Abbreviated Title: Machine learning based divergence detection

Detection of adaptive divergence in populations of the stream mayfly *Ephemera strigata* with machine learning

Bin Li[1,*], Sakiko Yaegashi[1,2], Thaddeus M Carvajal[1,3], Maribet Gamboa[1] and Kozo Watanabe[1,3]

[1]Department of Civil and Environmental Engineering, Ehime University, Bunkyo-cho 3, Matsuyama, 790-8577, Japan

[2]Department of Civil and Environmental Engineering, University of Yamanashi, 4-3-11 Takeda, Kofu, Yamanashi 400-851, Japan

[3]Biological Control Research Unit, Center for Natural Science and Environmental Research, De La Salle University, Taft Ave Manila, Philippines

Abstract

Adaptive divergence is a key mechanism shaping the genetic variation of natural populations. A central question linking ecology with evolutionary biology concerns the role of environmental heterogeneity in determining adaptive divergence among local populations within a species. In this study, we examined adaptive the divergence among populations of the stream mayfly *Ephemera strigata* in the Natori River Basin in northeastern Japan. We used a genome scanning approach to detect candidate loci under selection and then applied a machine learning method (i.e. Random Forest) and traditional distance-based redundancy analysis (dbRDA) to examine relationships between environmental factors and adaptive divergence at non-neutral loci. We also assessed spatial autocorrelation at neutral loci to quantify the dispersal ability of *E. strigata*. Our main findings were as follows: 1) random forest shows a higher resolution than traditional statistical analysis for detecting adaptive divergence; 2) separating markers into neutral and non-neutral loci provides insights into genetic diversity, local adaptation and dispersal ability and 3) *E. strigata* shows altitudinal adaptive divergence among the populations in the Natori River Basin.

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

Email address: [1]binglee527@gmail.com; [2]sakikoy@yamanashi.ac.jp; [3]tads.carvajal@gmail.com; [4]maribetg@gmail.com; [5]watanabe_kozo@cee.ehime-u.ac.jp

A central question linking ecology with evolutionary biology concerns the role of environmental heterogeneity in determining adaptive divergence among local populations within a species. Adaptive divergence in aquatic insects is usually reported to be influenced by altitudinal gradients at the river-corridor scale (Hughes et al. 2009, Keller et al. 2013, Polato et al. 2017). Altitude is often strongly related with a number of environmental factors, such as temperature and oxygen, which greatly influenced the biology of organisms (Keller and Seehausen 2012, Halbritter et al. 2015). Thermal regimes directly regulate species' growth, development and mating behaviour, thereby setting limits on species distributions and abundances across landscapes (Li et al. 2013). Oxygen availability also restricts species' distributions by affecting the respiratory metabolism of aquatic organisms (Rostgaard and Jacobsen 2005). Multiple studies have focused on the genetic basis of adaptive divergence in aquatic insects because of their importance in freshwater ecosystem biomonitoring. Altitudinal genetic divergence has been reported in aquatic insects including caddisflies (*Plectrocnemia conspersa* and *Polycentropus flavomaculatus* (Wilcock et al. 2007), *Stenopsyche maramorata* (Yaegashi et al. 2014), stoneflies (*Dinocras cephalotes*) (Elbrecht et al. 2014) and mayflies (Atalophlebia) (Baggiano et al. 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 and Freeland 2011). Therefore, based on the screening of a large numbers of candidate loci ('outlier' loci, reviewed by Nosil et al. 2009) and the estimation of the levels of genetic divergence, statistical methods can identify loci that are under direct selection or linked to loci under selection. Selected non-neutral loci can be used to test hypotheses about the adaptive process. Neutral loci may be available for accurate tests of neutral processes, such as isolation by distance (IBD) (Oleksa et al. 2013) and gene flow patterns, thereby avoiding the confounding effects of natural selection (Kirk and Freeland 2011).

In the ordinal analysis of genome scanning, non-neutral loci are detected based on genetic variation among populations with different phenotypes or ecotypes (Bonin et al. 2006, Nosil et al. 2008, Egan et al. 2008, Galindo et al. 2009) or allopatric populations among different geographic localities (Medugorac et

al. 2009, Gaggiotti et al. 2009, Renaut et al. 2011). Genome scanning can also be conducted using genetically defined populations with unknown phenotypes or ecotypes. For example, Bayesian clustering methods (Pritchard et al. 2000, Falush et al. 2003, 2007) 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).

The determination of 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 and Fortin 2010, Watanabe et al. 2014), distance-based redundancy analysis (dbRDA) (Watanabe and Monaghan 2017) and multivariate analysis of variance (MANOVA (Mccairns and Bernatchez 2008) have been widely applied, but these methods suffer from a number of limitations. First, associating genetic variance and environmental distances can result in bias and high error rates (Legendre and Fortin 2010, Guillot and Rousset 2013, Legendre et al. 2015). In addition, the Mantel test and dbRDA are limited to testing the linear independence between genetic and environmental distances among local populations. Fulfilling the underlying assumptions of conventional statistical methods (e.g. normal distribution and homogeneity of variance) can also be very difficult (Vittinghoff et al. 2012). On account 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 (Prasad et al. 2006, Biau and Scornet 2016).

Among the variety of machine learning methods available, Random Forest (RF) (Breiman 2001) is one of the most widely used modelling techniques to generate high-prediction accuracy and evaluate the relative importance of explanatory variables in the model (Biau and Scornet 2016). RF is an ensemble tree-based method that constructs multiple decision trees from a dataset 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 species' distributions and identify constrained environmental factors (Cutler 2007, Marmion et al. 2009, Evans et al. 2011). In most studies, environmental data have been used as independent variables to predict the presence or absence of species' (dependent variables). The relative contributions of

84 environmental variables to species distributions are quantified by their relative importance obtained from
85 the RF model. It may therefore be possible to extend the use of RF to population genetic studies where
86 environmental variables are used to predict the presence or absence of a haplotype or allele at outlier
87 loci. The relative importance of each environmental variable could be considered as its influence to outlier
88 loci, which may strongly drive adaptive divergence.

