Environmental association modelling with loci under divergent selection accurately predicts the distribution range of a lizard

Alejandro Llanos-Garrido[1], Andrea Briega-Álvarez[2], Javier Pérez-Tris[3], and José Díaz[3]

[1]Harvard University
[2]Museum für Naturkunde - Leibniz-Institut für Evolutions- und Biodiversitätsforschung
[3]UCM

June 17, 2020

Abstract

During geographical expansion of a species individual colonizers have to confront different ecological challenges, and the capacity of the species to broaden its range may depend on the total amount of adaptive genetic variation supplied by evolution. We set out to test whether the distribution of loci under selection along a contrasting environmental gradient can be turned into a model that accurately predicts a species' range. If positive, this may shed light on the genetic source of adaptive limits that shape range boundaries. We sampled five populations of the western Mediterranean lizard Psammodromus algirus that inhabit a noticeable environmental gradient of temperature and precipitation. We used 21 SNPs putatively under selection to correlate the genotypes of 95 individuals with environmental variation among their populations, using 1x1 km2 grid cells as sampling units. By extrapolating the resulting model to all possible combinations of alleles, we inferred the locations that were theoretically suitable for the species. The inferred distribution range overlapped to a large extent with the realized range of the species, including an accurate prediction of internal gaps and range borders. Our results suggest an adaptability threshold determined by the amount of genetic variation available that would be required to warrant adaptation beyond a certain limit of environmental variation. These results support the idea that the expansion of a species' range may be ultimately linked to the arising of new variants under selection.

Introduction

Aside from limits set by dispersal barriers, distribution range borders are commonly assumed to be the result of the constraints imposed by the ecological requirements of species, as environmental gradients change towards suboptimal conditions near range edges (Hutchinson 1957; Brown 2002). All in all, the factors that shape these distribution borders not due to dispersal barriers are ultimately linked to local adaptation dynamics; simply put, a species does not occur outside its distribution border because it is not adapted to the environmental conditions beyond it (Kirkpatrick and Barton 1997; Bridle and Vines 2007). However, the edge of a species' range is typically more abrupt than expected, given that environmental change towards suboptimal conditions or niche boundaries is usually gradual (Sexton et al. 2009). Moreover, all across their ranges species meet a range of conditions that is much greater than the gradient that takes place at the edge of the range (Kirkpatrick and Barton 1997). To understand these seemingly arbitrary boundaries to range expansion, Haldane (1956) proposed gene 'swamping' as a center-border effect by which gene flow from central to marginal habitats causes maladaptation at the edges of the range, reducing population density and constraining range expansion. This dynamic pattern would jeopardize adaptation at the edge of the range even if the genetic variants that could promote range expansion are present in the genetic pool of a species, because gene swamping would hamper a rise in the frequencies of adaptive alleles at range limits (Haldane 1956). However, this hypothesis has been subjected to continuous debate (Nosil and Crespi 2004;

Sexton et al. 2011; Polechová 2018).

Another possibility is that range limits arise because a species has fully colonized the spatial projection of its ecological niche, in such way that niche expansion must precede range enlargement (Hutchinson 1957). In such cases, since niche expansion implies adaptation to more extreme conditions along one or more environmental gradients, this process is limited by the magnitude of the additive genetic variance associated with adaptation to these gradients (Lande and Shannon 2006). Conversely, if habitat suitability remains high at and beyond range boundaries, then dispersal constraints, gene swamping and/or marginal demographic effects could be defining the location and shape of distribution limits (Kirkpatrick and Barton 1997; Bridle and Vines 2007; Charlesworth 2009; Peterson 2011). But if the edges of the range are tightly linked to the exhaustion of relevant genetic variance, then gene swamping cannot properly explain those limits, because it cannot operate beyond the limits imposed by additive genetic variance at relevant adaptive loci. In other words, if the genetic variability required for range expasion is not available in the genetic pool of a species, gene swamping cannot be invoked to explain range limits (Polechova and Barton 2015, Polechova 2018). Evaluating these different hypotheses is thus crucial to understanding the adaptive causes underlying the formation and shaping of range edges (Sexton et al. 2009; Lee-Yaw et al. 2018).

Landscape genomics approaches have boosted our understanding of how environmental variables drive the genetic dynamics of local adaptation (Hoban et al. 2016; Ahrens et al. 2018). These methods can be applied to model (and predict) potential range boundaries by looking at the shifts in allelic frequencies along environmental gradients (Eckert et al. 2008; Herrera and Bazaga 2008). Thus, it is possible to explore what loci govern the adaptability of a species, and to model the suitability of certain genotypes to different habitats all over a species' range (i. e, environmental association analyses; Rellstab et al. 2015; Whitlock and Lotterhos 2015). However, describing correlations between genotypes and environmental gradients is only one part of the challenge, because some loci could show strong but spurious associations with environmental gradients due to population history rather than natural selection. Thus, it is also paramount to identify the loci underpinning local adaptation, since they make the fraction of genetic variation that is relevant to explain an individual's ability to disperse to, and thrive in, new habitats (Dudaniec et al. 2018). Identifying these combinations of loci under selection is a prerequisite to understanding the adaptive basis of the origin and maintenance of new populations and, therefore, the genetic dynamics that shape range boundaries (Hargreaves et al. 2014). Yet, new populations of a species can be established either by 1) the arrival of individuals carrying genetic adaptations to that new site, or 2) the arrival of genetic variants that can recombine in situ to generate new locally adapted genotypes (Barton and Etheridge 2018). Discerning between these two possibilities is a hard challenge. In particular, the last scenario is controversial because it assumes that an individual is able to reproduce in a location to which it is not adapted. However, the potential to produce new genetic combinations increases with dispersal rate, by rising the probability that different (suboptimal) genotypes eventually co-occur at the same new habitats (Barton and Etheridge 2018; LaRue et al. 2018). Thus, it is important to consider dispersal ability in landscape genomic studies, and to compare systems with different dispersal rates. In particular, study organisms with low dispersal rates should reduce the confounding effects of dispersion, allowing us to focus on local adaptation dynamics as responsible of expansion constraints (Lee-Yaw et al. 2018).

In this study, we integrate genomic data into the distribution modelling of a lacertid lizard species, the Large Psammodromus *Psammodromus algirus* , whose phylogeographical and ecological differentiation is well characterized (Díaz et al. 2017, Llanos-Garrido et al. 2019). This lizard is widespread across the Western Mediterranean region, and its range encompasses contrasting environmental conditions, extending from northern Africa in the south to southwest France in the north, and from Portugal in the west to Tunisia in the east (Fig. 1). We used 21 loci putatively under selection (hereafter outliers; Llanos-Garrido et al. 2019) to model distribution boundaries on the basis of five closely located central populations that cover a representative fraction of the environmental variation faced by *P. algirus* across its entire distribution range. To this end, we run an environmental association analysis (Rellstab et al. 2015) with allelic variants at loci under selection as predictors, and we extrapolated, for all possible allelic combinations at those loci, the geographical locations with suitable environmental conditions. By doing so, we were able to infer not

only an ecological niche model of the whole distribution range of the species, but also the genotypes that would potentially be adapted to each geographical grid cell within it. We assumed a simple model without center-border biases, and in which every genotype is able to reach every geographic cell. Also, we only used the (adaptive) genetic variation likely to be associated with environmental conditions, in such way that we could know whether actual range limits are linked to adaptability thresholds determined by the amount of additive genetic variance available for selection. This approach allowed us to test whether a species' distribution range can be explained by the genetic dynamics that shape local adaptation, without invoking demographic processes such as gene swamping or increased homozigosity near the edge of the range (Herrera and Bazaga 2008; Polechová et al. 2009).

