Title: History or demography? Determining the drivers of genetic variation in North American plants

Short running title: Genetic variation in North American plants

Letters article

Count: Abstract: 154 words. Main text: 4169 words. 71 References. 5 figures and 1 table.

Keywords: Central-Marginal hypothesis, Core-periphery, Microsatellite, Species Distribution Modelling, Postglacial expansion, Range limits

Authors: Julia López-Delgado [1,2*](bsjld@leeds.ac.uk), Patrick G. Meirmans [1] (p.g.meirmans@uva.nl)

Author affiliations:

1 – Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

2 – Institute of Biology, Leiden University, Leiden, The Netherlands

Contact information of corresponding author:

School of Biology, University of Leeds, Manton Building, Leeds LS2 9NH, United Kingdom

Authorship:

PM and JLD designed the study and compiled the database. JLD built the species distribution models, performed the data analysis, and wrote the manuscript, with input from PM.

Data accessibility

The microsatellite data can be accessed from the corresponding peer-reviewed papers listed in Table S.1, and the occurrence data can be downloaded from the per-species DOI found in Table S.1.

Supplemental figures and tables are available in: EcolLett_Supplementary_Figures_López-Delgado_Meirmans.pdf and EcolLett_Supplementary_Tables_López-Delgado_Meirmans.pdf

SDM script and calculations of distance to edge range and distance to suitable area under LGM are available in: EcolLett_SDM_script_López-Delgado_Meirmans.R, which will be uploaded to GitHub.

Abstract

Understanding the impact of historical and demographic processes on genetic variation is essential for devising conservation strategies and predicting responses to climate change. Recolonizations after Pleistocene glaciations and population's positions within species ranges are expected to leave distinct genetic signatures. However, the general applicability of these patterns and relative importance of historical and demographic factors remains unknown. Here, we analysed the distribution of genetic variation in 91 native species of North American plants by coupling microsatellite data and Species Distribution Modelling. We tested the contributions of historical climatic shifts and the central-marginal hypothesis on genetic diversity and structure. Decreased diversity was found with increased distance from potential glacial refugia, coinciding with the expected make-up of postglacially colonised localities. At the range periphery, lower genetic diversity, higher inbreeding levels and genetic differentiation were reported, following the assumptions of the central-marginal hypothesis. History and demography were found to have approximately equal importance in shaping genetic variation.

Introduction

Disentangling the determinants shaping species' distributions and genetic diversity is key to understanding and conserving biodiversity and ultimately predicting responses to ongoing global change. The identification of the spatial distribution of genetic variation can elucidate the mechanisms of evolution and speciation, shed light on the processes maintaining geographical ranges, improve climate change forecasting, anticipate the spread of invasive species, and pinpoint conservation-priority populations (Hampe & Petit 2005; Howes & Lougheed 2008; Guo 2014). Phylogeographic analyses can be employed to study the role of ecological factors and mechanistic processes, such as past climatic shifts and demographic fluctuations, on the genetic structure of populations. Yet, despite a large body of research on the population genetics of individual species, few studies have looked at general patterns of genetic structure and variation across multiple species' ranges at large spatial scales (Gaston 2009).

Historical climate-driven changes in species range limits are well known to still affect present-day genetic diversity (Durka 1999; Hewitt 2000; Hampe & Petit 2005). The Pleistocene ice-ages led to southwards range contractions, brought by the extinction of northern populations when temperatures plummeted, with multiple areas across the continent serving as glacial refugia for species during the ice ages (Beatty & Provan 2011). The end of the Pleistocene led to a subsequent northward expansion of many species in the wake of deglaciation (Hewitt 1996). The shifting and fragmentation of species' ranges in the past 20,000 years have resulted in a legacy of genetic consequences for contemporary populations (Comes & Kadereit 1998; Taberlet *et al.* 1998; Alvarez *et al.* 2009). One major consequence is that previously glaciated areas are expected to show reduced genetic diversity as a result of sequential founder effects during the post-glacial recolonisation from refugia (Comes & Kadereit 1998; Taberlet *et al.* 1998; Schonswetter *et al.* 2005). The effect of the ice-ages was especially strong in temperate areas. In particular, the northern part of North America was

66 covered with two immense ice sheets (Cordilleran and Laurentide) at the Last Glacial Maximum
67 (LGM). The changing conditions following the melting of the glacial ice make North America an ideal
68 natural laboratory to study patterns of post-glacial colonisation and geographic variation (Hewitt
69 2000; Pulgarín-R & Burg 2012)

70 The distribution of genetic variation within a species is expected to also be shaped by demographic
71 and evolutionary processes at range margins. Declining environmental suitability towards the
72 periphery is predicted to result in decreasing population density (Kirkpatrick & Barton 1997), thus
73 reducing population size, gene flow and connectivity. Consequently, marginal populations tend to
74 exhibit low genetic diversity and high genetic differentiation (Brussard 1984; Sagarin & Gaines 2002;
75 Pfeifer *et al.* 2009). However, the broad applicability of this biogeographic paradigm –the central-
76 marginal hypothesis– is debated since patterns of population genetic variation across large spatial
77 scales are highly variable and usually species-specific (Sagarin & Gaines 2002; Yakimowski & Eckert
78 2007). For example, Eckert et al. criticised studies for not including a proper quantitative measure of
79 centrality/peripherality or estimates of population sizes. Additionally, the central-marginal
80 hypothesis assumes concordance between the geographic and environmental spaces, but this
81 assumption might not always hold, since ecological marginality does not always imply spatial
82 peripherality, and vice versa (Soule 1973; Pironon *et al.* 2015).

83 There are relatively few studies that have tested both the central-marginal hypothesis and historical
84 influences in a phylogeographical framework, or have attempted to distinguish historical effects on
85 genetic diversity from patterns caused by contemporary geographical variation in population
86 demography and dispersion (Eckert *et al.* 2008; Gaston 2009). This is problematic since the patterns
87 in genetic diversity resulting from these two processes can resemble each other. If populations at the
88 northern margin show reduced genetic diversity, is this due to founder effects of postglacial
89 recolonization or due to demographic effects related to the central-marginal hypothesis? Moreover,

almost all evidence regarding these fields is highly species-specific, which makes it difficult to draw any general conclusions.

Here, we use a novel phylogeographic framework to test the contributions of both historical climatic shifts and the central-marginal hypothesis on population genetic diversity across the ranges of 91 vascular plant species. We do this by coupling genetic data sourced from the literature and species distribution modelling to analyse the spatial structuring of population genetic variation across the North American continent (Fig. 1). Species distribution models were hereby employed as macroecological tools, used to estimate population suitability, act as surrogates of abundance, and develop proxies for colonisation history and population demography. We perform a continental-scale analysis to identify concordant patterns of population genetic diversity on a large number of unrelated taxa, empirically overcoming the drawbacks commonly associated with demographic and colonisation history studies.