89 In this study, we examined adaptive divergence using AFLP markers in populations of the stream mayfly
90 *E. strigata* from the Natori River Basin in northeastern Honshu Island, Japan (Fig.1). The primary aims of
91 the study were to determine the extent of local adaptation at the genome level in natural populations and
92 to quantify associations between environmental gradients and adaptive divergence. We first detected loci
93 under selection (non-neutral loci) based on locus-specific genetic differentiation among populations.
94 Rather than defining populations a priori using geographic or phenotypic information, we delineated
95 populations based on the discontinuities in the AFLP variation among individuals using a hierarchical
96 analysis of STRUCTURE (Pritchard et al. 2000, Falush et al. 2003, 2007, Vähä et al. 2007). Focusing on
97 non-neutral loci, we then applied RF to identify environmental variables most likely to contribute to
98 adaptive divergence and compared our results with a traditional dbRDA to examine the feasibility of the
99 method. Finally, we examined the dispersal patterns and dispersal distance in *E. strigata* using neutral
100 loci.

101 ## Methods

102 ### Study site and sampling

103 *E. strigata* is a burrowing mayfly well studied in Japan and Korea (Ban and Kawai, 1986; Lee et al.,
104 2008). In this study, sampling was carried out in the Natori River catchment in the Miyagi Prefecture in
105 northeastern Japan (Fig. 1). Larval samples were collected at 11 sites from October 26 to November 12,
106 2010. At each site, we collected *E. strigata* individuals using a Surber net (30 × 30 cm quadrat with mesh
107 size 250 μm) along 200–900 m stream reaches. All specimens were preserved in the field in 99.5%
108 ethanol, transported to the laboratory and identified to species level under a stereomicroscope (120×)
109 using taxonomic keys (Kawai and Tanida 2005).

We measured six geographical parameters at each site using standard ecological methods in stream surveys (Hauer and Lamberti 2007, Watanabe et al. 2008). Stream order was determined using a 1:25000 map. The width of the stream channel was measured 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 river side. The riverine distance between two sites was measured on Google Maps using the ruler function.

DNA extraction and AFLP fingerprinting

DNA from each individual was isolated from abdominal tissue by removing the digestive tract using the DNeasy 96 Blood & Tissue Kits (Qiagen). The concentration of extracted DNA was measured by Nano Drop 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 100μM MseI adapter, 0.2μM of EcoRI adapter, 2 μL T4 DNA ligase buffer (10×) (New England) and up to 20 μL dH_2O and incubating the solution at 16°C for 12 h. The sequences of the MseI adapter and EcoRI adapters followed Reisch (2007). The adapters were manually prepared as follows: 1) mixing equal molar amounts of adapter oligomer, 2) denaturing at 95°C for 5 min and 3) incubating for 10 min at room temperature. Restricted or ligated products were then diluted at a 1:19 ratio with 0.1× TE buffer. Pre-selective amplification was performed in a mixture of 0.06 μL of 100μM MseI and EcoRI primers (Reish 2007). 15 μL of AFLP Amplification Core Mix (Applied Biosystems), 4 μL of each restricted/ligated product and up to 29 uL dH_2O. Pre-selective polymerase chain reaction (PCR) parameters followed Reish (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 Dye2, 3 and 4 in 5'-end. The mixture of selective PCR was 0.1 μL of 100μM MseI and EcoRI primers, 15μL of AFLP Amplification Core Mix (Applied Biosystems) and up to 20 uL dH_2O. We followed Reich (2007) to set PCR reaction parameters.

The selective PCR products were separated by capillary gel electrophoresis using CEQ8000 (Beckman Coulter). To adjust fluorescent intensity, each fluorescent PCR product was mixed with the following proportion EcoRI-AGG & MseI-CAT 4 µL, EcoRI-ACC & MseI-CAC 2µL and EcoRI-AGG & MseI-CAC 1µL. Peak sizes of PCR products were calculated with DNA Size Standard 600 (Beckman Coulter) using the CEQ8000 software (Beckman Coulter) with default settings.

Hierarchal STRUCTURE analysis

We defined populations based on discontinuities in AFLP variation using the individual-based Bayesian clustering method implemented in STRUCTURE v. 2.3 (Pritchard et al. 2000, Falush et al. 2003, 2007). 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 Alphapropsd 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 et al. 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 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,

163 individuals were assigned to subpopulations based on the individual membership coefficient (Pritchard et
164 al. 2000).

165 **Outlier loci detection**

166 We used two different statistical methods to identify outlier loci. Dfdist (adapted from Fdist (Beaumont and
167 Nichols 1996)) uses coalescent simulations to generate thousands of loci evolving under a neutral model
168 of symmetrical islands with a mean global F_{ST} close to the observed global F_{ST}. Mean F_{ST} was calculated
169 using the default method by first excluding 30% of the highest and lowest observed values. Empirical loci
170 with F_{ST} values significantly greater ($p < 0.05$) than the simulated distribution (generated with 50,000 loci)
171 were considered to be outliers. Dfdist can detect both divergent selection and balancing selection, but we
172 focused only on divergent selection in this study.

173 BayeScan is a hierarchical Bayesian model-based method first described in Beaumont and Balding
174 (2004) and modified by Foll and Gaggiotti (2008) for dominant markers (available at
175 http://cmpg.unibe.ch/software/bayescan/). The Bayesian method is based on the concept that F_{ST} values
176 reflect contributions from locus-specific effects, such as selection, and population-specific effects, such as
177 local effective size and immigration rates. The main advantage of this approach is that it allows for
178 different demographic scenarios and different amounts of genetic drift in each population (Foll and
179 Gaggiotti 2006, 2008). Using a reversible jump Markov Chain Monte Carlo approach, the posterior
180 probability of each locus being subjected to selection is estimated. A locus is deemed to be influenced by
181 selection if its F_{ST} is significantly higher or lower than the expectation provided by the coalescent
182 simulations.

183 For all subsequent analyses, non-neutral loci were defined as outlier loci detected by the Dfdist and
184 BayeScan methods at the 95% confidence level. Neutral loci were defined as loci detected by neither
185 Dfdist nor BayeScan at the 95% thresholds. Loci detected as outliers by only one of the two methods
186 were not considered in the further analyses.

187 **Analysis of genetic diversity**

188 F_{ST} was calculated with ARLEQUIN v. 3.1 (Excoffier et al. 2009) using: 1) all loci, 2) only neutral loci and
189 3) only non-neutral loci. Global heterozygosity among all populations (H_t) and mean heterozygosity within
190 populations (H_w) were estimated separately for neutral and non-neutral loci with AFLP-SURV v. 1.0
191 (Vekemans 2002) using the Bayesian method with a uniform prior distribution of allele frequencies
192 (Zhivotovsky 1999). Molecular variance analysis (AMOVA) was also conducted using ARLEQUIN to
193 provide the estimates of genetic variations among and within sampling sites. For the test of IBD, we
194 examined the correlations of pairwise F_{ST} with geographical distance and riverine distance (i.e. distance
195 along the watercourse) between sites using GeneAlEx v. 6.5 (Peakall and Smouse 2012). The genetic
196 distance between each pair of sites was quantified using mean pairwise F_{ST} for neutral and non-neutral
197 loci using the Bayesian-estimated allele frequencies generated by AFLP-SURV.