Specifically, we aimed to answer the following questions: 1) Is it possible to infer an entire distribution range on the basis of a limited number of outlier loci? 2) How important are limitations to dispersal in defining a species' range limits? 3) Are there fewer genotypes adapted to marginal conditions than to core conditions? And 4) is there an adaptability threshold, determined by the availability of genetic variance under selection, that constrains the expansion of the range beyond its actual boundaries?

Materials and Methods

Study system

Psammodromus algirus is a ground-dwelling, heliothermic lizard from the Western Mediterranean region whose distribution range encompasses a wide variety of habitats, from arid shrublands to temperate forests (Diaz and Carrascal 1991). In the Iberian Peninsula, where *P. algirus* is the most abundant and widespread lizard species, climatic heterogeneity is mirrored by broad changes in vegetation patterns: forests dominate in the west of its range, whereas shrublands prevail in the east. The genetic diversity of the species is broadly structured in two mtDNA lineages, eastern and western, which diverged ca. 3-3.5 mya (Carranza et al. 2006; Verdú-Ricoy et al. 2010). These lineages show some degree of ecologically-driven divergence, because eastern lizards typically display a striped dorsal pattern absent among western ones, and striped and unstriped phenotypes seem to be adaptively linked to crypsis in the predominant habitat where lizards live (Díaz et al. 2017).

We sampled 95 lizards in five populations along a broad environmental gradient in the center of the Iberian Peninsula, covering both mtDNA lineages (Fig. 2). Three sampling sites housed populations of eastern mtDNA adscription: 1) Lerma (42.058 ^{o}N, -3.611 ^{o}E; 900 m asl), a fragmented mixed forest interspersed with grassland patches, 2) Aranjuez (40.016 ^{o}N, -3.586 ^{o}E; 594 m asl), a hot, dry site with a high cover of herbs and shrubs and no trees, and 3) Brihuega (40.778 ^{o}N, -2.911 ^{o}E; 1,009 m asl), a deciduous open forest with a mosaic of grassland and woodland patches. The two other sampling sites had populations of the western lineage: 4) El Pardo (40.511 ^{o}N, -3.755 ^{o}E; 658 m asl), a xeric, lowland evergreen forest, and 5) Navacerrada (40.726 ^{o}N, -4.023 ^{o}E; 1,230 m asl), a montane location covered by deciduous forest. Several particularities of these populations make them representative of a wide range of selective pressures gathered around the core of this species' range: 1) lizards from Lerma inhabit a very fragmented forest archipelago that resembles the typical habitat of western lizards (although they belong to the eastern lineage; see Díaz et al. 2005; Santos et al. 2008; Telleria et al. 2011; Pérez-Tris et al. 2019; for further information about habitat fragmentation effects in this system); 2) Aranjuez lizards inhabit the typical hot and dry habitat of eastern lizards, and although this locality is very close to the western populations included in this study (El Pardo and Navacerrada), it receives very little gene flow from them (Díaz et al. 2017), so that its isolated condition promotes the accumulation of genetic divergence subject to selection (Llanos-Garrido et al. 2019); and 3) the two western populations are separated by a significant altitudinal gradient, and although lizards from both populations show little genetic differentiation (Díaz et al. 2017), they differ in important phenotypic traits such as escape tactics, sexual dimorphism, sexual ornaments, ectoparasite loads and other life history traits (Iraeta et al. 2006, 2010, 2011; Llanos-Garrido et al. 2017).

DNA extraction, sequencing and variant calling, and outlier analyses

The 21 loci under selection used in this study were detected by outlier search analyses conducted in a previous

study (Llanos-Garrido et al. 2019). Shortly, we obtained tissue samples by clipping 2 cm of the tail tip of lizards, which were afterwards released at their site of capture. We purified DNA for library preparation using the Speedtools Tissue DNA Extraction kit (Biotools).

We used the restriction enzyme Pst1 for GBS library preparation. Sequencing was done in an Illumina HiSeq2500 sequencer. To recover SNPs we used the pipeline UNEAK, implemented in TASSEL v.3.0 (Bradbury et al. 2007), which is specifically designed for samples with no reference genome. We aligned sequence tags to each other to form 'networks' of tags, where each node is a single tag sequence, and each edge represents a single base pair difference between two tags. We pruned the networks to remove putative sequencing errors (low frequency alleles) using the error rate threshold parameter. We also discarded loci with minor allele frequencies < 0.01 or that could be successfully sequenced in less than 10% of individuals. The resulting dataset had 73,291 biallelic SNPs (Single Nucleotide Polymorphism), a site depth of 6.60 ± 6.75 and a site missingness of 0.42 ± 0.31.

To minimize false positives in outlier analyses, we discarded loci that could not be successfully sequenced from at least 75% of individuals in each population, and loci with minor allele frequencies < 0.05 in each population, thus excluding all private alleles from the dataset. Also, prior to performing outlier analyses, we used PLINK v1.9. (Purcell et al. 2007) to prune the SNP database for linkage disequilibrium (LD), according to observed sample correlation coefficients. This was necessary because if the outliers were found on highly correlated contigs, their non-independence could bias subsequent environmental association analysis (explained below). The resulting SNP dataset included 6,421 loci. We used a Bayesian approach to perform an outlier analysis as implemented in Bayescan v.2.1 (Foll and Gaggiotti 2008). Bayescan uses a logistic regression model to partition F_{ST} coefficients into a population-specific term (β) and a locus-specific term (α). We selected loci with $\alpha > 0$ as suggesting positive selection, and a false discovery rate (corrected for multiple testing) $q < 0.05$. To obtain these parameters, we ran the MCMC algorithm implemented in the program with a prior odd value of 10, and using 20 pilot runs of 5,000 iterations each, followed by 100,000 iterations with a burn-in of 50,000 interactions. In order to search for outliers while accounting for coancestry effects, we performed a second outlier analysis using the Bonhomme et al. (2010) extension of the Lewontin-Krakauer test. We also selected loci based on the statistical significance of the FLK statistic, with a restrictive significance threshold of $p < 0.001$ to account for multiple testing.

The outlier analysis performed with Bayescan detected 12 outlier loci with $\alpha > 0$ ($0.97 < \alpha < 1.35$) and $q < 0.05$, while the FLK analysis identified nine additional loci with $p < 0.001$, none of which was previously detected by Bayescan. An MDS analysis performed elsewhere with these 21 SNPs putatively under selection, placed sampled populations along a first major axis that recovered the same pattern of differentiation observed for mtDNA and for some relevant phenotypic traits (Llanos-Garrido et al., 2019). Moreover, these phenotypes were interpreted as adaptive after a process of ecologically-driven divergence in a much wider sample (Díaz et al., 2017).