Methods

Genetic Data

We compiled a genetic database consisting of microsatellite data for 91 native diploid species of North American angiosperms and gymnosperms, with each taxon containing at least three sampling sites including a minimum of five individuals per location. We searched the literature for studies employing microsatellite markers on native vascular plants from North America, which resulted in genetic data published in 67 peer-reviewed studies (see Table S1). Two species were represented by two studies. We also obtained four unpublished datasets by direct communication with the authors. For all studies we tried to obtain the raw allelic data by downloading them from data repositories or by contacting the authors. In total, we obtained the raw data for 52 species. The included species

display a wide range of ecological characteristics, abundances, range sizes and life history traits, thus aiming to serve as a representative sample of the continent's vascular flora. The sampling and genotyping protocols vary across species, being detailed in the corresponding publications.

Genetic Summary Statistics

We used the microsatellite data to calculate a set of four summary statistics per population, indicative of population genetic diversity and differentiation: the expected and observed heterozygosity (H_S and H_O respectively), the population inbreeding coefficient (F_{IS}), and beta (β) (Weir & Cockerham 1984; Nei 1987). The expected within-population diversity H_S is calculated only based on the allele frequencies within a population, whereas H_O considers the observed frequency of heterozygotes in the population. F_{IS} can be defined by comparing the above-mentioned heterozygosity measures: (H_S - H_O) / H_O, with high levels indicating an excess of homozygous genotypes compared to Hardy-Weinberg expectations. β is a population-specific estimator of the genetic differentiation statistic F_{ST} [24]. The summary statistics calculations were performed using the function basic.stats() from the 'hierfstat' R package (Goudet 2005), with every sampling location being treated as a separate population. Clone correction was performed for species with apomictic or vegetative reproduction, by removing all duplicated genotypes to ensure each genotype is represented by a single individual (one ramet per genet) per population. For the species for which the allelic data was not available, we used the published estimates of the summary statistics from the original studies, which were mostly limited to H_S, H_O and F_{IS}.

Species Distribution Modelling

Species distribution models were built for all 91 species to evaluate habitat suitability and develop proxies for colonisation history and population demography. The georeferenced occurrences from the sites included in the genetic dataset were complemented by downloading species records from the Global Biodiversity Information Facility (GBIF) data portal (GBIF DOIs in Table S.1). Occurrences were restricted to North America, and duplicate locations were removed to avoid pseudo-replication. All species were represented by at least ten unique presence records, which is the recommended minimum number required to calibrate a species distribution model (SDM) (Proosdij *et al.* 2016). Species occurrences totalled 97,074 unique records, ranging from 10 to 9,594 occurrences per species.

To model the species distributions, the 19 bioclimatic predictors of the WorldClim v.1.4 dataset (http://www.worldclim.org/) were obtained for past and present scenarios at a 2.5 arc-minute spatial resolution (Hijmans & Elith 2012). The used paleoclimate data resulted from simulations from a Global Climate Model (GCM) for the LGM (approximately 22,000 years ago), as estimated by the MIROC-ESM climate model (Watanabe *et al.* 2011). Current conditions represent interpolations of observed climatic data from 1960-1990. The GTOPO30, a global 30 arc-second digital elevation model (DEM) was retrieved from the USGS EROS archive (https://www.usgs.gov/centers/eros). The GTOPO30 was aggregated by a cell factor of five to achieve a 2.5 arc-minute resolution, resulting in a DEM for the present scenario. In order to obtain a DEM for the LGM, when the sea level was 120-135m lower than presently (Clark & Mix 2002), the GEBCO_2019 Grid, a global 15 arc-second bathymetry DEM (GEBCO, 2019) was downloaded. The GEBCO_2019 Grid was rasterised and aggregated by a cell factor of ten to achieve a 2.5 arc-minute resolution, followed by clipping it using an LGM bioclimatic layer as a mask. Employing altitude as a variable for modelling is advised against when the SDM aims to project to past climatic conditions (Raes & Aguirre-Gutiérrez 2018). Thus, slope and aspect, derived from the DEMs, were included as variables instead. All bioclimatic and topographic variables were clipped to the extent of the North American continent, from the

160 southernmost point of Panama to the northernmost point of Canada, excluding Greenland. Data
161 layer manipulations were performed with ArcGIS (ESRI).

162 To avoid collinearity amongst environmental predictors, which can result in overfitting (Graham
163 2003; Peterson *et al.* 2007), the number of predictor variables was reduced by removing highly
164 correlated parameters, as given by a Spearman's rank correlation test ($r_s > 0.7$). When deciding which
165 of two correlated variables to retain, we aimed to closely capture the key determinants of
166 physiological processes limiting distributions of plants, considering the ample range in ecological
167 preferences displayed by the species included. The retained variables for modelling were mean
168 diurnal temperature range (Bio2), temperature annual range (Bio7), mean temperature of wettest
169 quarter (Bio8), annual precipitation (Bio12), precipitation seasonality (Bio15), aspect, and slope.

170 Species distribution models were built for each species under present conditions and then projected
171 onto the LGM conditions, employing the 'sdm' R library (Naimi & Araújo 2016). Three modelling
172 methods were implemented: Domain (Carpenter *et al.* 1993), Generalised Linear Model (GLM;
173 McCullagh & Nelder 1989), and Maximum Entropy (MaxEnt; Phillips *et al.* 2006), each belonging to
174 one of the three main types of modelling algorithms, being 'profile', 'regression', and 'machine
175 learning', respectively. Cross-validation was performed to validate each model, with 70% of the data
176 being employed for calibration and 30% for evaluation, with ten bootstrap replications being run per
177 method. SDM accuracy was evaluated using the area under the curve (AUC) of the receiver operating
178 characteristic (ROC) plot (Hanley & McNeil 1982), a threshold-independent measure that is relatively
179 insensitive to prevalence (McPherson *et al.* 2004). AUC values range from zero to one; values close to
180 one indicate maximum fit, whereas values under 0.5 (half of the area under the ROC curve) indicate
181 the model prediction is no better than a random prediction. Ensembles of model forecasts were
182 fitted by combining the three modelling techniques, albeit models that performed poorly (i.e. AUC <
183 0.5) were not used to build the ensembles. By employing a consensus, errors (sensitivity to data, lack

184 of absence data, errors in environmental variables) tend to cancel each other out in ensembles, thus
185 producing a more robust and conservative solution (Araújo & New 2007; Diniz-Filho *et al.* 2009). The
186 ensembles were built using weighted averaging based on the AUC statistic to cope with model
187 variability and to improve the reliability of model predictions. The R script used for the species
188 distribution modelling is available in the supplement.