198 We conducted genetic spatial autocorrelation analysis using neutral loci for geographic distance. Eight
199 geographic distance classes defined every 4 km (from 0–4 km to 28–32 km) were used in the analysis.
200 Individuals within the same site were considered to be separated by a distance of 0 km. We calculated
201 Moran's I for each distance class using GeneAlEx, where I ranges from −1 to 1, and the positive values
202 indicate that sites within a given distance class have similar genetic structure. We used jackknifing to
203 estimate the 95% confidence intervals.

204 **Adaptive divergence modelling**

205 We determined the environmental variables that drive adaptive divergence at non-neutral loci using the
206 RF model (Chawla et al. 2002, Maciejewski and Stefanowski 2011, Blagus and Lusa 2013). All the six
207 environmental variables were used to predict the band presence/absence patterns at non-neutral loci. We
208 assigned individuals from the same site to the same environmental conditions. The dataset was
209 imbalanced because the number of individuals with band presence was not equal to that with band
210 absence. The individuals were thus classified in two classes (i.e. presence and absence). We solved the
211 data imbalance problem by oversampling for the minority class using the Synthetic Minority Oversampling
212 Technique (SMOTE) (Chawla et al. 2002). SMOTE creates synthetic minority class sample units by taking
213 the difference between the feature vector (sample) under consideration and its nearest neighbour. It then
214 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 process was conducted using the DMwR (Torgo 2013) and randomForest packages (Liaw and Wiener 2002) in the R programme (R Core Development Team 2015). Model performance was evaluated using the area under the receiver operating characteristic curve (AUC) (Janitza et al. 2013). The AUC value typically ranged from 0.5 (random prediction) to a maximum value of 1, which represents the perfect model theoretically. As rules of thumb, an AUC value > 0.9 indicates very good model quality, a value < 0.7 indicates poor model quality, and a value ranging from 0.7 to 0.9 indicates good model quality (Baldwin 2009).

We also conducted dbRDA as a comparative ordinal method. DbRDA was performed on the ordination solutions, rather than on the distance matrices (Legendre and Fortin 2010). In this study, pairwise genetic distances at non-neutral locus among sites were used to screen environmental factors that most closely relate to genetic divergence (Watanabe et al. 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. 2015) in the R programme (R Core Development Team 2015). Significant differences were tested with the anova.cca function in the vegan package.

Results

Hierarchical STRUCTURE analysis

Hierarchical iterations by STRUCTURE detected significant substructure until the 4th iteration beyond the initial analysis (Fig. 2). In total, 14 groups were defined for the 216 *E. strigata* individuals collected in 11 sites. Most groups were widespread over the sampling sites, whereas some groups were restricted to specific sites. For example, the members of groups 2, 3 and 8 occurred only in up- and middle-stream sites (Fig. 1: upstream sites, S1 and S6-8; middle-stream sites, S2-5).

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, respectively. Outlier loci under balancing selection were not involved in this study. All the 10 outlier loci under divergent selection were consistently identified by BayeScan, which alone identified 26 outliers (Table 1). 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 ($p < 0.01$; AMOVA). When measured using neutral or non-neutral loci, we found global F_{ST} values of 0.021 ($p < 0.01$) and 0.039 ($p < 0.01$), respectively (Table 2).

Detection of adaptive divergence

We separately built RF models for each of the 10 non-neutral loci. 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 (Fig. 3), With dbRDA, only genetic divergence at locus 254 was significantly predicted ($p < 0.05$) (Fig. 4) with altitude explaining 54% of the genetic divergence at this locus. For the other non-neutral loci, no significant relationship with environmental factors was found with dbRDA ($p > 0.05$). IBD was not significant for both geographic ($r = 0.11$, $p = 0.33$) and riverine distance ($r = 0.06$, $p = 0.49$) (Supplementary Fig. S1). The results of the spatial autocorrelation analysis based on neutral loci showed significant positive autocorrelation coefficients at the shortest range (0–4 km; Fig. 5).

Discussion

In this study, we used an RF model to examine the relationship between environmental factors and adaptive divergence at non-neutral loci in the stream mayfly *E. strigata*. Ordinal statistical tests of multiple linear regression method need assumptions that data are normally distributed with homogeneity of variance and are independent from one another (Vittinghoff et al. 2012), and this is often difficult to fulfil. The environmental factors investigated in this study did not show strong independency among variables. However, RF can overcome the limitations of regression models and accommodate pronounced nonlinearities in the exploration of gene-environment relationships in large genomic data sets (Breiman 2001, Fitzpatrick and Keller 2015, Biau and Scornet 2016). We developed RF models for each of the 10 non-neutral loci detected by both BayeScan and Dfdist. Three out of the 10 non-neutral loci (56, 89 and

254) showed good model prediction performance (AUC > 0.7), whereas the other seven could not be modelled well. This may be explained by natural selection at these seven loci being driven by environmental factors not included in our analysis (e.g. velocity and chlorophyll a) (Watanabe et al. 2014, Li et al. 2016, Brouwer et al. 2017). RF is recommended for future studies including huge numbers of environmental variables to assess their effects on adaptive divergence because RF can perform well with large numbers of variables (Genuer et al. 2010).

To compare the performance of RF with ordinal statistical analysis, we also conducted dbRDA analysis for all the 10 non-neutral loci. Only one locus (254) was well-modelled by dbRDA. This locus was one of the three loci accurately modelled by RF and the selected environmental factor (i.e. altitude) was consistent with results from RF. The low number of loci modelled in dbRDA may be because of its ability to only test linear independence[29]. The ranking of variable importance in RF relies on the principle that rearranging the values of unimportant variables should not degrade the predictive accuracy of the model (Breiman 2001). As a result, RF could reduce the influence of variable dependency on model results compared with dbRDA (Archer and Kimes 2008, Genuer et al. 2010).

To identify non-neutral loci, we used populations delineated by a hierarchal STRUCTURE analysis as an alternative to the geographic or phenotypic populations that are typically used in ordinal analysis of genome scanning. The STRUCTURE analysis successfully delineated populations with significant genetic differences, something that is difficult to achieve 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 the occurrence of observable phenotypic divergence and may provide a means to investigate the early stages of adaptive divergence prior to phenotypic divergence in population delineation and detection of non-neutral loci (Whiteley et al. 2011).