Quantification of environmental variation

To quantify environmental variation all over the potential range of the species, we selected an area that included its actual distribution range plus a 450-850 km wide perimeter belt around it (width variation depended on the geographical features of range edges). Within this area, we used data from the Bioclim 2.0 dataset (cell resolution = 1x1 km; Booth et al. 2014) to compute the score of each cell on a principal component analysis that combined all Bioclim environmental variables using R core. This PCA yielded a principal axis that opposed hot areas with low precipitation to temperate ones with high precipitation (Fig. 2).

Environmental Association Analysis (EAA)

Environmental values (dependent variable for EAA models) were assigned to the 1x1 km grid cells where individuals had been sampled, using QGIS v2.18.16 (QGIS Delopment Team 2015) and a layer of PCA-scores within polygons defined by sampling locations. The genotypes for each loci (independent variables for EAA models) were recoded as 0, 1 or 2 depending on whether they were homozygous for the reference

allele, heterozygous or homozygous for the alternative allele, respectively. As we pruned the SNP dataset by linkage disequilibrium, we avoided including any collinear predictors in the models.

Our EAA was constrained by the fact that the 95 individuals genotyped belonged to only five different populations. While this ensures a sufficient characterization of genetic variation within populations, it leads to unavoidable pseudoreplication of environmental data (and, depending on the extent of genetic differentiation and aggregation, also of genetic data). To make sure that this problem did not affect our conclusions, we used four randomization approaches, with the same model selection steps applied to each of them, so that they could be compared.

Firstly, we based our EAA on 1,000 random assignments of genotypes within populations to 1x1 km sampled grid cells (hereafter intra-population randomization). Given that environmental variation is several orders of magnitude larger among than within populations (97.6 % of the variance in PCA scores explained by population adscription), we chose to highlight the among-populations component of the models by assuming that all genotypes could occupy every grid cell within the geographical boundaries of their own population (which were determined by the discontinuous distribution of suitable habitat). This is more realistic than assuming a large component of genotype-environment covariation at a local, within-population scale. Moreover, the genetic component of such covariation could not be detected by our methods of outlier detection, which were specifically designed to search for genetic divergence among populations. Our set of 1,000 intra-population randomizations should therefore capture the actual pattern of genotypic and environmental covariation at the scale of the sampled gradient.

Secondly, we randomized 1,000 times the geographical grid cell assigned to each genotype without taking into account its population (hereafter inter-population randomization). This allowed us to produce a null hypothesis of no association between genotypic and environmental variation, but which takes into account the fact that environmental values are geographically structured by population of origin (and thus pseudoreplicated).

Thirdly, we used a randomized set of genetic data to control for the potential effects of genetic structure among populations and genetic aggregation within them (hereafter randomization by neutral loci). For that purpose, we constructed 1,000 new sets by randomly selecting 21 loci from each genotype (i.e. the number of detected outliers) but without taking into account whether they were characterized as outliers or not. This was done to account for the fact that neutral genetic variation (randomly selected SNPs) is expected to have the same degree of aggregation than variation under selection (outliers). However, while we should expect that at least a fraction of the genetic variation subject to selection should be correlated with environmental variation, the opposite is true for the neutral differentiation of populations.

Finally, we were aware that outlier analyses should effectively sort through the randomized SNP databases to identify those that explain the greatest variance among 'populations' (or subgroups). The projection of that variance into environmental PC-space could in turn draw some shape around the five sampled populations' environments that would be considered as suitable habitat, perhaps leading to significant genotype-environment associations. Because of that reason, the outlier selection step needs to be incorporated into our attempts to identify null expectations. To this end, our fourth randomization approach permuted the 6,421-SNP dataset and reran all the analysis from the outlier detection step onwards (hereafter complete randomization). We randomly assigned SNP genotypes to individuals and ran Bayescan to detect outliers putatively under selection with the same methods we used with real data. Then we took the top 21 outlier SNPs to conduct genotype-environment association models and predict the species' range. If the environmental signal of the SNPs generated by these simulations (and, as a consequence, expected by pure chance) is still smaller than that of real data, this must be interpreted as evidence that the SNPs selected by our EAA are not only correlated with the environmental gradient portrayed by our sampling populations, but also that they provide good proxies for genetic variants involved in local adaptation. We did not repeat the FLK extension test with this dataset because no significant phylogenetic (i.e. among-population) patterning is expected in a genetic dataset where genotypes are randomly assigned to individuals (and hence to populations). We performed 500 complete randomization tests instead of 1,000 due to computational limitations.

We performed a backward stepwise multiple regression analysis with each randomized data set (N = 3,500 EAAs in total), with SNPs as predictors and environmental scores as the dependent variable using *lm*function in R core. By doing this, we obtained a distribution of adjusted R^2 estimates and p-values for each set of randomizations. Final model building was achieved by considering the mean p-values of partial correlations calculated for all the datasets obtained with the intra-population randomization strategy. In each step, we removed all SNPs with a mean p-value > 0.5, and we recalculated all partial correlations with the remaining SNPs. In the last step, when all remaining markers had mean p-values < 0.5, we removed all SNPs with mean p-values > 0.05. Our final model (genotype-environment association model, or GEAM) was built with the mean intercept and mean beta values of the remaining SNPs.

Range inference

To infer the distribution range of the species, we followed a two-step procedure. Firstly, we included all the geographical cells that presented the same environmental scores as the sampled populations (predicted range #1). This first approach provides a baseline prediction with no genetic information that can be used to quantify the improvement in predictive ability supplied by GEAM. Secondly, we considered all possible combinations of alleles for the outlier loci selected by GEAM (i.e., all possible genotypes under selection) to predict all environmental values suitable for at least one genotype according to GEAM. By fulfilling all grid cells with those environmental values, we could extrapolate our prediction to the overall distribution range of the species. Finally, we removed from the inferred range a few disconnected patches (in central France, coastal Italy, and the Mediterranean islands) that were too far from the main distribution range of *P. algirus* (>8 km from the nearest inferred distribution limit, which is the distance to the largest patch disconnected from the continuous range of the species), whose low dispersal rate (Santos et al. 2009) is supported by the fact that genetic differentiation can be detected even among forest fragments separated by 350 m of unsuitable arable land (Pérez-Tris et al. 2019). This produced our second (and final) inferred distribution range (predicted range #2). The extent of overlap between real and predicted distribution ranges was estimated using QGIS v2.18.16 (QGIS Development Team 2018).

Results

Environmental Association Analysis

The PCA with all Bioclim environmental variables (N = 19 variables) yielded a single principal component (eigenvalue = 0.679) that retained four variables using the scree plot criterion: annual mean temperature (BIO1), max temperature of the warmest month (BIO5), mean temperature of the warmest quarter (BIO10), and annual precipitation (BIO12). Thereby, this component defined a bioclimatic gradient with a hot and dry extreme in the area occupied by the Sahara desert (highest values) and a temperate and wet extreme in northwestern Spain (lowest values; Fig. 2).