189

190 *Ecological data*

191 We used the output of the SDM to calculate for every population in each species four measures that
192 quantify ecological suitability, colonisation history, and population demography: 1) habitat suitability
193 under the current conditions, 2) habitat suitability during the LGM, 3) distance to range edge under
194 current conditions, and 4) distance to potential glacial refugium. The habitat suitability was taken
195 directly from the ensemble forecasting produced as output of the different SDMs (Anderson &
196 Martínez-Meyer 2004; Diniz-Filho *et al.* 2009). Thus, the ecological suitability *S* of each population
197 under the two modelled time frames was defined as the average value of occurrence provided by
198 each model in the ensemble. For estimating population centrality/peripherality, an innovative
199 quantitative measure was developed. For this, the suitability data was transformed into
200 presence/absence data by setting a threshold, using the *Max SSS* approach of Liu et al. [41], which
201 maximises the sum of model sensitivity and specificity. Out of thirteen threshold selection methods,
202 *Max SSS* was found to perform best when only presence data is available [42]. The distance *D* to the
203 species' range edge under present bioclimatic conditions was then computed by calculating the
204 closest distance between each population and the contour of the generated binary presence map.
205 Distances to range edge had negative values if populations were found outside the predicted species
206 range. Finally, distance from a potential glacial refugium was calculated as the closest distance from
207 each population to a suitable area under LGM after creating a binary presence map based on the

ensemble prediction for the LGM. Distance to range edge computations were performed in ArcGIS v10.2 (ESRI) and distance to suitable areas under different scenarios were calculated using the 'geosphere' R package (Hijmans *et al.* 2019), both employing a geodesic method.

Hypothesis testing

Linear mixed effect models were used to test the relative contribution of distance to edge range, distance to a suitable area under the LGM, and present and LGM suitability in shaping the four population genetic parameters. Additional linear mixed effects models were used to investigate the relationship between the distance to edge range and distance to a suitable area under the LGM with present and LGM suitability. The species name was used as a nested random effect in the analyses to account for phylogenetic non-independence. The linear mixed effects models were performed using the 'lme4' package (Bates *et al.* 2013) in R v3.5. Model selection was performed in terms of parsimony (based on AIC) and variance explained. The variance explained was calculated using the methods proposed by Nakagawa & Schielzeth 2013 as implemented in the 'MuMIn' package, which provides the total variance explained by fixed and random effects and allows the calculation of variance explained by each fixed effect (Barton 2011). P-values were calculated using the Satterthwaite (Satterthwaite 1946) approximations, using the 'lmerTest' R package (Kuznetsova *et al.* 2017), standard Bonferroni correction was then applied (Bonferroni 1936).

Results

The database we compiled includes 1,406 populations across 91 vascular plant species, spanning the whole North American continent, except for the arctic regions. For 829 populations, the genetic summary statistics were computed using the raw allelic microsatellite data, whereas for the

231 remaining populations the available published estimates were taken (Table S.2). Overall minimum
232 and maximum estimates were 0.000 and 0.919 for the expected heterozygosity (H_S), 0.000 and 1.000
233 for the observed heterozygosity (H_O), -1.000 and 1.000 for the inbreeding coefficient (F_{IS}), and -
234 0.2681 and 1.000 for the population differentiation statistic β.

235 The species distribution modelling performed on all 91 species had high predictive power for the
236 relationship between the species' distribution and the bioclimatic variables (Table S.1). Of the 2,730
237 species distribution models (three SDMs x 10 bootstrap replications x 91 species), 88% had AUC
238 values above 0.85 and all models had AUC values above 0.6. AUC values of the MaxEnt models were
239 generally higher than those of GLM models, which in turn were higher than those of Domain,
240 pointing towards a higher predictive power of MaxEnt over GLM and Domain for most species. The
241 data proved highly robust, as the identified general patterns remained even when removing the
242 species for which at least one of the three modelling procedures had AUC values lower than 0.8.

243 The analyses combining the genetic summary statistics with the output of the SDM revealed a legacy
244 of past glaciation in the genetic data. The genetic diversity –as measured with H_S– significantly
245 decreased with increasing distance to suitable areas during the LGM (Fig. 2). Furthermore, there was
246 a significant positive relationship between the suitability of populations under LGM conditions and
247 both H_S and H_O (Fig. 3). No significant patterns were found for F_{IS} and β, (Figs. 2 & 3) even though F_{IS}
248 showed a slight decrease with increasing distance to suitable areas under the LGM. Additionally,
249 populations displayed significantly lower ecological suitability when further away from suitable areas
250 under the LGM (Fig. S1).

251 The data also showed clear support for the central-marginal hypothesis. Marginal populations indeed
252 had a lower ecological suitability as evidenced by a significant positive relationship between distance
253 to range edge and suitability under the current climatic conditions (Fig. S2). This was accompanied by
254 a significant increase in genetic diversity (H_S) as distance to the range edge incremented (Fig. 4).

255 Furthermore, the value of H_S showed a significant negative correlation with the present suitability of
256 the populations (Fig. 5). No significant patterns were observed for H_O, β and F_{IS} (Fig. 5).

257 The standardised variance explained in the global models of the genetic parameters totalled 22.1%
258 by the distance to suitable area under the LGM, 27.5% by LGM suitability, 22.6% by the distance to
259 edge range, and 19.1% by present suitability (Table 1). For the expected heterozygosity, the minimal
260 model included both effects of the ice-ages and effects of the central-marginal hypothesis; this
261 minimal model was composed of the distance to edge range, the distance to a suitable area under
262 the LGM, and the present suitability, with fixed and random effects explaining a total variance of
263 79.4%. The observed heterozygosity's minimal model was given by the present suitability, accounting
264 for 88.2% of the data variation. Model selection was not performed for the inbreeding coefficient, as
265 no variables were deemed significant. Finally, the minimal model for genetic differentiation was
266 given by present suitability, explaining 63.4% of the variance. For all models, the reported values for
267 the percentage of variance explained also includes the nested random effect of species name,
268 showing that there are strong differences between species in genetic diversity and differentiation.

269

270 Discussion

271 Here, we used a phylogeographical framework combining species distribution modelling and
272 population genetics to assess historical and demographic influences on contemporary genetic
273 patterns. Our analysis of genetic data of 91 vascular plant species clearly shows the effects of both
274 Pleistocene climatic events and central-marginal processes in spatially structuring genetic variation in
275 North America. Of the more than 1,400 included populations, those located at species' range margins
276 and most distant from glacial refugia were found to have significantly reduced genetic diversity.