We introduced a hierarchical approach to the STRUCTURE analysis that enabled us to look at the finer population structure (i.e. higher K) than the ordinal STRUCTURE analysis, which stops once the uppermost hierarchical level is found. The number of populations (K) is an important determinant in outlier detection (Foll and Gaggiotti 2008). We conducted outlier loci detection based on the geographical populations and uppermost hierarchical level of the STRUCTURE analysis that delineated two

293 populations, but we could not detect any outlier loci. This clearly shows the advantages of using a
294 hierarchical approach to STRUCTURE analysis. However, a deeper hierarchical level (e.g. the 4th
295 iteration in the hierarchy) will define a weaker structure at the risk of detecting extremely fine population
296 substructures.

297 By employing a genome scanning approach, we comparatively used neutral and non-neutral loci in
298 examining genetic diversity and genetic distance. Importantly, we found greater genetic divergence at
299 non-neutral loci than neutral loci. This pattern is consistent with the study of three caddisflies species and
300 one mayfly species in the same catchment system (Watanabe et al. 2014). Other studies also found
301 similar pattern of lower levels genetic divergence in neutral DNA markers compared with morphological
302 traits (analogues to non-neutral markers) in macroinvertebrate species such as snails (Cook 1992),
303 spiders (Gillespie and Oxford 1998) and damselflies (Wong et al. 2003). Based on the results of Dfdist,
304 the 10 non-neutral loci were under divergent selection rather than stabilising selection, and hence
305 presented greater genetic divergence compared with neutral loci (Table 2).

306 One of the main findings of this study is that the mountain burrowing mayfly *E. strigata* presents an
307 adaptive divergence along an altitudinal gradient. Altitude is often reported to be closely related with a
308 number of environmental factors that influence the life cycle and development of organisms (Múrria et al.
309 2013, Halbritter et al. 2015). For example, altitude influences insect phenology, restricting the mating
310 period to only a few days, thus leading to asynchronous emergence, which may act as a reproductive
311 barrier between populations (Yaegashi et al. 2014, Watanabe et al. 2017) or as metabolism regulator
312 (Gamboa et al. 2017). Altitude also influences air density which affects both respiration and the power
313 required for flight. The haemoglobin gene and other genes with a potential role for adaptation to low O_2
314 may show divergence between populations along an altitude gradient (Keller et al. 2013).

315 As opposed to non-neutral makers, neutral markers are suitable for examining neutral process occurring
316 under the drift-migration balance. Previous population genetic studies have inferred dispersal patterns of
317 stream insects without differentiating neutral and non-neutral loci (Miller et al. 2002, Mila et al. 2010). This
318 may cause an overestimation of genetic drift because non-neutral loci under divergent selection will

increase the estimates of genetic divergence (Kirk and Freeland 2011). Therefore, we used only neutral makers to infer dispersal patterns.

We did not find significant IBD for both geographic and riverine distances based on neutral loci, suggesting that populations are not in a genetic drift–migration equilibrium at the studied geographic scale (Supplementary Fig. S1). The results of the spatial autocorrelation analysis based on neutral loci showed significant positive autocorrelation coefficients at the shortest range (0−4 km; Fig. 5a), indicating low dispersal ability for *E. strigata.* Mayflies are generally considered to have a very low dispersal ability in mountain streams (Barber-James et al. 2007). Limited dispersal distances were also observed in stoneflies owing to their poor dispersal ability (Briers et al. 2003, 2004). In contrast, caddisflies were frequently reported to show strong dispersal ability. Yaegashi et al. (2014) reported that the caddisfly *Stenopsyche marmorata* showed pronounced dispersal ability along stream corridors up to 12 km.

In conclusion, the RF approach applied in this study performed better than the ordinal dbRDA in determining the influence of environmental factors on outlier loci under selection. Using neutral and non-neutral methods, we showed that the mountain burrowing mayfly *E. strigata* presents adaptive divergence along an altitudinal gradient. The hierarchical STRUCTURE analysis detected finer population structures and increased the power of outlier detection. A limitation of this study was that our study did not include many environmental factors, which may also be constrained factors and help to improve the model performance. In addition, sequencing the detected outlier loci would provide a deeper understanding of altitudinal adaptation in *E. strigata*.

References

Archer, K. J. and R. V. Kimes. 2008. Empirical characterization of random forest variable importance measures. Computational statistics & data analysis 52:2249–2260.

Baggiano, O., D. J., Schmidt, and J. M., Hughes. 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.

344 Baldwin, R. A. 2009. Use of Maximum Entropy Modeling in Wildlife Research. Entropy 11:854–866.

345 Ban, R., and T. Kawai. 1986. Comparison of the life cycles of two mayfly species between upper and
346 lower parts of the same stream. Aquatic Insects 8:207–215.

347 Barber-James, H. M., J-L. Gattolliat, and M. D. Hubbard. 2007. Global diversity of mayflies
348 (Ephemeroptera, Insecta) in freshwater. Hydrobiologia 595:339–350.

349 Beaumont, M. A. and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations
350 from genome scans. Mol. Ecol 13:969–980.

351 Beaumont, M. A. and R. A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population
352 structure. Proc. R Soc. Lond B 263:1619–1626.

353 Biau, G. and E. Scornet. 2016. A random forest guided tour. TEST 25:197–227.

354 Blagus, R. and L. Lusa. 2013. SMOTE for high–dimensional class–imbalanced Data. BMC Bioinformatics
355 14: 106.

356 Bonin, A., P. Taberlet, and F. Pompanon. 2006. Explorative genome scan to detect candidate loci for
357 adaptation along a gradient of altitude in the commonfrog (*Rana temporaria*). Mol Biol Evol 23:773–
358 783.

359 Breiman, L. 2001. Random Forests. Machine Learning 45:5-32.

360 Briers, R. A., H. R. Gee, and R. Geoghegan. 2004. Inter–population dispersal by adult stoneflies detected
361 by stable isotope enrichment. Freshwater biology 49:425–431.

362 Briers, R., H. Cariss, and J. H. R. Gee. 2003. Flight activity of adult stoneflies in relation to weather.
363 Ecological Entomology 28:31–40.

364 Brouwer, J. H. F., A. Bessee-Lototskaya, and P. F. M. Verdonschot. 2017. Flow velocity tolerance of
365 lowland stream caddisfly larvae (Trichoptera). Aquat Sci 79:419–425.

366 Chawla, N. V., K. W. Bowyer, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over–sampling
367 Technique. Journal of Artificial Intelligence Research 16:321–357.