A vast majority of inter-population and loci randomizations resulted in non-significant models (Fig. 3). Mean adjusted R^2 for the random assignment of genotypes to populations (inter-population randomization) was 0.002 (SD = 0.038, range = 0 - 0.199), with a mean p-value of 0.487 (SD = 0.292, range = 0.0004 - 1); only 6% of the 1,000 randomized datasets yielded significant results. Mean adjusted R^2 for loci randomization (i.e. random within-population selection of SNPs, either outliers or not) was 0.002 (SD = 0.105, range = 0 - 0.324), with a mean p-value of 0.494; SD = 0.296, range = 0.001- 1); only 5.6 % of the datasets yielded significant models. In the case of the complete randomizations (i.e. fully permuted SNP datasets prior to outlier analysis), mean adjusted R^2 was 0.047 (SD = 0.092, range = 0 - 0.788), with a mean p-value of 0.616 (SD = 0.451, range = 10^{-7} - 1).

All the datasets built by intra-population randomization produced highly significant models (mean adjusted R^2 = 0.646 ± 0.007, range = 0.623 - 0.670; mean p-value ± SD =3.5 x 10^{-17}± 2.8 x 10^{-17}, range = 2.02 x 10^{-18}– 2.93 x 10^{-16}; Fig. 3). Only 2 out of the 500 complete randomizations yielded a model that explained more environmental variability than those built with intra-population randomized datasets. Thus, the environmental association models including the SNPs under selection (intra-population randomization: mean R^2 = 0.646) had on average between one and three orders of magnitude more predictive power than those

built with the other three randomization strategies (mean $R^2 = 0.002$, 0.002, and 0.047, for inter-population, by neutral loci, and complete randomizations, respectively). The final GEAM included four SNPs with significant partial correlations (Table 1).

Range inference

Of the total number of cells that form the real species' range, 27.83% had the same environmental scores as the sampled populations (or, in other words, predicted range #1 allowed to forecast 27.83% of the species' range); 25.56% of predicted range #1 fell outside real range limits (Fig. 4). Predicted range #2 (the range inferred by extrapolating GEAM to include all grid cells suitable for any possible combination of alleles at the four loci in the final model) was similar to the real species' range. All grid cells in predicted range #1 were included in predicted range #2, accounting for 36.64 % of its total amount. In turn, predicted range #2 captured 75.09 % of the actual distribution of *P. algirus* , with 13.41 % of inferred presences beyond real range limits. The 24.91% fraction of the real distribution range unpredicted by GEAM mostly corresponded to the northwest corner of the range, as well as to a large number of small predicted gaps within it. Nevertheless, predicted range #2 accurately reflected not only the northern and southern edges of the real distribution range, but also many of the gaps within it, both in northwest Africa (i.e. far from the sampled populations) and in many Iberian mountain ranges (Fig. 4).

Because complete randomizations provided the most reliable null hypothesis for EAAs (see above), we used them to infer the species range following the same procedure as described above. We ran a different EAA per fully randomized dataset because each of them included a different set of randomly generated 'outliers' (i.e. SNPs putatively under selection). The mean percentage of inferred range was 29.61 % (SD = 5.581, range = 27.83 – 79.67), roughly equivalent to predicted range #1 but smaller than predicted range #2. Of the 500 complete randomization EEAs, only two were able to infer a larger proportion of the species range than the one predicted by GEAM (P = 0.004, Fig. 5).

Discussion

Extrapolating genotype-environment association analyses to all possible combinations of alleles at a few outlier loci (i.e. supposed to be locally adaptive) allowed us to explore how much environmental diversity could be exploited by a given amount of genetic variation. This, in turn, revealed an adaptability threshold that ultimately defined the distribution boundaries of *P. algirus* . Thus, our successful inference of a species' range from the geographical distribution of a few adaptive loci uncovered the role of genotypic variation at the inter-individual level in shaping a species' distribution range. Our approach should therefore provide compelling evidence of how the genetic dynamics of local adaptation underlie distribution patterns.

Environmental Association Analysis

Our intra-population randomization approach showed that the predictive power of GEAM was much larger than expected by chance. Such high predictive power was based on the genetic diversity found in five closely-located populations, that included both the eastern and western lineages into which *P. algirus* is divided (Carranza et al. 2006; Díaz et al. 2017); geographical distances among these populations are not correlated with either genetic or environmental distances (Llanos-Garrido et al. 2019). Yet, a small number of inter-population randomizations and randomizations by neutral loci also yielded significant models, as expected from a certain degree of environmental pseudoreplication and genetic aggregation in our data. However, the rate of significance was close to 5%, i.e. the conventional level of type I error rate for significance in statistical tests. On the other hand, our complete randomization approach, which included the critical outlier selection step, produced a relatively large number of significant models (25%, still much lower than the 100% obtained by the 'correct' intra-population approach). This confirmed that outlier analyses were effectively able to sort through the randomized SNP databases identifying those that explain the greatest variance among arbitrary subgroups, in such way that the projection of that genetic variance into the environmental PC-space around the five sampled populations resulted into significant association models. However, the environmental signal of these randomly genotyped SNPs was significantly smaller than that of real data. This provides strong support for the idea that the particular SNPs selected by our EAA are good proxies for genetic variants

involved in local adaptation. In addition, given the low standard error of parameter estimates (Table 1), our final genotype-environment association model should be regarded as robust.

Range inference accuracy

Our approach combines the advantages of correlational and mechanistic distribution models (Kearney and Porter 2009), because it relies on genotype-environment correlations based on the geographical distribution of field-captured individuals, but deals with SNPs putatively under selection that should ultimately be associated with functional differences in morphology, physiology, and/or behavior. This should allow us to enlarge the scale of our analysis to the species level (Buckley 2010), because the extrapolation of our results to all possible allele combinations at loci under selection should cover a much wider range of adaptive phenotypic variation than the one revealed by the physiological measurements of a restricted sample of individuals or populations. The only assumption behind this assertion is that all allelic combinations are actually plausible, as it should be expected if loci are correctly pruned by linkage disequilibrium. Moreover, as long as local adaptation leads to a heterogeneous distribution of the genotypes adapted to different parts of the range, our GEAM should be more realistic than mechanistic models because mechanistic models are based on physiological measurements of individuals, and these may lack the specific adaptations required to thrive in specific habitats different from their own (such as the ones that determine range boundaries; Svardal et al. 2015).

Genotype-based range inference was especially accurate at the southern edge of the species' range, including a precise delimitation of range gaps in Morocco, where detailed chorological information is available (Bons and Geniez, 1996). However, we could not test the accuracy of our model for the rest of North Africa due to the lack of detailed distribution maps of *Psammodromus algirus* in this area. The only information about these locations was obtained from the IUCN Red List database (which does not provide data about within-range gaps) and the GBIF database (which has only 22 records in this area, all of which were predicted by our model). Nevertheless, the distribution borders suggested by these databases were accurately predicted by our range inference.