The effects of the Pleistocene ice-ages are evident in the decrease in genetic diversity, as measured using H_S, with increased distance from suitable areas under the LGM, i.e. glacial refugia. This reduction in diversity is similar to what is observed in simulation models on the genetic make-up of postglacially colonised regions (Hewitt 1999, 2000; Pironon *et al.* 2015). The observed genetic patterns are most likely due to the larger population sizes and more stable population dynamics at refugial localities, as well as repeated bottlenecks at the advancing (generally northern) range edge during postglacial colonisation (Hewitt 1996; Comes & Kadereit 1998). Our results therefore indicate that the shifting and fragmentation of species' geographical ranges in the past 20,000 years has played an important role in moulding genetic variation across large taxonomic and spatial scales (Parks *et al.* 1994; Sewell *et al.* 1996; Soltis *et al.* 1997; King & Ferris 1998). Nonetheless, we must be cautious when interpreting diversity patterns, as patterns do not solely emerge from simple models of postglacial colonisation (Petit *et al.* 2003).

The effects of central-marginal dynamics are evident in the negative correlation of genetic diversity with both the distance from the range edge and ecological suitability. The central-marginal hypothesis predicts range limits arise because peripheral populations occur in marginal habitats and cannot adapt to conditions beyond the range edges (Haag & Ebert 2004; Kawecki 2008). This demographic instability can induce low effective population sizes and frequent bottlenecks, leading to reduced genetic diversity in margins, such as observed in our data (Sagarin & Gaines 2002; Sexton *et al.* 2009; Micheletti & Storfer 2015). Although we do not have population size estimates, we assessed the ecological suitability of populations within their ranges. Diniz-Filho *et al.* 2009 propose employing this variable as a macroecological surrogate of abundance, which ensures the logical application of the central-marginal model to complex spatial abundance patterns as it considers species' ecological, and not geographical, ranges. The applicability of this approach in our data was supported by a clear central-peripheral pattern that was observed in the ecological suitability for the analysed populations as this was negatively correlated with the distance to the range margins.

Though we found strong patterns in genetic diversity (H_S and to some extent H_O) related to history and demography proxies, no significant associations were found for the inbreeding coefficient F_{IS}; the minimum model in our Linear Mixed Model analysis of F_{IS} did not include any of the explanatory variables. Theoretical models have suggested that self-fertilisation may be favoured at the range margins, especially at postglacially expanding range fronts, given the possible advantages of reproductive assurance or local adaptation (Arnaud-Haond *et al.* 2006; Hargreaves & Eckert 2014). Indeed, several studies on single species have shown selfing to be more frequent in marginal populations, due to low population sizes and environmental stress at range margins, leading to elevated inbreeding levels (Schoen Daniel J. *et al.* 1996; Aldrich & Hamrick 1998; Barrett 2002). Our results across a taxonomically wide range of species suggests that this may not be a general phenomenon.

Similar to F_{IS}, no significant relationships were found for the genetic differentiation (β, a single population estimate of F_{ST}) of populations with the rest of the metapopulation, though the minimum model according to AIC did include the present suitability as an explanatory variable. This lack of a significant relationship contrasts with numerous theoretical models and empirical studies reporting a decrease in genetic differentiation along postglacial expansion routes (Austerlitz *et al.* 1997; Excoffier *et al.* 2009). On the other hand, marginal populations are expected to be more isolated from the central ones, so it is possible that the effects of central-marginal processes and the Pleistocene recolonisation are counteracting each other. It is also possible that major population structure lies elsewhere, such as among different genetic clusters associated with recolonisation from multiple isolated refugia (Ursenbacher *et al.* 2015).

We employed a species distribution modelling approach to develop effective proxies for colonisation history and population demography. This is a major step up from many previous studies that have worked with the simple assumption that latitude is an adequate surrogate for either recolonisation

history or central-marginal processes (Eckert *et al.* 2008); instead, our SDM-based approach allowed us to test and separate the effects of both these phenomena. It is noteworthy that SDMs do not account for biotic interactions, potential for rapid adaptations or time lag. It is also important to realise that both the SDM and the modelled climate data for the LGM represent extreme extrapolations, and therefore can only be taken as a rough approximation of the situation during the ice-ages. For a better overview of the locations of glacial refugia, SDM results should be combined with palynological data, though this has also been proven to be challenging, due to the coarseness of the palynological record (Birks 2019). Nonetheless, our results show that the models, which had high power for predicting the present distribution, can be used as a valuable tool to infer causation of extant genetic variation patterns across a large scale.

Both historical and demographical processes appear to have equal and non-exclusive importance in shaping genetic variation. Most genetic summary statistics displayed the expected responses to these proxies; however, only a quarter of the comparisons were significant and the variation explained by the fixed effect factors was relatively low. Most variation in the summary statistics comes from within-species determinants, as a large proportion of the variance was explained by the species random effects. This was not unexpected given the widely different demographic histories of the species studied, and it highlights the role of potential additional variables acting at an intraspecific level. Numerous studies have emphasized how the genetics of populations is shaped by ecological factors, anthropogenic factors –such as land use and habitat fragmentation–, biotic interactions, and life-history traits –including the aforementioned breeding strategy and dispersal capacities (Kuittinen *et al.* 1997; Yeaman & Jarvis 2006; Alvarez *et al.* 2009; Meirmans *et al.* 2011). Additionally, each study included in our meta-analysis used different genotyping and sampling strategies and covered varying proportions of the range, as well as employing different molecular markers, characterised by their own mutation rates. Despite the large influence of this random factor

on our analysis, we were able to detect the effects of two vital processes on species ranges and their intraspecific genetic variation.

Bridging the effects of historical range shifts and contemporary demographic processes on species' genetic constitutions over large scales has many implications to evolutionary biology and addresses conservation needs in the face of ongoing environmental change. This study provides the foundation for further phylogeographical analyses to continue integrating the processes dictating genetic variation across species ranges.

References

Aldrich, P.R. & Hamrick, J.L. (1998). Reproductive Dominance of Pasture Trees in a Fragmented Tropical Forest Mosaic. *Science*, 281, 103–105.

Alvarez, N., Thiel-Egenter, C., Tribsch, A., Holderegger, R., Manel, S., Schönswetter, P., *et al.* (2009). History or ecology? Substrate type as a major driver of spatial genetic structure in Alpine plants. *Ecol. Lett.*, 12, 632–640.

Anderson, R.P. & Martínez-Meyer, E. (2004). Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biol. Conserv.*, 116, 167–179.

Araújo, M.B. & New, M. (2007). Ensemble forecasting of species distributions. *Trends Ecol. Evol.*, 22, 42–47.