368 Cook, L. 1992. The neutral assumption and maintenance of color morph frequency in mangrove snails.
369 Heredity 69:184–189.

370 Cutler, D. R. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

371 Egan, S. P., P. Nosil, and D. J. Funk. 2008. Selection and genomic differentiation during ecological
372 speciation: isolating the contributions of host association via a comparative genome scan of
373 *neoclamisus bebbianae* leaf beetles. Evolution 62:1162–1181.

374 Elbrecht, V., C. K. Feld, and F. Leese. 2014. Genetic diversity and dispersal potential of the stonefly
375 Dinocras cephalotes in a central European low mountain range. Freshwater Science 33:181–192.

376 Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the
377 software structure: a simulation study. Molecular Ecology 14:2611–2620.

378 Evans, J. S., M. A. Murphy, and S. A. Cushman. 2011. Modeling Species Distribution and Change Using
379 Random Forest. Pages 139–159 *in* C. Drew, Y. Wiersma, F. Huettmann (eds). Predictive Species and
380 Habitat Modeling in Landscape Ecology. Springer, New York, NY.

381 Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured
382 population. Heredity 103:285–298.

383 Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus
384 genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

385 Falush, D., M. Stephens, and J. K. Pritchard. 2007. Inference of population structure using multilocus
386 genotype data: dominant markers and null alleles. Molecular Ecology Notes. 7:574–578.

387 Fitzpatrick, M. C. and S. R. Keller. 2015. Ecological genomics meets community–level modelling of
388 biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecology
389 Letters 18:1–6.

390 Foll, M. and O. E. Gaggiotti. 2006. Identifying the environmental factors that determine the genetic
391 structure of populations. Genetics 174:875–891.

392 Foll, M., and O. E. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both
393 dominant and codominant markers: a Bayesian perspective. Genetics 180:977–993.

394 Gaggiotti, O. E., D. Bekkevold, and D. E. Ruzzante. 2009. Disentangling the effects of evolutionary,
395 demographic, and environmental factors influencing genetic structure of natural populations: Atlantic
396 herring as a case study. Evolution 63:2939–2951.

397 Galindo, J., and E. Rolán-Alvarez. 2009. Comparing geographical genetic differentiation between
398 candidate and noncandidate loci for adaptation strengthens support for parallel ecological divergence
399 in the marine snail *Littorina saxatilis*. Molecular Ecology 18:919–930.

400 Gamboa, M., M. C. Tsuchiya, and K. Watanabe. 2017. Differences in protein expression among five
401 species of stream stonefly (Plecoptera) along a latitudinal gradient in Japan. Insect biochemistry and
402 physiology. 96:e21422.

403 Genuer, R., J-M. Poggi, and C. Tuleau–Malot. 2010. Variable selection using random forests. Pattern
404 Recognition Letters 31:2225–2236.

405 Gillespie, R. G. and G. S. Oxford. 1998. Selection on the color polymorphism in hawallan happy–face
406 spiders: evidence from genetic structure and temporal fluctuations. Evolution 52:775–783.

407 Guillot, G. and F. Rousset. 2013. Dismantling the Mantel tests. Methods in Ecology and Evolution 4:336–
408 344.

409 Halbritter, A. H., R. Billeter, and J. M. Alexander. 2015. Local adaptation at range edges: comparing
410 elevation and latitudinal gradients. Journal of Evolutionary Biology 28:1849–1860.

411 Hauer, F. R., G. A. Lamberti. 2007. Methods in stream ecology. 3rd edition. Academic Press, London.

412 Hughes, J. M., D. J. Schmidt, and D. S. Finn. 2009. Genes in Streams: Using DNA to Understand the
413 Movement of Freshwater Fauna and Their Riverine Habitat. BioScience 59:573–583.

414 Hwang, J. M. and Y. J. Bae. 2008. Review of the tropical Southeast Asian Ephemera (Ephemeroptera:
415 Ephemeridae). Aquatic Insects. 30, 105–126.

416 Janitza, S., C. Strobl. And A. L. Boulsesteix. 2013. An AUC–based permutation variable importance
417 measure for random forests. BMC Bioinformatics 14:119.

418 Kawai T., and K. Tanida. 2005. Aquatic insects of Japan: manuals with keys and illustration (in
419 Japanese). Tokai University Press, Tokyo.

420 Keller, I., and O. Seehausen. 2012. Thermal adaptation and ecological speciation. Molecular Ecology
421 21:782–799.

422 Keller, I., J. M., Alexander, and P. J., Edwards. 2013. Widespread phenotypic and genetic divergence
423 along altitudinal gradients in animals. Journal of Evolutionary Biology 26:2527–2543.

424 Kirk, H., and J. R. Freeland. 2011. Applications and Implications of Neutral versus Non-neutral Markers in
425 Molecular Ecology. International Journal of Molecular Sciences 12:3966–3988.

426 Lee, S. J., J. M. Hwang, and Y. J. Bae. 2008. Life history of a lowland burrowing mayfly, *Ephemera*
427 *orientalis* (Ephemeroptera: Ephemeridae), in a Korean stream. Hydrobiologia 596:279–288.

428 Legendre, P., and M. J. Fortin. 2010. Comparison of the Mantel test and alternative approaches for
429 detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology
430 Resources 10: 831–844.

431 Legendre, P., M-J. Fortin, and D. Borcard. 2015. Should the Mantel test be used in spatial analysis?
432 Methods in Ecology and Evolution 6:1239–1247.

433 Li, B. K. Watanabe, and T-S. Chon. 2016. Identification of Outlier Loci Responding to Anthropogenic and
434 Natural Selection Pressure in Stream Insects Based on a Self–Organizing Map. Water 8:188.

435 Li, F. Q., N. Chung, and Y-S. Park. 2013. Temperature change and macroinvertebrate biodiversity:
436 assessments of organism vulnerability and potential distributions. Climatic Change 119:421–434.

437 Liaw A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.
438 http://CRAN.R–project.org/doc/Rnews/.

439 Maciejewski, T. and J. Stefanowski. 2011. Local neighbourhood extension of SMOTE for mining
440 imbalanced data. Computational Intelligence and Data Mining. 104-111.

441 Marmion, M., M. Parviainen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive
442 species distribution modelling. Diversity and Distributions 15:59–69.

443 Mccairns, R. J. S. and L. Bernatchez. 2008. Landscape genetic analyses reveal cryptic population
444 structure and putative selection gradients in a large–scale estuarine environment. Molecular Ecology
445 17:3901–3916.