Regarding northern boundaries in the Iberian Peninsula, where detailed corological data are also available (Pleguezuelos 1997), we did not recover the presence of the species in a relatively large NW area where lizard populations do occur, inhabiting suitable habitat patches near the cool, humid end of the tested environmental gradient. This is probably because our outlier analyses did not capture all the genetic variation under selection that is associated with such gradient (see below). Across southern France, the real distribution range of the species does not exceed the Rhône River delta, a geographical barrier which could not be predicted by our method of range inference. However, in the Iberian Peninsula our model successfully recovered the central and eastern parts of the northern range boundary, as well as several within-range gaps associated with mountain ranges (around central plateaus and river valleys) and arid regions in south eastern Spain.

The role of niche boundaries and dispersal limitations in shaping range limits

The accuracy of our prediction of range limits suggests that these were revealing ecological niche boundaries. If other environmental factors (e.g. prey, competitors, predators or parasites) had been constraining range expansions, our inferred range would have extended beyond real range boundaries, and realized range edges would be explained by the existence of limitations to expansion before fulfilling all cells within the spatial projection of the species' niche (Holt 2003). In fact, this happened only in the northeastern border of the range, where GEAM predicted the presence of *P. algirus* beyond the barrier imposed by the Rhône River delta (niche boundaries would actually allow the species to reach Italy).

Also, our results led us to dismiss the existence of demographical center-border effects such as gene swamping or increased homozygosity near the edge of the range (Herrera and Bazaga 2008; Polechová et al. 2009; Pironon et al. 2017). This is because if these processes were acting, they would be limiting the persistence of marginal populations near range boundaries, and we would systematically infer false positives beyond range limits (Case and Taper 2000; Bridle and Vines 2007; Lee-Yaw et al. 2018). However, we did wrongly infer a relatively large inland area of false positives at the northern side of the Pyrenees (Figs. 4 and 5).

Interestingly, this area was suitable for a small number (< 3) of genotypes, which provides a reasonable explanation for these false positives, due to the low probability that the few genotypes that could be adapted to these unoccupied areas were available in nearby marginal populations (Pujol et al. 2009; Dawson et al. 2010; Barton and Etheridge 2018).

The dispersal ability of the genotypes arising at range margins plays an important role in the colonization of new areas beyond range limits (Simmons and Thomas 2004; Hardie and Hutchings 2010). In our system, for example, the range of *P. algirus* would extend ca. 13% beyond its eastern European border if lizards were able to disperse across the Rhône River. Furthermore, genetic diversity would be fostered by greater dispersion abilities (Duckworth 2008), which should facilitate the co-occurrence of adaptive allele combinations at range margins. For instance, the low dispersal rate of these small terrestrial ectotherms would complicate their expansion towards suitable but unoccupied areas north of the Pyrenees (false positives in our model), given the synergistic effects of low dispersal rates and a low probability of finding adapted genotypes at nearby marginal populations. To clarify the role of dispersal ability in the colonization of new locations beyond range limits, our genotype-based modelling approach should be applied to species showing different dispersal abilities (Sanford et al. 2006; Dawson et al. 2010), or we could perform transplant experiments beyond range limits (i.e. manipulating species' dispersal ability; Hargreaves et al. 2014). By doing this, we should be able to discern between the role of dispersal ability *per se* and the genetic contribution of pre-adapted genotypes arising (or not) at marginal populations (Bridle and Vines 2007; Sexton et al. 2009; Phillipsen et al. 2015).

Number of adapted genotypes per cell across the species' range

Besides the general pattern by which fewer genotypes were adapted to range boundaries than to core areas, we found two distinct scenarios within the Iberian Peninsula (Fig. 6). On the one hand, in the northern half of the peninsula there were many areas suitable for a high number of genotypes. In such context, there is a high probability of finding the few genotypes that are adapted to the challenging environmental conditions characteristic of northern boundaries, which should facilitate the establishment of the marginal populations that shape the corresponding edge (Kawecki 2008; Hardie and Hutchings 2010; Halbritter et al. 2015). Conversely, most of the southern half of Iberia seemed to be suitable only for a small number of genotypes, despite the fact that*P. algirus* is abundant in this area. However, several small areas locally suitable for many genotypes were interspersed all across the region. Such areas could therefore play an important role as sources of specific genetic diversity adapted to the demanding, singular environments that surround them (Holt and Keitt 2005; Sagarin et al. 2006). The genotyping of populations that inhabit demanding environments, suitable for a small number of allele combinations, would be crucial to sustain this assertion (Eckert et al. 2008; Gallet et al. 2018). Similarly, a model simulating the intensity of selection in both sources and sinks of genetic diversity should be useful to test whether the genetic variants adapted to demanding environments arise with higher probability in source populations with more relaxed selection regimes (Alleaume-Benharira et al. 2006).

Adaptability thresholds constrain range limits

Our results suggest that species' ranges are determined by the maximum possible span of environmental variation to which adaptive genetic variants are suited. As a consequence, range expansions should be constrained by adaptability thresholds. In our system, it seems that the environmental range to which *P. algirus* is adapted is ultimately linked to the amount of genetic variance under selection associated to a specific bioclimatic gradient. If positive, a range expansion promoted by adaptations towards more extreme environments should entail the selection of new genetic variants. Moreover, such range expansion would require that the effect of the new adaptive mutations is additive with respect to the ones that define the adaptability threshold (Polechová and Barton 2015; Polechová 2018). Whilst our results support this line of reasoning, further theoretical exploration is needed to uncover the hypothesized positive relationship between the magnitude of the increment in environmentally correlated additive genetic variance, and the extent of range expansion that can be achieved (Angert et al. 2008; Polechová et al. 2009).

Overall, we have shown that inferring species' ranges from the geographical distribution of SNPs under

selection can be not only very accurate, but also informative about the genetic dynamics that underlie local adaptation all over a species' range. Our results suggest that the amount of genetic variability subject to selection is a determinant of the location and shape of range boundaries. This conclusion sheds light on the key processes that determine the configuration of distribution ranges, putting forward the importance of inherent limits to adaptation as an ultimate explanation for the evolution of their shape and boundaries (Connallon and Sgrò 2018).

Acknowledgements

This research was funded by the Spanish Ministry of Science and Technology (grant CGL2013-41642-P).

Ethical treatment of animals

Sampling work, measuring methods and tissue collection were performed in accordance with relevant guidelines and regulations approved by the Junta de Castilla y León (Consejería de Medio Ambiente) under license E.P. 117/2001, the Junta de Castilla La Mancha (Consejería de Agricultura) under license DGMEN/SEN/avp_15_096_aut, and the Comunidad de Madrid (Consejería de Medio Ambiente y Ordenación del Territorio) under license 10/089301.9/15.

Data availability statement

Data for this study are available at PANGAEA (Llanos-Garrido, PérezTris, & Díaz, 2019): https://doi.org/10.1594/PANGAEA.908220.

Table 1. Parameter estimates for the regression coefficients of the SNPs under selection that entered the final model.

