provides an alternative method in determining constraint environmental factors for outlier loci under selection. We found that the mountain burrowing mayfly *E. strigata* present adaptive divergence along an altitude gradient using neutral and non-neutral methods. The hierarchical STRUCTURE analysis could help to detect finer populations and increase the power of outlier detection. One limitation in this study is that we did not include many environmental factors that may also have the chance to be constrained factors and help to improve the model performance. Assessing a larger number of non-neutral loci or do some simulations with known constraint variable use our modified RF approach will make it more applicable. Alternatively, sequencing the detected outlier loci would provide a deeper understanding of elevational adaptation of this species. In addition, besides the research in Natori River system, a comparative study in other similar area could help to provide a more comprehensive understanding of genetic adaptive divergence of *E. strigata*.

ACKNOWLEDGMENTS

This research was financially supported by the Japan Society for the Promotion of Science (JSPS) (grant numbers: 16H04437, 17H01666, and 16K18174). We thank K. Nagamine, S. Takahashi, and Y. Kumagai for assistance with field sampling and laboratory works and T. Omura for useful suggestions. H. Harada, Tohoku University, allowed us to use their DNA sequencer and analyzing system.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Bin Li: Formal analysis (lead); methodology (lead); writing–review and editing (lead). **Sakiko Yaegashi:** Data curation (lead); investigation (lead). **Thaddeus M. Carvajal:** Conceptualization (equal); methodology (equal). **Maribet Gamboa:** Conceptualization (supporting); writing–original draft (supporting); writing–review and editing (supporting). **Ming-Chih Chiu:** Conceptualization (supporting); methodology (supporting); writing–review and editing (supporting).

Zongming Ren: Conceptualization (supporting); writing–review and editing (equal). **Kozo Watanabe:** Conceptualization (supporting); funding acquisition (lead); project administration (lead); writing–review and editing (supporting).

DATA AVAILABILITY STATEMENT

Data supporting the results of the paper uploaded on Dryad: <https://doi.org/10.5061/dryad.hmgqnk9d0>. The reviewer URL is available: https://datadryad.org/stash/share/LTdzZZxLDvEzI93auRjM_U_OYijDBpCuaUCypjZnJ60