446 Medugorac, I., A. Medugorac, and M. Förster. 2009. Genetic diversity of European cattle breeds
447 highlights the conservation value of traditional unselected breeds with high effective population size.
448 Molecular Ecology 18:3394–3410.

449 Mila, B., S. Carranza, O. Guillaume, and J. Clobert. 2010. Marked genetic structuring and extreme
450 dispersal limitation in the Pyrenean brook newt *Calotriton asper* (Amphibia: Salamandridae) revealed
451 by genome–wide AFLP but not mtDNA. Molecular Ecology 19:108–120.

452 Miller, M. P., D. W. Blinn, & P. Keim. 2002. Correlations between observed dispersal capabilities and
453 patterns of genetic differentiation in populations of four aquatic insect species from the Arizona White
454 Mountains, U.S.A. Freshwater Biology 47:1660–1673.

455 Múrria, C., N. Bonada, A. P. Vogler. 2013. Higher β- and γ-diversity at species and genetic levels in
456 headwaters than in mid-order streams in Hydropsyche (Trichoptera). Freshwater Biology 58:2226–
457 2236.

458 Nosil, P., D. J. Funk, and D. Ortiz-Barrientos. 2009. Divergent selection and heterogeneous genomic
459 divergence. Molecular Ecology 18:375–402.

460 Nosil, P., S. P. Egan, and D. J. Funk. 2008. Heterogeneous genomic differentiation between walking–
461 stick ecotypes: “Isolation by adaptation” and multiple roles for divergent selection. Evolution 62:316–
462 336.

463 Oksanen, J. et al. 2018. Vegan: community ecology package. R package vegan, vers. 2.4–6.
464 https://CRAN.R-project.org/package=vegan.

465 Oleksa, A., I. J. Chybicki, and J. Burczyk. 2013. Isolation by distance in saproxylic beetles may increase
466 with niche specialization. J Insect Conserv 17:219–233.

467 Peakall, R. and P. E. Smouse. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software
468 for teaching and research–an update. Bioinformatics 28:2537–2539.

469 Polato, N. R. M. M., Gray, and K. R. Zamudio. 2017. Genetic diversity and gene flow decline with
470 elevation in montane mayflies. Heredity 119:107–116.

471 Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques:
472 bagging and random forests for ecological prediction. Ecosystems 9:181–199.

473 Pritchard, J. K., M. Stephens, P. Donnelly. 2000. Inference of population structure using multilocus
474 genotype data. Genetics 155:945–959.

475 R Development Core Team. 2015. R: A Language and Environment for Statistical Computing. (R
476 Foundation for Statistical Computing).

477 Reisch, C. 2007. Genetic structure of *Saxifraga tridactylites* (Saxifragaceae) from natural and man-made
478 habitats. Conserv Genet 8:893-902.

479 Renaut, S., A. W. Nolte, and L. Bernatchez. 2011. SNP signatures of selection on standing genetic
480 variation and their association with adaptive phenotypes along gradients of ecological speciation in
481 lake white fish species pairs (*Coregonus spp.*). Molecular Ecology 20:545–559.

482 Rosenberg, N. A. T. Burke, and S. Weigend. 2001. Empirical evaluation of genetic clustering methods
483 using multilocus genotypes from 20 chicken breeds. Genetics 159:699–713.

484 Rostgaard, S., and D. Jacobsen. 2005. Respiration rate of stream insects measured in situ along a large
485 altitude range. Hydrobiologia 549:79–98.

486 Torgo, L. 2013. Package 'DMwR'. Comprehensive R Archive Network. http://cran.r–
487 project.org/web/packages/DMwR/DMwR.pdf.

488 Vähä, J., J. Erkinaro, and C. R. Primmer. 2007. Life–history and habitat features influence the within–river
489 genetic structure of Atlantic salmon. Molecular Ecology 16:2683–2654.

490 Vekemans X., T. Beauwens, and I. Roldan-Ruiz. 2002. Data from amplified fragment length
491 polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between
492 degree of homoplasy and fragment size. Molecular Ecology 11:139-151.

493 Vittinghoff, E., D. V. Glidden, and C. E. Mcculloch. 2012. Regression Methods in Biostatistics: Linear,
494 Logistic, Survival, and Repeated Measures Models. *In* Vittinghoff, E, D. V. Glidden, and C. E.
495 McCulloch (eds). Springer Science & Business Media. Springer, New York.

496 Vos, P., R. Hogers, and M. Zabeau. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids
497 Research 23:4407–4414.

498 Watanabe, K. and M. T. Monaghan. 2017. Comparative tests of the species-genetic diversity correlation
499 at neutral and non-neutral loci in four species of stream insect. Evolution 71:1755−1764.

500 Watanabe, K., M. T. Monaghan, and T. Omura. 2008. Longitudinal patterns of genetic diversity and larval
501 density of the riverine caddisfly *Hydropsyche orientalis* (Trichoptera). Aquatic insects. 70:377–387.

502 Watanabe, K., S. Kazama, and M. T. Monaghan. 2014. Adaptive Genetic Divergence along Narrow
503 Environmental Gradients in Four Stream Insects. PLoS ONE 9:e93055.

504 Whiteley, A. R. A. Bhat, and L. Bernatchez. 2011. Population genomics of wild and laboratory zebrafish
505 (*Danio rerio*). Molecular Ecology 20:4259–4276.

506 Wilcock, H. R., M. W., Bruford, and A. G., Hildrew. 2007. Landscape, habitat characteristics and the
507 genetic population structure of two caddisflies. Freshwater Biology 52:1907–1929.

508 Wong, A., M. L. Smith, and M. R. Forbes. 2003. Differentiation between subpopulations of a
509 polychromatic damselfly with respect to morph frequencies, but not neutral genetic markers. Molecular
510 Ecology 12:3505–3513.

511 Yaegashi, S., K. Watanabe, and T. Omura. 2014. Fine–scale dispersal in a stream caddisfly inferred from
512 spatial autocorrelation of microsatellite markers. Molecular approaches in freshwater ecology 33:172–
513 180.

514 Zhivotovsky, L. A. 1999. Estimating population structure in diploids with multilocus dominant DNA
515 markers. Molecular Ecology 8:907–913.

516

517 Acknowledgements

518 This research was financially supported by the Japan Society for the Promotion of Science (JSPS) (grant
519 numbers: 16H04437, 17H01666, 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.