Arnaud-Haond, S., Teixeira, S., Massa, S.I., Billot, C., Saenger, P., Coupland, G., *et al.* (2006). Genetic structure at range edge: low diversity and high inbreeding in Southeast Asian mangrove (*Avicennia marina*) populations. *Mol. Ecol.*, 15, 3515–3525.

Austerlitz, F., Jung-Muller, B., Godelle, B. & Gouyon, P.-H. (1997). Evolution of Coalescence Times, Genetic Diversity and Structure during Colonization. *Theor. Popul. Biol.*, 51, 148–164.

Barrett, S.C.H. (2002). The evolution of plant sexual diversity. *Nat. Rev. Genet.*, 3, 274.

Barton, K. (2011). *Model selection and model averaging based on information criteria (AICc and alike).*

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., *et al.* (2013). *lme4: Linear Mixed-Effects Models using "Eigen" and S4*.

Beatty, G.E. & Provan, J. (2011). Phylogeographic analysis of North American populations of the parasitic herbaceous plant *Monotropa hypopitys* L. reveals a complex history of range expansion from multiple late glacial refugia. *J. Biogeogr.*, 38, 1585–1599.

Birks, H.J.B. (2019). Contributions of Quaternary botany to modern ecology and biogeography. *Plant Ecol. Divers.*, 12, 189–385.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni R Ist. Super. Sci. Econ. E Commericiali Firenze*, 8, 3–62.

Brussard, P.F. (1984). Geographic Patterns and Environmental Gradients: The Central-Marginal Model in *Drosophila* Revisited. *Annu. Rev. Ecol. Syst.*, 15, 25–64.

Carpenter, G., Gillison, A.N. & Winter, J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodivers. Conserv.*, 2, 667–680.

391 Clark, P.U. & Mix, A.C. (2002). Ice sheets and sea level of the Last Glacial Maximum. *Quat. Sci. Rev.*,
392 EPILOG, 21, 1–7.
393 Comes, H.P. & Kadereit, J.W. (1998). The effect of Quaternary climatic changes on plant distribution
394 and evolution. *Trends Plant Sci.*, 3, 432–438.
395 Diniz-Filho, J.A.F., Nabout, J.C., Bini, L.M., Soares, T.N., de Campos Telles, M.P., de Marco, P., *et al.*
396 (2009). Niche modelling and landscape genetics of *Caryocar brasiliense* ("Pequi" tree:
397 Caryocaraceae) in Brazilian Cerrado: an integrative approach for evaluating central–
398 peripheral population patterns. *Tree Genet. Genomes*, 5, 617–627.
399 Durka, W. (1999). Genetic diversity in peripheral and subcentral populations of *Corrigiola litoralis* L.
400 (Illecebraceae). *Heredity*, 83, 476–484.
401 Eckert, C.G., Samis, K.E. & Lougheed, S.C. (2008). Genetic variation across species' geographical
402 ranges: the central–marginal hypothesis and beyond. *Mol. Ecol.*, 17, 1170–1188.
403 Excoffier, L., Foll, M. & Petit, R.J. (2009). Genetic Consequences of Range Expansions. *Annu. Rev. Ecol.*
404 *Evol. Syst.*, 40, 481–501.
405 Gaston K.J. (2009). Geographic range limits of species. *Proc. R. Soc. B Biol. Sci.*, 276, 1391–1393.
406 Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol. Ecol.*
407 *Notes*, 5, 184–186.
408 Graham, M.H. (2003). Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, 84,
409 2809–2815.
410 Guo, Q. (2014). Central-marginal population dynamics in species invasions. *Front. Ecol. Evol.*, 2.
411 Haag, C.R. & Ebert, D. (2004). A new hypothesis to explain geographic parthenogenesis. *Ann. Zool.*
412 *Fenn.*, 41, 539–544.
413 Hampe, A. & Petit, R.J. (2005). Conserving biodiversity under climate change: the rear edge matters:
414 Rear edges and climate change. *Ecol. Lett.*, 8, 461–467.
415 Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating
416 characteristic (ROC) curve. *Radiology*, 143, 29–36.
417 Hargreaves, A.L. & Eckert, C.G. (2014). Evolution of dispersal and mating systems along geographic
418 gradients: implications for shifting ranges. *Funct. Ecol.*, 28, 5–21.
419 Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405, 907–913.
420 Hewitt, G.M. (1996). Some genetic consequences of ice ages, and their role in divergence and
421 speciation. *Biol. J. Linn. Soc.*, 58, 247–276.
422 Hewitt, G.M. (1999). Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.*, 68, 87–112.
423 Hijmans, R.J. & Elith, J. (2012). *Species Distribution Modelling with R*.

Hijmans, R.J., Williams, E. & Vennes, C. (2019). *geosphere: Spherical Trigonometry*.

Howes, B.J. & Lougheed, S.C. (2008). Genetic diversity across the range of a temperate lizard. *J. Biogeogr.*, 35, 1269–1278.

Kawecki, T.J. (2008). Adaptation to Marginal Habitats. *Annu. Rev. Ecol. Evol. Syst.*, 39, 321–342.

King, R.A. & Ferris, C. (1998). Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. *Mol. Ecol.*, 7, 1151–1161.

Kirkpatrick, M. & Barton, N.H. (1997). Evolution of a Species' Range. *Am. Nat.*, 150, 1–23.

Kuittinen, H., Mattila, A. & Savolainen, O. (1997). Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis thaliana*. *Heredity*, 79, 144.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.*, 82.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28, 385–393.

Liu, C., White, M. & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *J. Biogeogr.*, 40, 778–789.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Chapman and Hall, Boca Raton; London; New York.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.*, 41, 811–823.

Meirmans, P.G., Goudet, J., IntraBioDiv Consortium & Gaggiotti, O.E. (2011). Ecology and life history affect different aspects of the population structure of 27 high-alpine plants. *Mol. Ecol.*, 20, 3144–3155.

Micheletti, S.J. & Storfer, A. (2015). A test of the central-marginal hypothesis using population genetics and ecological niche modelling in an endemic salamander (*Ambystoma barbouri*). *Mol. Ecol.*, 24, 967–979.

Naimi, B. & Araújo, M.B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, 39, 368–375.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.*, 133–142.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.