	Parameter estimate ± SD (min, max)	P-value ± SD (min, max)
Intercept	- 3.630 ± 0.033 (- 3.739, - 3.539)	$2.42 \times 10^{-22} \pm 2.03 \times 10^{-22}$ (1.08×10^{-23}, 1.92×10^{-21})
SNP1	+ 0.605 ± 0.012 (0.567, 0.642)	$3.31 \times 10^{-7} \pm 1.85 \times 10^{-7}$ (6.62×10^{-8}, 1.46×10^{-6})
SNP2	+ 0.488 ± 0.019 (0.430, 0.551)	$5.55 \times 10^{-4} \pm 2.66 \times 10^{-4}$ (1.01×10^{-4}, 1.71×10^{-3})
SNP3	+ 0.347 ± 0.011 (0.315, 0.385)	$3.87 \times 10^{-4} \pm 1.78 \times 10^{-4}$ (7.31×10^{-5}, 1.49×10^{-3})
SNP4	+ 0.366 ± 0.010 (0.328, 0.396)	$2.81 \times 10^{-4} \pm 1.03 \times 10^{-4}$ (7.41×10^{-5}, 8.68×10^{-4})

Figure legends

Figure 1. Known distribution range of *Psammodromus algirus* (based on Bons and Geniez 1996 [north-west Africa] and Pleguezuelos 1997 [Iberian Peninsula]). The gaps within the range that correspond to croplands or cities are not considered. A question mark is placed where the species is known to inhabit but there is no accurate information about its distribution (22 presences scattered all over North Africa). The discontinuous line defines the assumed southern edge of the distribution range according to IUCN.

Figure 2. Bioclimatic gradient defined by the environmental PCA for all the western Mediterranean region (potential distribution range of *P. algirus*). Black circles mark the location of the sampled populations (L = Lerma, B = Brihuega, N = Navacerrada, P = El Pardo, and A = Aranjuez). On the right, the location of these populations within the environmental gradient.

Figure 3. Distribution of adjusted R^2 values for regression models obtained with four different randomized datasets (see text for details).

Figure 4. Inferred distribution range. In dark green, predicted range #1 (grid cells with the same environmental scores than the sampled populations); and in clear green, predicted range #2 (inferred by extrapolating GEAM to any possible combination of alleles at the loci included in the final model).

Figure 5. Distribution of the proportion of the actual species' range that was inferred by completely randomized datasets incorporating the outlier selection step (see text for details). The proportion of range inferred

by GEAM is marked with an arrow.

Figure 6. Temperature map representing the number of adapted genotypes per grid cell according to our GEAM.

Figure 1.

Hosted file

`image1.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Figure 2.

Hosted file

`image2.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Figure 3.

Hosted file

`image3.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Fig. 4

Hosted file

`image4.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Fig. 5

Hosted file

`image5.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Figure 6.

Hosted file

`image6.emf` available at https://authorea.com/users/334163/articles/460218-environmental-association-modelling-with-loci-under-divergent-selection-accurately-predicts-the-distribution-range-of-a-lizard

Bibliography

Ahrens, C. W., P. D. Rymer, A. Stow, J. Bragg, S. Dillon, K. D. L. Umbers, and R. Y. Dudaniec. 2018. The search for loci under selection: trends, biases and progress. Molecular Ecology.

Alleaume-Benharira, M., I. R. Pen, and O. Ronce. 2006. Geographical patterns of adaptation within a species' range: Interactions between drift and gene flow. Journal of Evolutionary Biology.

Angert, A. L., H. D. Bradshaw, and D. W. Schemske. 2008. Using experimental evolution to investigate geographic range limits in monkeyflowers. Evolution.

Barton, N. H., and A. M. Etheridge. 2018. Establishment in a new habitat by polygenic adaptation. Theoretical Population Biology.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. SanCristobal. 2010. Detecting selection in population trees: The Lewontin and Krakauer test extended. Genetics.

Booth, T. H., H. A. Nix, J. R. Busby, and M. F. Hutchinson. 2014. Bioclim: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Diversity and Distributions.

Bradbury, P. J., Z. Zhang, D. E. Kroon, R. M. Casstevens, Y. Ramdoss, and E. S. Buckler. 2007. TASSELL Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Bridle, J. R., and T. H. Vines. 2007. Limits to evolution at range margins: when and why does adaptation fail? Trends in Ecology and Evolution.

Brown, J. H. 2002. On the Relationship between Abundance and Distribution of Species. The American Naturalist.

Buckley, L. B. 2010. The range implications of lizard traits in changing environments. Global Ecology and Biogeography.

Carranza, S., D. J. Harris, E. N. Arnold, V. Batista, and J. P. Gonzalez De La Vega. 2006. Phylogeography of the lacertid lizard, Psammodromus algirus, in Iberia and across the Strait of Gibraltar. Journal of Biogeography.

Case, T. J., and M. L. Taper. 2000. Interspecific Competition, Environmental Gradients, Gene Flow, and the Coevolution of Species' Borders. The American Naturalist.

Charlesworth, B. 2009. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics.

Connallon, T., and C. M. Sgrò. 2018. In search of a general theory of species' range evolution. PLoS Biology.

Dawson, M. N., R. K. Grosberg, Y. E. Stuart, and E. Sanford. 2010. Population genetic analysis of a recent range expansion: Mechanisms regulating the poleward range limit in the volcano barnacle Tetraclita rubescens. Molecular Ecology.

Diaz, J. A., and L. M. Carrascal. 2006. Regional Distribution of a Mediterranean Lizard: Influence of Habitat Cues and Prey Abundance. Journal of Biogeography.

Díaz, J. A., J. Pérez-Tris, J. L. Tellería, R. Carbonell, and T. Santos. 2005. Reproductive investment of a lacertid lizard in fragmented habitat. Conservation Biology 19:1578–1585.

Díaz, J. A., J. Verdú-Ricoy, P. Iraeta, A. Llanos-Garrido, A. Pérez-Rodríguez, and A. Salvador. 2017. There is more to the picture than meets the eye: adaptation for crypsis blurs phylogeographical structure in a lizard. Journal of Biogeography.

Duckworth, R. A. 2008. Adaptive Dispersal Strategies and the Dynamics of a Range Expansion. The American Naturalist.

Dudaniec, R. Y., C. J. Yong, L. T. Lancaster, E. I. Svensson, and B. Hansson. 2018. Signatures of local adaptation along environmental gradients in a range-expanding damselfly (Ischnura elegans). Molecular Ecology.

Eckert, C. G., K. E. Samis, and S. C. Lougheed. 2008. Genetic variation across species' geographical ranges: The central-marginal hypothesis and beyond. Molecular Ecology.

Foll, M., and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics.

Gallet, R., R. Froissart, and V. Ravigné. 2018. Experimental demonstration of the impact of hard and soft selection regimes on polymorphism maintenance in spatially heterogeneous environments. Evolution.

Halbritter, A. H., R. Billeter, P. J. Edwards, and J. M. Alexander. 2015. Local adaptation at range edges: Comparing elevation and latitudinal gradients. Journal of Evolutionary Biology.

Haldane, J. B. S. 1956. The Relation between Density Regulation and Natural Selection. Proceedings of the Royal Society B: Biological Sciences.

Hardie, D. C., and J. A. Hutchings. 2010. Evolutionary ecology at the extremes of species' ranges. Environmental Reviews.