ORCID

Bin Li  <https://orcid.org/0000-0001-5455-6969>

REFERENCES

- Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*, 2249–2260. <https://doi.org/10.1016/j.csda.2007.08.015>
- Baggiano, O., Schmidt, D. J., Sheldon, F., & Hughes, J. M. (2011). The role of altitude and associated habitat stability in determining patterns of population genetic structure in two species of *Atalophlebia* (Ephemeroptera: Leptophlebiidae). *Freshwater Biology*, *56*, 230–249. <https://doi.org/10.1111/j.1365-2427.2010.02490.x>
- Baldwin, R. A. (2009). Use of maximum entropy modeling in wildlife research. *Entropy*, *11*, 854–866. <https://doi.org/10.3390/e11040854>
- Ban, R., & Kawai, T. (1986). Comparison of the life cycles of two mayfly species between upper and lower parts of the same stream. *Aquatic Insects*, *8*, 207–215. <https://doi.org/10.1080/01650428609361255>
- Barber-James, H. M., Gattolliat, J.-L., Sartori, M., & Hubbard, M. D. (2007). Global diversity of mayflies (Ephemeroptera, Insecta) in freshwater. *Hydrobiologia*, *595*, 339–350. <https://doi.org/10.1007/s10750-007-9028-y>
- Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*, 969–980. <https://doi.org/10.1111/j.1365-294X.2004.02125.x>
- Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *263*, 1619–1626.
- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*, 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
- Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced Data. *BMC Bioinformatics*, *14*, 106. <https://doi.org/10.1186/1471-2105-14-106>
- Bonin, A., Taberlet, P., Miaud, C., & Pompanon, F. (2006). Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the commonfrog (*Rana temporaria*). *Molecular Biology and Evolution*, *23*, 773–783. <https://doi.org/10.1093/molbev/msj087>
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
- Briers, R., Cariss, H., & Gee, J. H. R. (2003). Flight activity of adult stoneflies in relation to weather. *Ecological Entomology*, *28*, 31–40. <https://doi.org/10.1046/j.1365-2311.2003.00480.x>
- Briers, R. A., Gee, H. R., Cariss, H. M., & Geoghegan, R. (2004). Interpopulation dispersal by adult stoneflies detected by stable isotope enrichment. *Freshwater Biology*, *49*, 425–431. <https://doi.org/10.1111/j.1365-2427.2004.01198.x>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. <https://doi.org/10.1613/jair.953>
- Cook, L. (1992). The neutral assumption and maintenance of color morph frequency in mangrove snails. *Heredity*, *69*, 184–189.
- deBrouwer, J. H. F., Bessee-Lototskaya, A. A., ter Braak, C. J. F., Kraak, M. H. S., & Verdonchot, P. F. M. (2017). Flow velocity tolerance of lowland stream caddisfly larvae (Trichoptera). *Aquatic Sciences*, *79*, 419–425. <https://doi.org/10.1007/s00027-016-0507-y>
- Egan, S. P., Nosil, P., & Funk, D. J. (2008). Selection and genomic differentiation during ecological speciation: Isolating the contributions of host association via a comparative genome scan of *neoclamisus bebiana* leaf beetles. *Evolution*, *62*, 1162–1181.
- Elbrecht, V., Feld, C. K., Gies, M., Hering, D., Sondermann, M., Tollrian, R., & Leese, F. (2014). Genetic diversity and dispersal potential of the stonefly *Dinocras cephalotes* in a central European low mountain range. *Freshwater Science*, *33*, 181–192.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*, 2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
- Evans, J. S., Murphy, M. A., Holden, Z., & Cushman, S. A. (2011). Modeling species distribution and change using random forest. In C. Drew, Y. Wiersma, & F. Huetmann (Eds.), *Predictive species and habitat modeling in landscape ecology* (pp. 139–159). New York, NY: Springer.
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analysis under Linux and Windows. *Molecular Ecology Resources*, *10*, 564–567.
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.
- Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, *7*, 574–578. <https://doi.org/10.1111/j.1471-8286.2007.01758.x>
- Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*, 1–6. <https://doi.org/10.1111/ele.12376>
- Foll, M., & Gaggiotti, O. E. (2006). Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, *174*, 875–891. <https://doi.org/10.1534/genetics.106.059451>
- Foll, M., & Gaggiotti, O. E. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, *180*, 977–993. <https://doi.org/10.1534/genetics.108.092221>
- Gaggiotti, O. E., Bekkevold, D., Jørgensen, H. B. H., Foll, M., Carvalho, G. R., André, C., & Ruzzante, D. E. (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, *63*, 2939–2951. <https://doi.org/10.1111/j.1558-5646.2009.00779.x>
- Galindo, J., & Rolán-Alvarez, E. (2009). Comparing geographical genetic differentiation between candidate and noncandidate loci for adaptation strengthens support for parallel ecological divergence in the marine snail *Littorina saxatilis*. *Molecular Ecology*, *18*, 919–930.
- Gamboa, M., Tsuchiya, M. C., Matsumoto, S., Iwata, H., & Watanabe, K. (2017). Differences in protein expression among five species of stream stonefly (Plecoptera) along a latitudinal gradient in Japan. *Insect Biochemistry and Physiology*, *96*, e21422. <https://doi.org/10.1002/arch.21422>
- Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*, 2225–2236. <https://doi.org/10.1016/j.patrec.2010.03.014>
- Gillespie, R. G., & Oxford, G. S. (1998). Selection on the color polymorphism in hawallan happy-face spiders: Evidence from genetic structure and temporal fluctuations. *Evolution*, *52*, 775–783.

- Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4, 336–344. <https://doi.org/10.1111/2041-210x.12018>
- Halbritter, A. H., Billeter, R., Edwards, P. J., & Alexander, J. M. (2015). Local adaptation at range edges: Comparing elevation and latitudinal gradients. *Journal of Evolutionary Biology*, 28, 1849–1860. <https://doi.org/10.1111/jeb.12701>
- Hauer, F. R., & Lamberti, G. A. (2007). *Methods in stream ecology* (3rd ed.). London, UK: Academic Press.
- Hughes, J. M., Schmidt, D. J., & Finn, D. (2009). Genes in streams: Using DNA to understand the movement of freshwater fauna and their riverine habitat. *BioScience*, 59, 573–583. <https://doi.org/10.1525/bio.2009.59.7.8>
- Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14, 119. <https://doi.org/10.1186/1471-2105-14-119>
- Kawai, T., & Tanida, K. (2005). *Aquatic insects of Japan: Manuals with keys and illustration (in Japanese)*. Tokyo, Japan: Tokai University Press.
- Keller, I., Alexander, J. M., Holderegger, R., & Edwards, P. J. (2013). Widespread phenotypic and genetic divergence along altitudinal gradients in animals. *Journal of Evolutionary Biology*, 26, 2527–2543. <https://doi.org/10.1111/jeb.12255>
- Keller, I., & Seehausen, O. (2012). Thermal adaptation and ecological speciation. *Molecular Ecology*, 21, 782–799. <https://doi.org/10.1111/j.1365-294X.2011.05397.x>
- Kirk, H., & Freeland, J. R. (2011). Applications and Implications of Neutral versus Non-neutral Markers in Molecular Ecology. *International Journal of Molecular Sciences*, 12, 3966–3988. <https://doi.org/10.3390/ijms12063966>
- Lee, S. J., Hwang, J. M., & Bae, Y. J. (2008). Life history of a lowland burrowing mayfly, *Ephemera orientalis* (Ephemeroptera: Ephemeridae), in a Korean stream. *Hydrobiologia*, 596, 279–288. <https://doi.org/10.1007/s10750-007-9103-4>
- Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, 10, 831–844. <https://doi.org/10.1111/j.1755-0998.2010.02866.x>
- Legendre, P., Fortin, M. J., & Borcard, D. (2015). Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, 6, 1239–1247. <https://doi.org/10.1111/2041-210X.12425>
- Li, B., Watanabe, K., Kim, D.-H., Lee, S.-B., Heo, M., Kim, H.-S., & Chon, T.-S. (2016). Identification of outlier loci responding to anthropogenic and natural selection pressure in stream insects based on a self-organizing map. *Water*, 8, 188. <https://doi.org/10.3390/w8050188>
- Li, F. Q., Chung, N., Bae, M.-J., Kwon, Y.-S., Kwon, T.-S., & Park, Y.-S. (2013). Temperature change and macroinvertebrate biodiversity: Assessments of organism vulnerability and potential distributions. *Climatic Change*, 119, 421–434. <https://doi.org/10.1007/s10584-013-0720-9>
- Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest*. R News 2:18–22. Retrieved from <http://CRAN.R-project.org/doc/Rnews/>
- Lytle, D. A., & Poff, N. L. (2004). Adaptation to natural flow regimes. *TRENDS in Ecology and Evolution*, 19(2), 94–100. <https://doi.org/10.1016/j.tree.2003.10.002>
- Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. *Computational Intelligence and Data Mining*, 104–111.
- Mccairns, R. J. S., & Bernatchez, L. (2008). Landscape genetic analyses reveal cryptic population structure and putative selection gradients in a large-scale estuarine environment. *Molecular Ecology*, 17, 3901–3916. <https://doi.org/10.1111/j.1365-294X.2008.03884.x>