Author Contributions

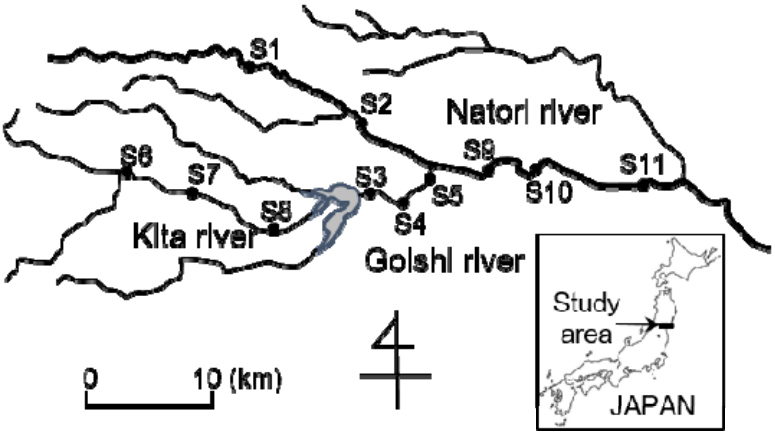
B.L. analysed the data and wrote the manuscript; S.Y. conducted fieldwork, DNA extraction and AFLP experiments; T.M.C. contributed to developing the analytical methods; G.M. and K.W. edited and revised the manuscript. All authors contributed to writing the manuscript.

526 ## Additional Information

527

528 ## Figures

529



530

531

532 **Figure 1.** Map of 11 sampling sites for *Ephemera strigata* in the Natori River Basin in northeastern Japan.

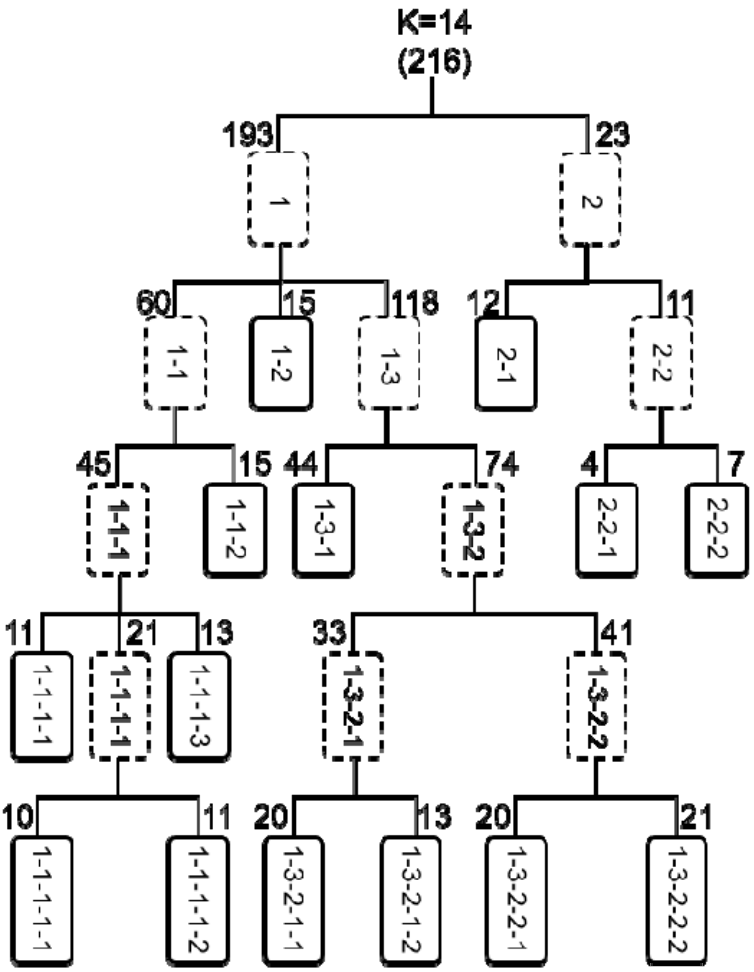
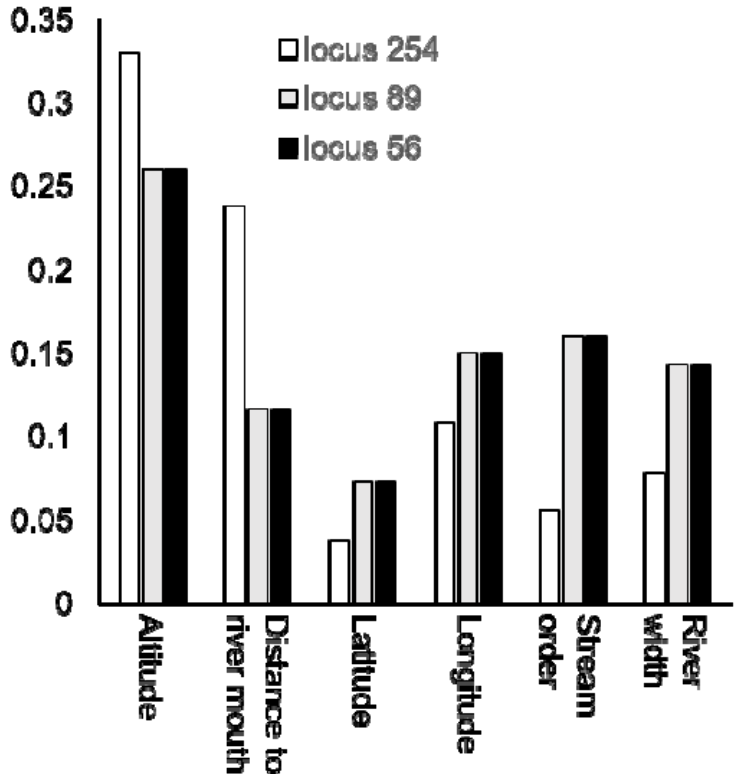


Figure 2. Subpopulation structure of *Ephemera strigata* as determined using STRUCTURE with hierarchical iterations. Dashed boxes indicate subpopulations and solid boxes indicate final populations. Numbers at the top of boxes indicate the number of individuals assigned to the populations. A total of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



538

539 **Figure 3.** Relative importance of environmental variables based on the random forest model for three

540 non-neutral loci (56, 89 and 254).