Hargreaves, A. L., K. E. Samis, and C. G. Eckert. 2014. Are Species' Range Limits Simply Niche Limits Writ Large? A Review of Transplant Experiments beyond the Range. The American Naturalist.

Herrera, C. M., and P. Bazaga. 2008. Adding a third dimension to the edge of a species' range: Altitude and genetic structuring in mountainous landscapes. Heredity.

Hoban, S., J. L. Kelley, K. E. Lotterhos, M. F. Antolin, G. Bradburd, D. B. Lowry, M. L. Poss, et al. 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions (Supplemental Material). The American Naturalist.

Holt, R. D. 2003. On the evolutionary ecology of species' ranges. Evolutionary Ecology Research.

Holt, R. D., and T. H. Keitt. 2005. Species' borders: A unifying theme in ecology. Oikos.

Hutchinson, G. E. 1957. Concluding remarks on the cold Spring Harbor Symposia on Quantitative Biology. Population Studies: Animal Ecology and Demography. Cold Spring Harbor SYmposia on Quantitative Biology.

Iraeta, P., C. Monasterio, A. Salvador, and J. A. Díaz. 2006. Mediterranean hatchling lizards grow faster at higher altitude: A reciprocal transplant experiment. Functional Ecology 20:865–872.

Iraeta, P., C. Monasterio, A. Salvador, and J. A. Díaz. 2011. Sexual dimorphism and interpopulation differences in lizard hind limb length:Locomotor performance or chemical signalling? Biological Journal of the Linnean Society.

Iraeta, P., A. Salvador, C. Monasterio, and J. A. Díaz. 2010. Effects of gravidity on the locomotor performance and escape behaviour of two lizard populations: The importance of habitat structure. Behaviour.

Kawecki, T. J. 2008. Adaptation to Marginal Habitats. Annual Review of Ecology, Evolution, and Systematics.

Kearney, M., and W. Porter. 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecology Letters.

Kirkpatrick, M., and N. H. Barton. 1997. Evolution of a Species' Range. The American Naturalist.

Lande, R., and S. Shannon. 2006. The Role of Genetic Variation in Adaptation and Population Persistence in a Changing Environment. Evolution.

LaRue, E. A., J. D. Holland, and N. C. Emery. 2018. Environmental predictors of dispersal traits across a species' geographic range. Ecology.

Lee-Yaw, J. A., M. Fracassetti, and Y. Willi. 2018. Environmental marginality and geographic range limits: a case study with Arabidopsis lyrata ssp. lyrata. Ecography.

Llanos-Garrido, A., J. A. Díaz, A. Pérez-Rodríguez, and E. Arriero. 2017. Variation in male ornaments in two lizard populations with contrasting parasite loads. Journal of Zoology.

Llanos-Garrido, A., Pérez-Tris, J. and Díaz, J. A. 2019. The combined use of raw and phylogenetically independent methods of outlier detection uncovers genome-wide dynamics of local adaptation in a lizard. Ecology and Evolution.

Nosil, P., and B. J. Crespi. 2004. Does gene flow constrain adaptive divergence or vice versa? A test using ecomorphology and sexual isolation in Timema cristinae walking-sticks. Evolution.

Perez-Tris, J., Llanos-Garrido, A., Bloor, P., Carbonell, R., Telleria, J. L., Santos, T., and Diaz, J. A. 2019. Increased individual homozygosity is correlated with low fitness in a fragmented lizard population. Biological Journal of the Linnean Society.

Peterson, A. T. 2011. Ecological niche conservatism: A time-structured review of evidence. Journal of Biogeography.

Phillipsen, I. C., E. H. Kirk, M. T. Bogan, M. C. Mims, J. D. Olden, and D. A. Lytle. 2015. Dispersal ability and habitat requirements determine landscape-level genetic patterns in desert aquatic insects. Molecular Ecology.

Pironon, S., G. Papuga, J. Villellas, A. L. Angert, M. B. Garcia, and J. D. Thompson. 2017. Geographic variation in genetic and demographic performance: new insights from an old biogeographical paradigm. Biological Reviews.

Polechova, J. 2018. Is the sky the limit? On the expansion threshold of a species' range. PLoS Biology.

Polechova, J., and N. H. Barton. 2015. Limits to adaptation along environmental gradients. Proceedings of the National Academy of Sciences.

Polechova, J., N. Barton, and G. Marion. 2009. Species' Range: Adaptation in Space and Time. The American Naturalist.

Pujol, B., S.-R. Zhou, J. Sanchez Vilas, and J. R. Pannell. 2009. Reduced inbreeding depression after species range expansion. Proceedings of the National Academy of Sciences.

QGIS Delopment Team. 2015. QGIS Geographic Information System. Open Source Geospatial Foundation.

Rellstab, C., F. Gugerli, A. J. Eckert, A. M. Hancock, and R. Holderegger. 2015. A practical guide to environmental association analysis in landscape genomics. Molecular Ecology.

Sagarin, R. D., S. D. Gaines, and B. Gaylord. 2006. Moving beyond assumptions to understand abundance distributions across the ranges of species. Trends in Ecology and Evolution.

Sanford, E., S. B. Holzman, R. A. Haney, D. M. Rand, and M. D. Bertness. 2006. Larval tolerance, gene flow, and the northern geographic range limit of fiddler crabs. Ecology.

Santos, T., J. A. Diaz, J. Perez-tris, R. Carbonell, and J. L. Telleria. 2008. Habitat quality predicts the distribution of a lizard in fragmented woodlands better than habitat fragmentation. Animal Conservation.

Santos, T., J. Perez-Tris, R. Carbonell, J. L. Telleria, and J. A. Diaz. 2009. Monitoring the performance of wild-born and introduced lizards in a fragmented landscape: Implications for ex situ conservation programmes. Biological Conservation.

Sexton, J. P., P. J. McIntyre, A. L. Angert, and K. J. Rice. 2009. Evolution and Ecology of Species Range Limits. Annual Review of Ecology, Evolution, and Systematics.

Sexton, J. P., S. Y. Strauss, and K. J. Rice. 2011. Gene flow increases fitness at the warm edge of a species' range. Proceedings of the National Academy of Sciences.

Simmons, A. D., and C. D. Thomas. 2004. Changes in Dispersal during Species' Range Expansions. The American Naturalist.

Svardal, H., C. Rueffler, and J. Hermisson. 2015. A general condition for adaptive genetic polymorphism in temporally and spatially heterogeneous environments. Theoretical Population Biology.

Telleria, J. L., J. A. Diaz, J. Perez-Tris, E. de Juana, I. de la Hera, P. Iraeta, A. Salvador, et al. 2011. Barrier effects on vertebrate distribution caused by a motorway crossing through fragmented forest landscape. Animal Biodiversity and Conservation 2:331–340.

Verdu-Ricoy, J., S. Carranza, A. Salvador, S. D. Busack, and J. A. Diaz. 2010. Phylogeography of Psammodromus algirus (Lacertidae) revisited: Systematic implications. Amphibia Reptilia.