- Medugorac, I., Medugorac, A., Russ, I., Veit-Kensch, C. E., Taberlet, P., Luntz, B., ... Förster, M. (2009). Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Molecular Ecology*, 18, 3394–3410. <https://doi.org/10.1111/j.1365-294X.2009.04286.x>
- Mila, B., Carranza, S., Guillaume, O., & Clobert, J. (2010). Marked genetic structuring and extreme dispersal limitation in the Pyrenean brook newt *Calotriton asper* (Amphibia: Salamandridae) revealed by genome-wide AFLP but not mtDNA. *Molecular Ecology*, 19, 108–120.
- Miller, M. P., Blinn, D. W., & Keim, P. (2002). Correlations between observed dispersal capabilities and patterns of genetic differentiation in populations of four aquatic insect species from the Arizona White Mountains, U.S.A. *Freshwater Biology*, 47, 1660–1673. <https://doi.org/10.1046/j.1365-2427.2002.00911.x>
- Múrria, C., Bonada, N., Arnedo, M. A., Prat, N., & Vogler, A. P. (2013). Higher β - and γ -diversity at species and genetic levels in headwaters than in mid-order streams in Hydropsyche (Trichoptera). *Freshwater Biology*, 58, 2226–2236.
- Nosil, P., Egan, S. P., & Funk, D. J. (2008). Heterogeneous genomic differentiation between walking-stick ecotypes: “Isolation by adaptation” and multiple roles for divergent selection. *Evolution*, 62, 316–336. <https://doi.org/10.1111/j.1558-5646.2007.00299.x>
- Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18, 375–402. <https://doi.org/10.1111/j.1365-294X.2008.03946.x>
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlenn, D., ... Wagner, H. (2018). *Vegan: Community ecology package*. R package vegan, vers. 2.4–6. Retrieved from <https://CRAN.R-project.org/package=vegan>
- Oleksa, A., Chybicki, I. J., Gawroński, R., Svensson, G. P., & Burczyk, J. (2013). Isolation by distance in saproxylic beetles may increase with niche specialization. *Journal of Insect Conservation*, 17, 219–233. <https://doi.org/10.1007/s10841-012-9499-7>
- Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, 28, 2537–2539.
- Pelletier, T. A., Carstens, B. C., Tank, D. C., Sullivan, J., & Espíndola, A. (2018). Predicting plant conservation priorities on a global scale. *Proceedings of the National Academy of Sciences*, 115(51), 13027–13032. <https://doi.org/10.1073/pnas.1804098115>
- Polato, N. R., Gray, M. M., Gill, B. A., Becker, C. G., Casner, K. L., Flecker, A. S., ... Zamudio, K. R. (2017). Genetic diversity and gene flow decline with elevation in montane mayflies. *Heredity*, 119, 107–116. <https://doi.org/10.1038/hdy.2017.23>
- Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9, 181–199. <https://doi.org/10.1007/s10021-005-0054-1>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.
- R Development Core Team. (2015). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.
- Reisch, C. (2007). Genetic structure of *Saxifraga tridactylites* (Saxifragaceae) from natural and man-made habitats. *Conservation Genetics*, 8, 893–902.
- Renaut, S., Nolte, A. W., Rogers, S. M., Derome, N., & Bernatchez, L. (2011). SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake white fish species pairs (*Coregonus* spp.). *Molecular Ecology*, 20, 545–559. <https://doi.org/10.1111/j.1365-294X.2010.04952.x>
- Rosenberg, N. A. T., Burke, K., Elo, M. W., Feldman, P. J., Freidlin, M. A. M., Groenen, J., ... Weigend, S. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, 159, 699–713.