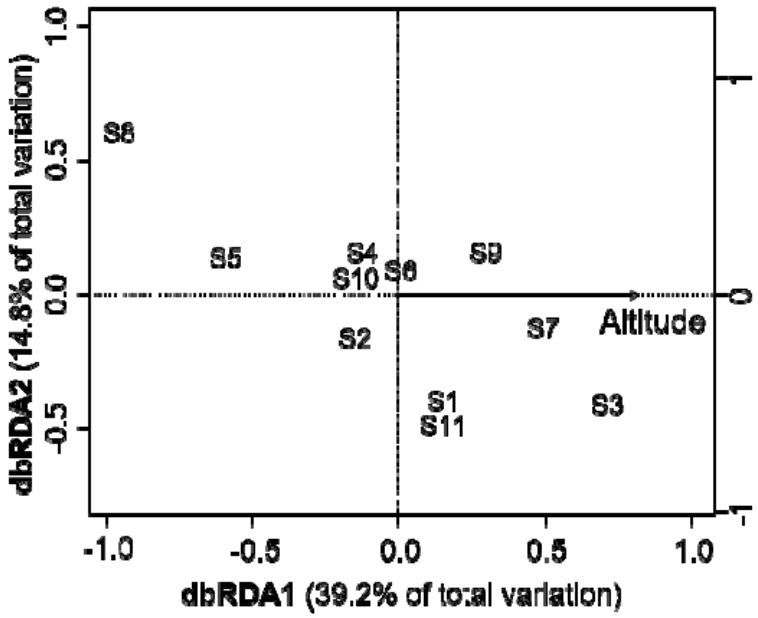
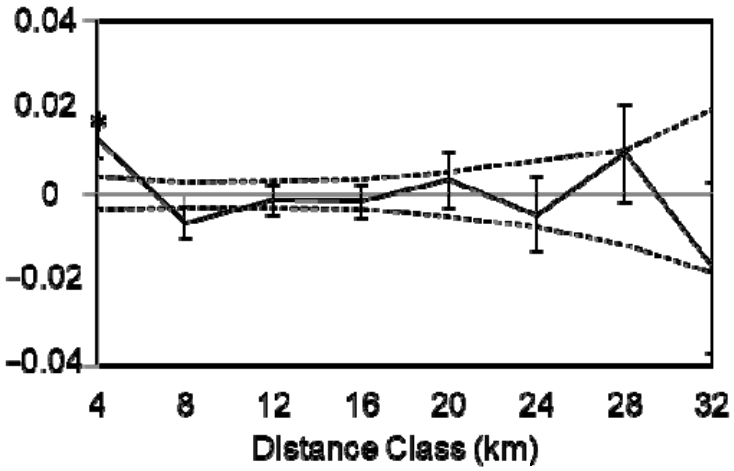


Figure 4. Distance-based redundancy analysis (dbRDA) describing the influence of environmental heterogeneity on genetic variation at a non-neutral locus (254).



545

546 **Figure 5.** Spatial autocorrelation at 4-km distance classes based on geographic distance for neutral loci.
547 Dashed lines indicate permutated 95% confidence intervals and error bars indicate jackknifed 95%
548 confidence intervals. * indicates significant spatial autocorrelation ($p < 0.05$).

Tables

Table 1. Results of outlier loci detection and model construction based on three population definitions and two adaptive divergence models. Out of the 10 non-neutral loci identified from the 14 populations delineated by the hierarchical STRUCTURE analysis, three loci (56, 89 and 254) were modelled by random forest (AUC > 0.7) and one locus (254) was modelled by dbRDA ($p < 0.05$).

Population definition	Number of populations	Non-neutral loci			Neutral loci	Adaptive divergence model	
		Dfdist	BayeScan	Both		Random forest	dbRDA
Sites	11	0	0	0	0	-	-
STRUCTURE	2	0	0	0	0	-	-
Hierarchical STRUCTURE	14	21	26	10	346	3	1

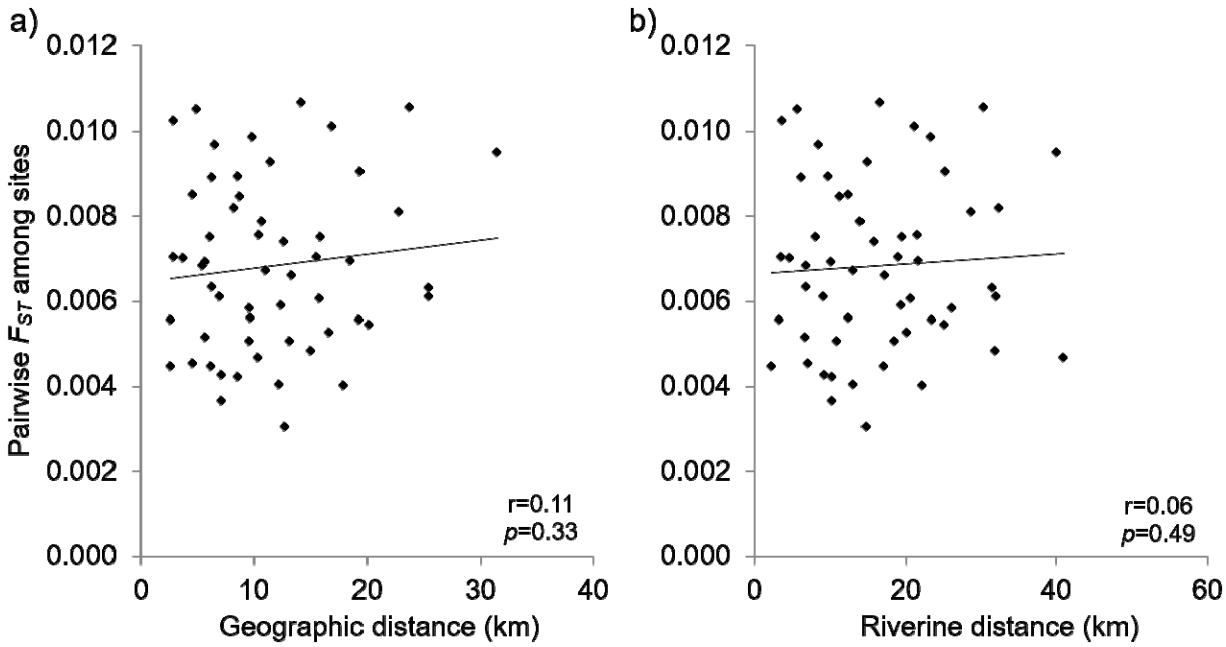
555 **Table 2**. Genetic diversity and divergence measured using the following: 1) all loci, 2) only neutral loci
556 and 3) only non-neutral loci. H_t = total expected heterozygosity; H_w = mean expected heterozygosity
557 within sites; F_{ST} = Wright's fixation index among sites.

	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

558

559 # Supplementary Information

560



561

562 **Figure S1.** Isolation by distance calculated using geographic (a) and riverine (b) distance. Solid lines
563 indicate correlations between Wright's fixation index (F_{ST}) and geographic (r = 0.11, p = 0.33) or riverine
564 distance (r = 0.06, p = 0.49) calculated with Mantel tests.