455 Parks, C.R., Wendel, J.F., Sewell, M.M. & Qiu, Y.-L. (1994). The significance of allozyme variation and
456 introgression in the *Liriodendron tulipifera* complex (*Magnoliaceae*). *Am. J. Bot.*, 81, 878–
457 889.
458 Peterson, A.T., Papeş, M. & Eaton, M. (2007). Transferability and model evaluation in ecological niche
459 modeling: a comparison of GARP and Maxent. *Ecography*, 30, 550–560.
460 Petit, R.J., Aguinagalde, I., Beaulieu, J.-L. de, Bittkau, C., Brewer, S., Cheddadi, R., *et al.* (2003). Glacial
461 Refugia: Hotspots But Not Melting Pots of Genetic Diversity. *Science*, 300, 1563–1565.
462 Pfeifer, M., Schatz, B., Xavier Pico, F., Passalacqua, N.G., Fay, M.F., Carey, P.D., *et al.* (2009).
463 Phylogeography and genetic structure of the orchid *Himantoglossum hircinum* (L.) Spreng.
464 across its European central–marginal gradient. *J. Biogeogr.*, 36, 2353–2365.
465 Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006). Maximum entropy modeling of species
466 geographic distributions. *Ecol. Model.*, 190, 231–259.
467 Pironon, S., Villellas, J., Morris, W.F., Doak, D.F. & García, M.B. (2015). Do geographic, climatic or
468 historical ranges differentiate the performance of central versus peripheral populations?
469 *Glob. Ecol. Biogeogr.*, 24, 611–620.
470 Proosdij, A.S.J. van, Sosef, M.S.M., Wieringa, J.J. & Raes, N. (2016). Minimum required number of
471 specimen records to develop accurate species distribution models. *Ecography*, 39, 542–552.
472 Pulgarín-R, P.C. & Burg, T.M. (2012). Genetic Signals of Demographic Expansion in Downy
473 Woodpecker (*Picoides pubescens*) after the Last North American Glacial Maximum. *PLoS*
474 *ONE*, 7, e40412.
475 Raes, N. & Aguirre-Gutiérrez, J. (2018). A Modeling Framework to Estimate and Project Species
476 Distributions in Space and Time. In: *Mountains, Climate and Biodiversity*.
477 Sagarin, R.D. & Gaines, S.D. (2002). The 'abundant centre' distribution: to what extent is it a
478 biogeographical rule? *Ecol. Lett.*, 5, 137–147.
479 Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biom.*
480 *Bull.*, 2, 110–114.
481 Schoen Daniel J., Morgan Martin T. & Bataillon Thomas. (1996). How does self-pollination evolve?
482 Inferences from floral ecology and molecular genetic variation. *Philos. Trans. R. Soc. Lond. B.*
483 *Biol. Sci.*, 351, 1281–1290.
484 Schonswetter, P., Stehlik, I., Holderegger, R. & Tribsch, A. (2005). Molecular evidence for glacial
485 refugia of mountain plants in the European Alps. *Mol. Ecol.*, 14, 3547–3555.
486 Sewell, M.M., Parks, C.R. & Chase, M.W. (1996). Intraspecific Chloroplast DNA Variation and
487 Biogeography of North American *Liriodendron* L. (*Magnoliaceae*). *Evolution*, 50, 1147–1154.

Sexton, J.P., McIntyre, P.J., Angert, A.L. & Rice, K.J. (2009). Evolution and Ecology of Species Range Limits. *Annu. Rev. Ecol. Evol. Syst.*, 40, 415–436.

Soltis, D.E., Gitzendanner, M.A., Strenge, D.D. & Soltis, P.S. (1997). Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Syst. Evol.*, 206, 353–373.

Soule, M. (1973). The Epistasis Cycle: A Theory of Marginal Populations. *Annu. Rev. Ecol. Syst.*, 4, 165–187.

Taberlet, P., Fumagalli, L., Wust-Saucy, A. & Cosson, J. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.*, 7, 453–464.

Ursenbacher, S., Guillon, M., Cubizolle, H., Dupoué, A., Blouin-Demers, G. & Lourdais, O. (2015). Postglacial recolonization in a cold climate specialist in western Europe: patterns of genetic diversity in the adder (*Vipera berus*) support the central-marginal hypothesis. *Mol. Ecol.*, 24, 3639–3651.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., *et al.* (2011). MIROC-ESM: model description and basic results of CMIP5-20c3m experiments.

Weir, B.S. & Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38, 1358–1370.

Yakimowski, S.B. & Eckert, C.G. (2007). Threatened Peripheral Populations in Context: Geographical Variation in Population Frequency and Size and Sexual Reproduction in a Clonal Woody Shrub. *Conserv. Biol.*, 21, 811–822.

Yeaman Sam & Jarvis Andy. (2006). Regional heterogeneity and gene flow maintain variance in a quantitative trait within populations of lodgepole pine. *Proc. R. Soc. B Biol. Sci.*, 273, 1587–1593.

Acknowledgements

We thank the authors who shared the microsatellite raw data, as well as Maarten van der Sande and Yorick Coolen for assisting in the compilation of the database. JLD also thanks Niels Raes and Emiel van Loon for advice on building the SDMs.

Competing interests

The authors declare no competing interests.

