Hybridization in late stages of speciation: Strong but incomplete genome-wide reproductive isolation and 'large Z-effect' in a moving hybrid zone

Guilherme Caeiro-Dias, Alan Brelsford, Mariana Meneses-Ribeiro, Pierre-André Crochet, Catarina Pinho

Hybridization in late stages of speciation: strong but incomplete genome-wide reproductive isolation and "large Z-effect" in a moving hybrid zone

Guilherme Caeiro-Dias [1,2,3,a,*], Alan Brelsford [4,5], Mariana Meneses-Ribeiro [1,2,6], Pierre-André Crochet[4], Catarina Pinho [1,6]

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

[3]CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France

[4]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[5]Biology Department, University of California Riverside, California, USA

[6]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal

[a] (Present Address) Department of Biology and Museum of Southwestern Biology, University of New Mexico, Albuquerque, New Mexico, USA.

* Corresponding author: Telephone: (505) 277-3411, E-mail: gcaeirodias@unm.edu

Key words: Introgression, Interspecific gene flow, Geographic clines, Genomic clines; Lacertidae; ddRAD sequencing

Running title: Hybridization in late stages of speciation

ORCID:

Guilherme Caeiro-Dias: 0000-0002-5771-6208

Alan Brelsford: 0000-0002-5074-7765

Mariana Meneses-Ribeiro: 0000-0001-7340-301X

Catarina Pinho: 0000-0001-5150-2850

Pierre-André Crochet: 0000-0002-0422-3960

Abstract

In organisms reproducing sexually, speciation occurs when increasing divergence results in pre- or post-zygotic reproductive isolation between lineages. Studies focusing on reproductive isolation origin in early stages of speciation are common; and many rely on genomic scans to infer introgression providing limited information on the genomic architecture of reproductive isolation long-term maintenance. This study analyses a natural hybrid zone between two species in a late stage of speciation. We used ddRADseq genotyping in the contact between *Podarcis bocagei* and *P. carbonelli* to examine admixture extent, analyse hybrid zone stability, and assess genome-wide variation in selection against introgression. We confirmed strong but incomplete reproductive isolation in a bimodal hybrid zone. New findings revealed population genetic structure within *P. carbonelli* in the contact zone; geographic and genomic clines analysis suggested strong selection against gene flow, but a relatively small proportion of the loci can introgress, mostly within the narrow contact zone. However, geographic clines revealed that a few introgressed loci show signs of potential positive selection, particularly into *P. bocagei*. Geographic clines also detected a signal of hybrid zone movement towards *P. bocagei* distribution. Genomic cline analysis revealed heterogeneous patterns of introgression among loci within the syntopy zone, but the majority maintain a strong association with the genomic background of origin. However, incongruences between both cline approaches were found, potentially driven by confounding effects on genomic clines. Last, an important role of the Z chromosome in reproductive isolation is suggested. Importantly, overall patterns of restricted introgression seem to result from numerous strong intrinsic barriers across the genome.

Introduction

The evolution of reproductive isolation is a gradual process (Templeton, 1992) and hybridization often persists from the initial stages of speciation (Mallet, 2005; Rieseberg, 2009) to more advanced stages (Nosil, Harmon, & Seehausen, 2009). Hybridization and incomplete reproductive isolation may thus result in genetic introgression during most of the duration of the speciation process (Melo-Ferreira, Alves, Freitas, Ferrand, & Boursot, 2009; Mullen, Dopman, & Harrison, 2008; Pinho & Hey, 2010; Tarroso, Pereira, Martínez-Freiría, Godinho, & Brito, 2014). However, as divergence time increases, the build-up of deleterious epistatic interactions between independently evolving loci, also called Bateson-Dobzhansky-Muller incompatibilities (BDMI), impedes gene flow, gradually causing complete reproductive isolation via hybrid unviability or sterility (Bateson, 1909; Dobzhansky, 1937; Gourbiere & Mallet, 2010; Matute, Butler, Turissini, & Coyne, 2010; Muller, 1942). Many empirical studies using genomic tools (Ellegren et al., 2012; Gompert et al., 2010; Larson, White, Ross, & Harrison, 2014; Parchman et al., 2013) support the idea that genomes are not homogenous ensembles (Nosil & Feder, 2012; Wu, 2001) and that reproductive isolation affects individual loci differently depending on their linkage to incompatibility genes. Deciphering the nature, number, and distribution of such genetic incompatibilities across the genome remains one of the main challenges in the study of speciation.