Verdu-Ricoy, J., P. Iraeta, A. Salvador, and J. A. Diaz. 2014. Phenotypic responses to incubation conditions in ecologically distinct populations of a lacertid lizard: A tale of two phylogeographic lineages. Journal of Zoology.

Whitlock, M. C., and K. E. Lotterhos. 2015. Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F ST . The American Naturalist.

Supplementary material 1: Description of sequencing, variant calling, filtering and outlier analysis.

The methods that are described here are fully detailed in Llanos-Garrido et al. 2019.

We used the restriction enzyme Pst1 for GBS library preparation. Sequencing was done in an Illumina HiSeq2500 sequencer. To recover SNPs we used the pipeline UNEAK, implemented in TASSEL v.3.0 (Bradbury et al. 2007), which is specifically designed for samples with no reference genome. We aligned sequence tags to each other to form 'networks' of tags, where each node is a single tag sequence, and each edge represents a single base pair difference between two tags. We pruned the networks to remove putative sequencing errors (low frequency alleles) using the error rate threshold parameter. We also discarded loci with minor allele frequencies < 0.01 or that could be successfully sequenced in less than 10% of individuals. The resulting dataset had 73,291 biallelic SNPs (Single Nucleotide Polymorphism), a site depth of 6.60 ± 6.75 and a site missingness of 0.42 ± 0.31.

To minimize false positives in outlier analyses, we discarded loci that could not be successfully sequenced from at least 75% of individuals in each population, and loci with minor allele frequencies < 0.05 in each population, thus excluding all private alleles from the dataset. Also, prior to performing outlier analyses, we used PLINK v1.9. (Purcell et al. 2007) to prune the SNP database for linkage disequilibrium (LD), according to observed sample correlation coefficients. This was necessary because if the outliers were found on highly correlated contigs, their non-independence could bias subsequent environmental association analysis (explained below). The resulting SNP dataset included 6,421 loci. We used a Bayesian approach to perform an outlier analysis as implemented in Bayescan v.2.1 (Foll and Gaggiotti 2008). Bayescan uses a logistic

regression model to partition F_{ST} coefficients into a population-specific term (β) and a locus-specific term (α). We selected loci with $\alpha > 0$ as suggesting positive selection, and a false discovery rate (corrected for multiple testing) $q < 0.05$. To obtain these parameters, we ran the MCMC algorithm implemented in the program with a prior odd value of 10, and using 20 pilot runs of 5,000 iterations each, followed by 100,000 iterations with a burn-in of 50,000 interactions. In order to search for outliers while accounting for coancestry effects, we performed a second outlier analysis using the Bonhomme et al. (2010) extension of the Lewontin-Krakauer test. We also selected loci based on the statistical significance of the FLK statistic, with a restrictive significance threshold of $p < 0.001$ to account for multiple testing.

Supplementary material 2: Description of the code used to perform intrapopulation randomizations to build a GEAM and use it to infer an entire species range (ref GitHub: to be determined)

Firstly, we randomized the distribution of samples by using the *sample* function from R base among the geographic cells within the area of their respective populations. We obtained 1000 randomized distributions, which consist in datasets with three fields: geographic cell, population and individual. Then, for each resulting table, we added another column with the environmental value that corresponded to each geographic cell (from the score value of a previously performed environmental PCA, see main text). We also included the dose of the (arbitrarily assigned) reference allele for each of the 21 previously detected SNPs under selection. Then, we applied a backwards stepwise regression with the genotypes of these loci as independent variable and the environmental value at each geographical cell as dependent variable. To do this, we built a basic function called "GEAM", which simply consisted in a *while* function with the following elements: 1) a loop that iterates through all randomized tables to perform each multiple regression

with the R function *lm*, 2) computation of the mean p values for each partial correlation between SNPs and environmental values, and 3) selection of the SNPs with the most significant association by discarding the ones with progressively higher p values (threshold set at p-value = 0.5 with a final step set at p-value = 0.05). This process was repeated until all SNPs contained in the model showed p values < 0.05 (which was the condition implemented within the *while* function).

During this process, we saved the p value and R squared of each model in tables with each row containing the results from each iteration of the loop (i.e. the regression model of a particular distribution table). This was in turn stored in a list where each element contained the results from a complete loop (i.e. one step of the *while* function).

Once we identified the most fitted SNPs, we computed their mean partial correlation coefficients from the resulting model after all the iterations of the *while* function to build the final GEAM. Then, we extrapolated it to all possible genotypic combinations of the SNPs that showed a significant partial correlation with the environmental gradient. We obtained all these combinations by using the R base function *expand.grid* (with 0, 1 and 2 representing all possible genotypes as if they were doses of an arbitrary reference allele, and a grid length of 4). The output of this model were the environmental values associated with each combination of SNPs. The inferred distribution was obtained by fulfilling all the geographic cells that presented any of these environmental values.

Supplementary table 1: Inferred suitable areas for each possible genotype (asterisks mark sampled genotypes) and their overlap with the actual distribution range of *Psammodromus algirus*.

Genotypes (number of reference alleles at each locus)	Suitable geographical cells according to GEAM	Overlap with actual distribution range
0000	21031	0.008
0001	22185	0.009
0002	20553	0.011
0010*	24254	0.009
0011	19386	0.009
0012	22914	0.014
0020*	19665	0.009
0021	22585	0.013
0022	23891	0.019
0100*	19665	0.009
0101	22585	0.013
0102	23891	0.019
0110*	21930	0.012
0111*	23438	0.018
0112	26991	0.023
0120*	23521	0.016
0121	25565	0.02
0122	29375	0.024
0200*	23521	0.016
0201*	25565	0.02
0202*	29375	0.024
0210*	24707	0.019
0211*	29064	0.024
0212	29729	0.024
0220*	28115	0.023
0221	29320	0.024
0222*	37252	0.03
1000	19665	0.009
1001	22585	0.013
1002	23891	0.019
1010	21930	0.012
1011	23438	0.018
1012	26991	0.023
1020*	23521	0.016
1021	25565	0.02

1022	29375	0.024
1100	23521	0.016
1101*	25565	0.02
1102*	29375	0.024
1110	24707	0.019
1111*	29064	0.024
1112	29729	0.024
1120	28115	0.023
1121*	29320	0.024
1122*	37252	0.03
1200*	28115	0.023
1201*	29320	0.024
1202*	37252	0.03
1210*	29525	0.025
1211*	33824	0.027
1212*	49514	0.041
1220*	31236	0.024
1221*	44152	0.037
1222*	55956	0.044
2000	23521	0.016
2001	25565	0.02
2002	29375	0.024
2010	24707	0.019
2011	29064	0.024
2012	29729	0.024
2020	28115	0.023
2021	29320	0.024
2022	37252	0.03
2100	28115	0.023
2101	29320	0.024
2102	37252	0.03
2110	29525	0.025
2111	33824	0.027
2112	49514	0.041
2120	31236	0.024
2121	44152	0.037
2122	55956	0.044
2200*	31236	0.024
2201*	44152	0.037
2202*	55956	0.044
2210*	39886	0.033
2211	53885	0.043

2212	50911	0.031
2220*	51655	0.042
2221*	52501	0.035
2222*	49938	0.011