- Rostgaard, S., & Jacobsen, D. (2005). Respiration rate of stream insects measured in situ along a large altitude range. *Hydrobiologia*, 549, 79–98. <https://doi.org/10.1007/s10750-005-4165-7>
- Smith, M. L., & Carstens, B. C. (2019). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, 74(2), 216–229. <https://doi.org/10.1111/evo.13878>
- Torgo, L. (2013). Package 'DMwR'. *Comprehensive R Archive Network*. Retrieved from <http://cran.r-project.org/web/packages/DMwR/DMwR.pdf>
- Vähä, J., Erkinaro, J., Niemelä, E., & Primmer, C. R. (2007). Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. *Molecular Ecology*, 16, 2638–2654. <https://doi.org/10.1111/j.1365-294X.2007.03329.x>
- Vekemans, X., Beauwens, T., Lemaire, M., & Roldán-Ruiz, I. (2002). Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Molecular Ecology*, 11, 139–151. <https://doi.org/10.1046/j.0962-1083.2001.01415.x>
- Vittinghoff, E., Glidden, D. V., & McCulloch, C. E. (2012). In E. Vittinghoff, D. V. Glidden, & C. E. McCulloch (Eds.), *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. New York, NY: Springer Science & Business Media. Springer.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., van deLee, T., Hornes, M., ... Zabeau, M. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, 23, 4407–4414. <https://doi.org/10.1093/nar/23.21.4407>
- Watanabe, K., Kazama, S., Omura, T., & Monaghan, M. T. (2014). Adaptive genetic divergence along narrow environmental gradients in four stream insects. *PLoS ONE*, 9, e93055. <https://doi.org/10.1371/journal.pone.0093055>
- Watanabe, K., & Monaghan, M. T. (2017). Comparative tests of the species-genetic diversity correlation at neutral and non-neutral loci in four species of stream insect. *Evolution*, 71, 1755–1764. <https://doi.org/10.1111/evo.13261>
- Watanabe, K., Monaghan, M. T., & Omura, T. (2008). Longitudinal patterns of genetic diversity and larval density of the riverine caddisfly *Hydropsyche orientalis* (Trichoptera). *Aquatic Insects*, 70, 377–387.
- Wedger, M. J., Topp, C. N., & Olsen, K. M. (2019). Convergent evolution of root system architecture in two independently evolved lineages of weedy rice. *New Phytologist*, 223, 1031–1042. <https://doi.org/10.1111/nph.15791>
- Whiteley, A. R., Bhat, A., Martins, E. P., Mayden, R. L., Arunachalam, M., Uusi-Heikkilä, S., ... Bernatchez, L. (2011). Population genomics of wild and laboratory zebrafish (*Danio rerio*). *Molecular Ecology*, 20, 4259–4276. <https://doi.org/10.1111/j.1365-294X.2011.05272.x>
- Wilcock, H. R., Bruford, M. W., Nichols, R. A., & Hildrew, A. G. (2007). Landscape, habitat characteristics and the genetic population structure of two caddisflies. *Freshwater Biology*, 52, 1907–1929. <https://doi.org/10.1111/j.1365-2427.2007.01818.x>
- Wong, A., Smith, M. L., & Forbes, M. R. (2003). Differentiation between subpopulations of a polychromatic damselfly with respect to morph frequencies, but not neutral genetic markers. *Molecular Ecology*, 12, 3505–3513. <https://doi.org/10.1046/j.1365-294X.2003.02002.x>
- Yaegashi, S., Watanabe, K., Monaghan, M. T., & Omura, T. (2014). Fine-scale dispersal in a stream caddisfly inferred from spatial autocorrelation of microsatellite markers. *Molecular Approaches in Freshwater Ecology*, 33, 172–180. <https://doi.org/10.1086/675076>
- Zhivotovsky, L. A. (1999). Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, 8, 907–913. <https://doi.org/10.1046/j.1365-294x.1999.00620.x>

How to cite this article: Li B, Yaegashi S, Carvajal TM, et al. Machine-learning-based detection of adaptive divergence of the stream mayfly *Ephemera strigata* populations. *Ecol Evol*. 2020;10:6677–6687. <https://doi.org/10.1002/ece3.6398>