Figures

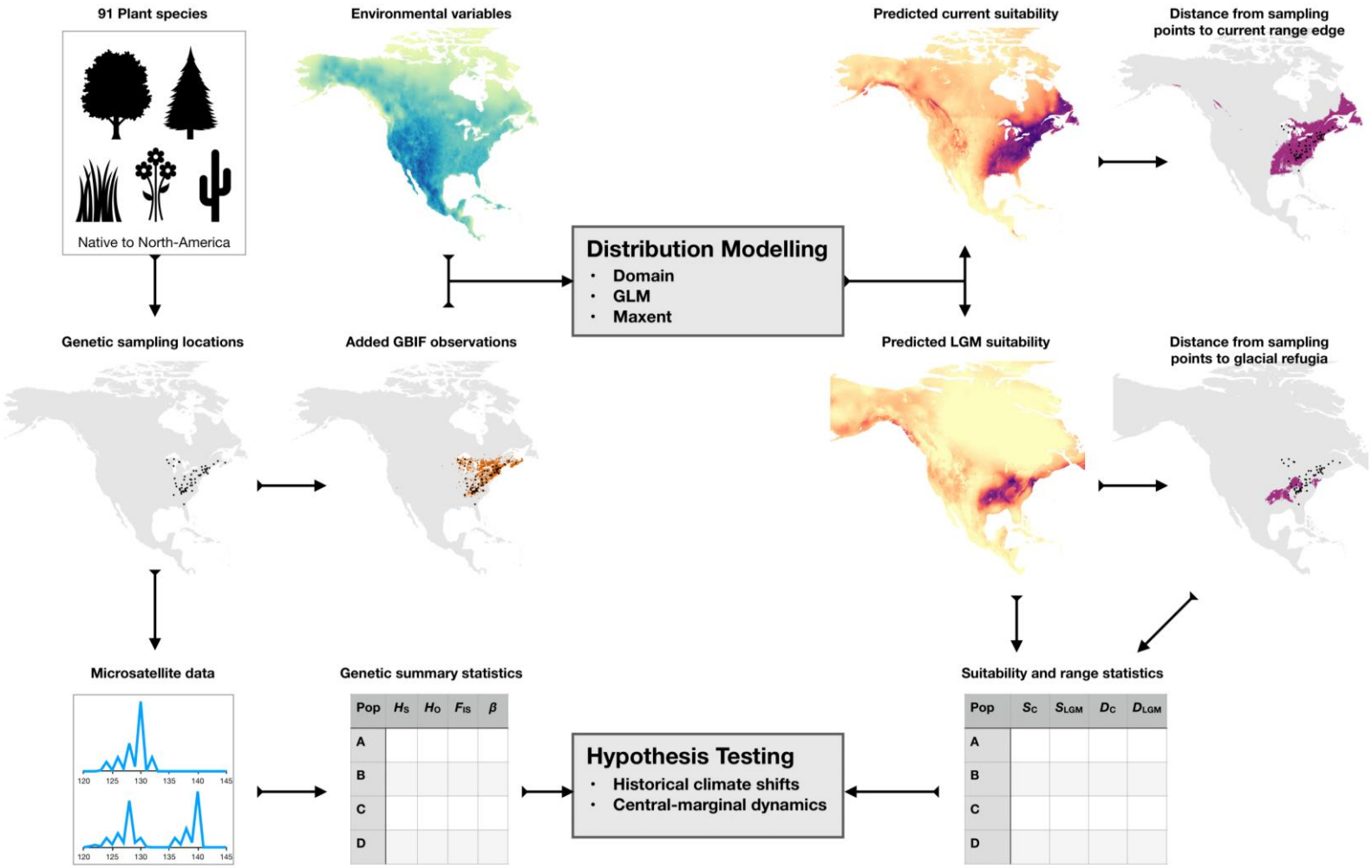
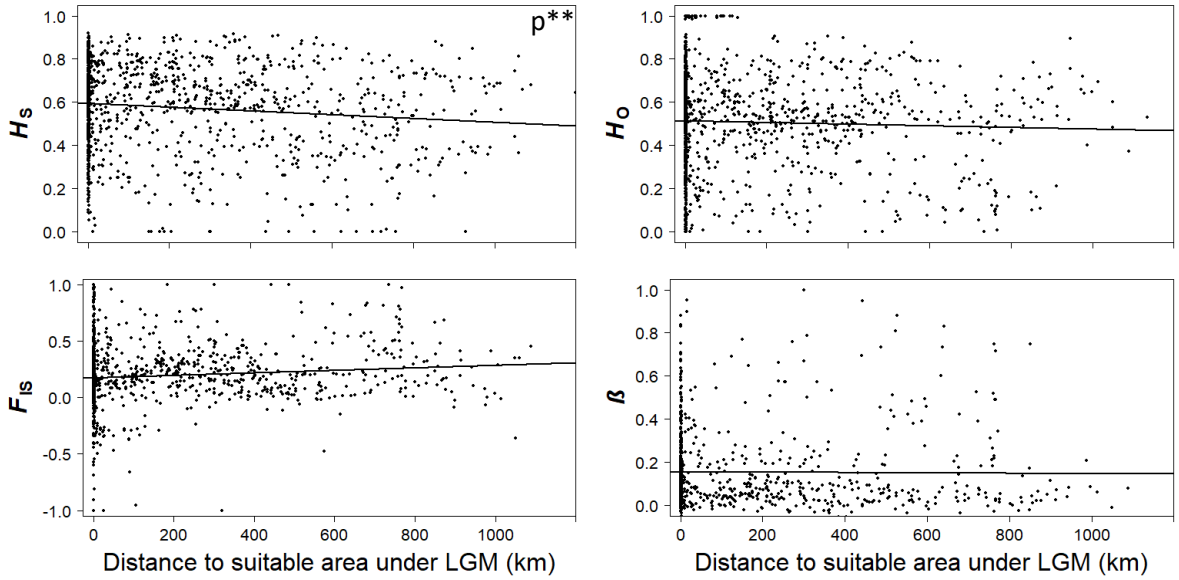


Figure 1: Overview of the methodological approach, which couples genetic data and species distribution modelling to test the contributions of historical climatic shifts and the central-marginal hypothesis on the spatial distribution of genetic variation. The shown distribution data and modelling output are for *Tsuga canadensis* (see Tables S.1 and S.2).

526

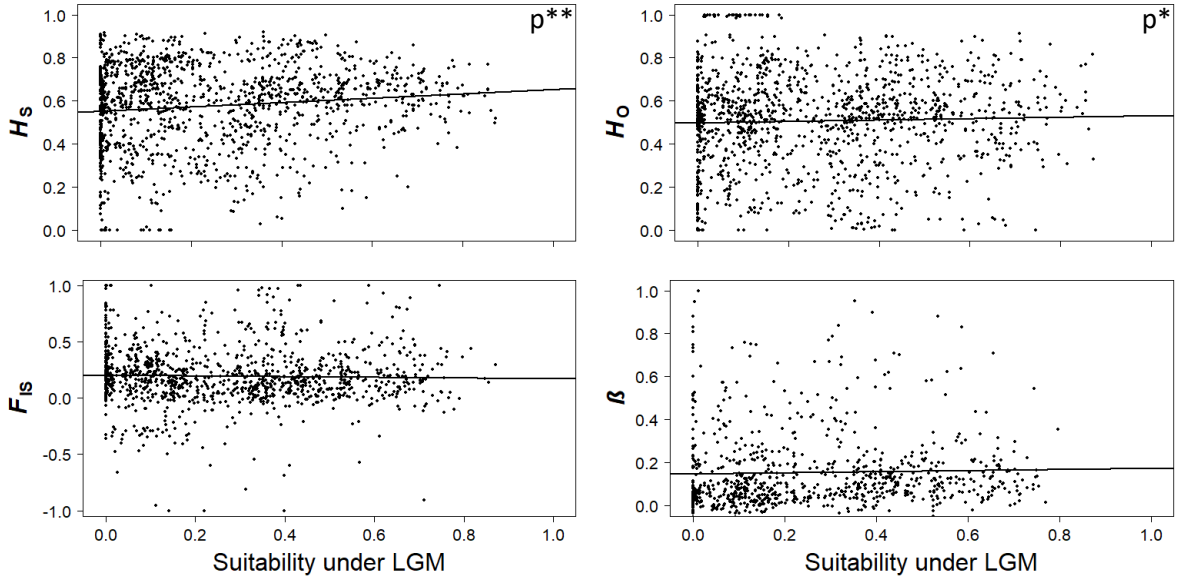


527

528 Figure 2: Relationship between the distance from populations to a suitable area under LGM
529 conditions and the four genetic parameters.