The build-up of reproductive isolation during divergence can occur through several processes. For example, it can occur due to the accumulation of BDMI starting in 'genomic islands' of reduced gene flow and elevated differentiation (Ellegren et al.,

2012; Feder & Nosil, 2010; Turner, Hahn, & Nuzhdin, 2005), becoming more extensive when divergent selection reduces effective gene flow in the surrounding genomic regions, i.e., divergence hitchhiking (Feder & Nosil, 2010; Via, 2012), and may be particularly important for speciation with gene flow if blocks of associated loci that contribute to isolation are co-inherited (Feder, Egan, & Nosil, 2012). Alternative processes that lead to the development of reproductive isolation involve gene incompatibilities arising in many unlinked loci scattered throughout the genome (Feder, Gejji, Yeaman, & Nosil, 2012; Feder & Nosil, 2010) and barrier loci appear later in the speciation process (Dufresnes et al., 2021). Often the evolution of strong reproductive isolation involves multiple barrier effects, interactions among them and the arose of new interactions (Bierne, Welch, Loire, Bonhomme, & David, 2011; Butlin & Smadja, 2018; Kulmuni, Butlin, Lucek, Savolainen, & Westram, 2020), but ultimately, as the speciation process proceeds and regardless of the mechanisms involved, the evolution of reproductive isolation is expected to result in identical situations where genomes are too divergent to be compatible.

Many studies on the genomics of divergence focus on the early stages of speciation (e.g., Andrew & Rieseberg, 2013; Janoušek et al., 2012; Luttikhuizen, Drent, Peijnenburg, Van der Veer, & Johannesson, 2012; Parchman et al., 2013; Teeter et al., 2010). These early stages are clearly the best to document genomic architectures at the onset of differentiation. However, the factors influencing isolation may vary during different stages of speciation (Ravinet et al., 2017). During the early stages, genetic divergence is low and long-term isolation often rests on pre-mating barriers if divergent populations are sympatric (Egan et al., 2015; van Schooten et al., 2020), but in later stages, the relative contribution of pre- and post-mating barriers often remains an open question. As such, studying genomic mechanisms by which reproductive

isolation is maintained (and not only initiated) is equally important. The study of genomic divergence and gene flow in the later stages of speciation is thus crucial to understand which factors influence the genomic architecture of reproductive isolation across the whole process of speciation.

Another common question in the genomics of speciation is whether sex chromosomes have a disproportional higher influence on reproductive isolation when compared to autosomes, often formalized as the "large X-effect" and Haldane's rule. The large X-effect is the higher influence of X (or Z; here referred to as "large Z-effect") chromosome on hybrid incompatibilities (Coyne & Orr, 2004; Masly & Presgraves, 2007; Orr, 1987), as evidenced by restricted introgression on X or Z-linked loci (Payseur, Krenz, & Nachman, 2004; Storchová, Reif, & Nachman, 2010; Tucker, Sage, Warner, Wilson, & Eicher, 1992). Haldane's rule stems from the observation that fitness reductions are often stronger in hybrids of the heterogametic sex (Haldane, 1922), and can cause reduced introgression of loci linked to sex chromosomes and mitochondrial genomes when compared to autosomes (Carling & Brumfield, 2008; Gowen et al., 2014; Walsh, Shriver, Olsen, & Kovach, 2016). Thus, differential patterns of introgression between sex-linked, autosome and mtDNA markers can inform us on the evolutionary forces underlying hybrid incompatibilities.

Here we study the genomics of the late stages of speciation by quantifying global and locus-specific patterns of introgression in the only known hybrid zone between two species exhibiting strong but incomplete reproductive isolation: *Podarcis bocagei* and *P. carbonelli*. These two species are part of the *Podarcis hispanicus* complex (Iberian and North African wall lizards) and despite their deep divergence (5.6 Mya; Kaliontzopoulou, Pinho, Harris, & Carretero, 2011) they still hybridize in a narrow and bimodal hybrid zone (Caeiro-Dias et al., 2021; Pinho, Kaliontzopoulou, Carretero,

Harris, & Ferrand, 2009). These species' distributions are allopatric except in this area where both species come into secondary contact. Based on biogeographic and phylogenetic patterns (Pinho, Harris, & Ferrand, 2007; Sá-Sousa, 2001) secondary contact likely happened after post-glacial expansions of both *P. bocagei* and northern populations of *P. carbonelli*. Whether such events of expansion imprinted a genomic signature in the contact zone is still unclear.