530

531



532

533 Figure 3: Relationship between the suitability of populations under LGM conditions and the four
534 genetic parameters.

535

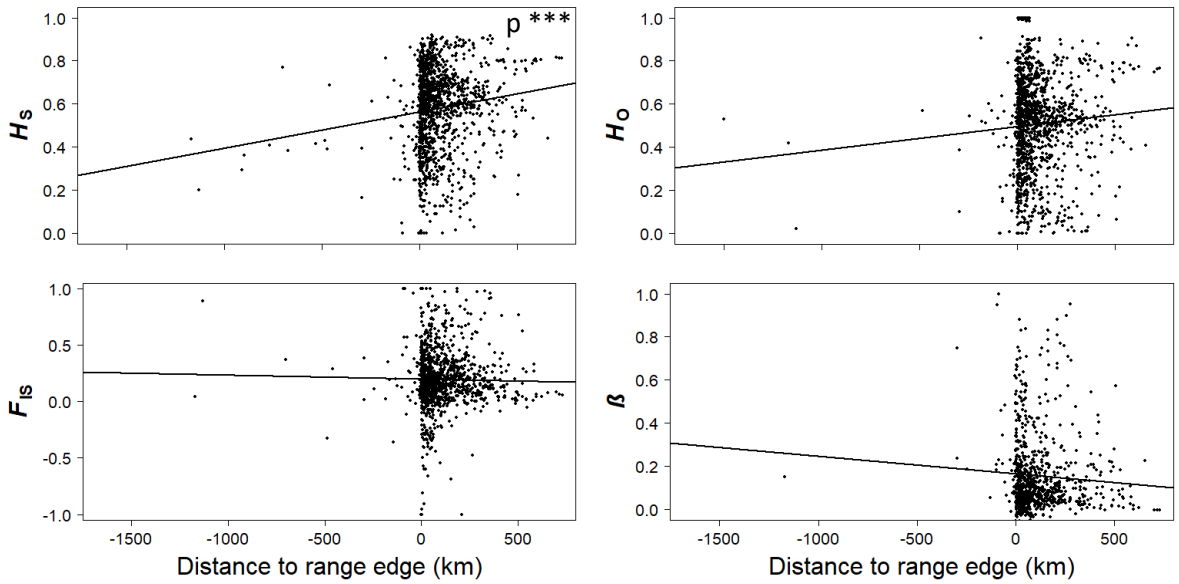


Figure 4: Relationship between the distance from populations to the range edge and the four genetic parameters.

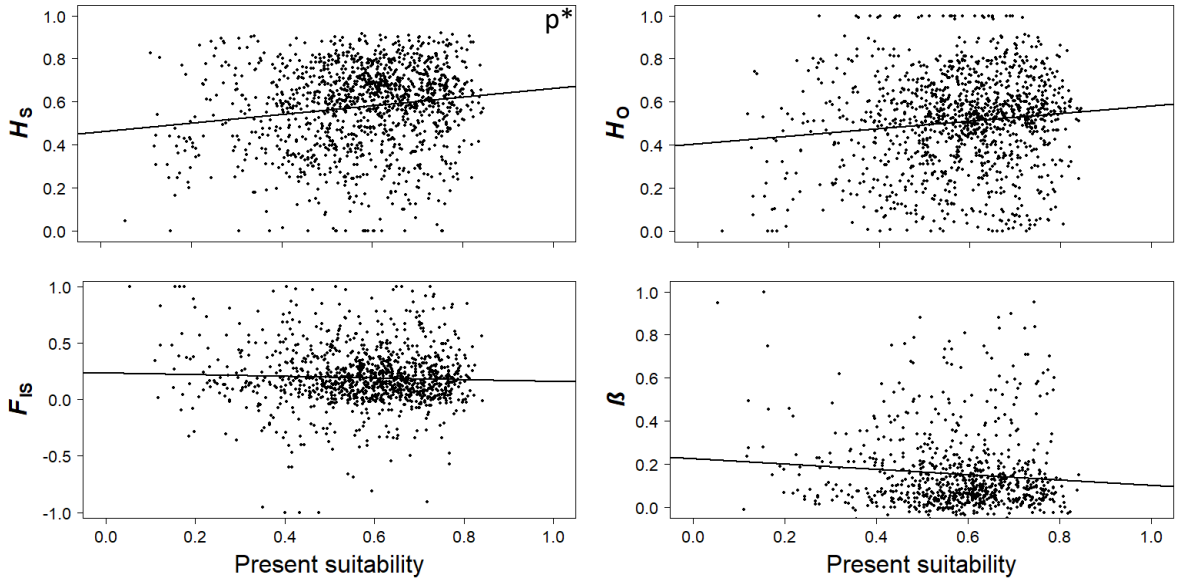


Figure 5: Relationship between population present ecological suitability and the four genetic parameters.

Tables

Table 1: Output of Linear Mixed Model analysis: degrees of freedom (d.f.), percentages of variance explained (Variance) and P-values are given for the global models of each genetic summary statistic including the distance to suitable area under the LGM, the suitability under the LGM, the distance to range edge, and present suitability. The total variance explained by the global and minimal models of each genetic summary statistic are reported in the bottom rows. The total variance explained by the historical and demographic proxies is given in the rightmost column, calculated as the sum of the standardised variances explained for the four genetic summary statistics. For each summary statistic, the variables included in the minimum model, according to AIC, are given in black; variables excluded from the minimum model are given in grey.

		H_S	H_O	F_{IS}	β	**Total variance of parameters (%)**
Distance to suitable area under LGM	d.f.	1310	1198	1110	822	22.1
	Variance (%)	18	1	1.1	2	
	P-value	0.0026	0.3690	1.0000	1.0000	
Suitability under LGM	d.f.	1287	1185	1090	817	27.5
	Variance (%)	10.8	2.2	0	14.5	
	P-value	0.0030	0.0464	0.8088	0.3846	
Distance to range edge	d.f.	1323	1217	1126	814	22.6
	Variance (%)	14.8	2.1	1.9	3.8	
	P-value	0.0007	1.0000	1.0000	1.0000	
Present suitability	d.f.	1327	1209	1114	819	19.1
	Variance (%)	1.3	4.5	3.6	9.7	
	P-value	0.0149	0.3472	1.0000	0.1691	
Total variance of global model (%)		79.4	88.2	85.3	62.2	
Total variance of minimal model (%)		79.4	88.2	-	63.4	