More specifically, we used ddRADseq genotyping data to i) investigate if the bimodality of the hybrid zone, found in previous studies, can be explained alone by low frequency of hybridization, i.e., if pre-mating barriers between parental species are enough to explain such bimodality; ii) evaluate the extent of introgression and the nature of possible barriers to interspecific gene flow; iii) assess potential effects of population expansion on the stability of the hybrid zone; iv) investigate if loci that show reduced introgression are spread over the genome or grouped in particular regions; and v) if sexual chromosomes have a higher contribution to reproductive isolation than autosomes.

Materials and methods

Transect sampling

We collected tissue samples across a 44 km transect including the contact zone between *P. bocagei* and *P. carbonelli* previously studied by Pinho et al. (2009) and Caeiro-Dias et al. (2021, see Figure 1). The contact zone was sampled in a way to include the area of sympatry and adjacent areas to north and south where only *P. bocagei* and *P. carbonelli* were found, respectively, to ensure that the complete contact zone was sampled. Here we collected 193 individuals across 2.3 km. We also sampled

P. carbonelli in three sites south of ESP (SIL, ESM and TOR) and *P. bocagei* in four sites to the north (AGU, FRA, MAD and POR). Sampling details are provided in Figure 1, Supplemental Information File 1 (Methods) and Supplemental Information File 2.

Extraction of DNA, RAD sequencing, and SNP discovery

Genomic DNA was extracted using the EasySpin® Genomic DNA Tissue Kit (Citomed, Odivelas, Portugal) following the supplier's protocol. The quality and quantity of extracted DNA was evaluated on a 2% agarose gel and a QUBIT 2.0 fluorimeter (Life Technologies, Grand Island, NY, USA).

We obtained RAD sequencing data using modifications to protocols from Parchman *et al.* (2012), Peterson *et al.* (2012) and Purcell *et al.* (2014). The complete protocol was described by Brelsford *et al.* (2016) and summarized in Supplemental Information File 1 (Methods). A single library containing all samples was sequenced on eight Illumina® (San Diego, CA, USA) HiSeq 2000 lanes in the Lausanne Genetic Technology Facility (Lausanne, Switzerland), with single-end 100 bp reads. Four individuals were included in triplicate to assess sequencing errors.

We demultiplexed the raw reads using the process_radtags module of Stacks version 1.3 (Catchen, 2013) allowing one mismatch for barcodes, to remove low-quality reads, reads with uncalled bases and reads that failed the Illumina 'chastity' filter. Adapters were trimmed from demultiplexed reads using a custom shell script. Reads were then aligned in each individual *de novo* to assemble loci using the denovo_map.pl wrap up program. A *de novo* assembly was chosen over the possibility of an assembly guided by a reference genome since the *P. muralis* genome used latter in this study was not yet available at this stage of the work. Furthermore, we discovered enough SNPs using a *de novo* assembly for the type of analysis employed in this study. Each

locus was assembled with a minimum depth of sequencing coverage of one and allowing two mismatches between stacks within individuals and two between individuals. We then used VCFtools version 0.1.15 (Danecek *et al.*, 2011) to filter the resulting variants. Loci with depth coverage <8, with missing data >20% or with alleles with minimum frequency <0.05 were removed. To avoid possible paralogs, we removed loci containing more than five SNPs and with more than 70% heterozygosity. We selected one SNP per locus by choosing the SNPs maximizing frequency differences between species with a custom Python script available at https://github.com/catpinho/filter_RADseq_data/blob/master/removes_loc_multipleSNPs%2Bleaves1SNP_highfreq.py. This final dataset (complete dataset) was then used to generate two additional datasets, prior to apply a missing data filter by individual. One for loci with allele frequencies higher than 0.8 in reference population containing only parental individuals of one species and lower than 0.2 in the reference population of the other species ("80/20" dataset). The other dataset includes loci with fixed differences between parental populations of *P. bocagei* and *P. carbonell* (diagnostic dataset). The sites POR or TOR were used as the reference populations for *P. bocagei* and *P. carbonelli*, respectively, to produce both datasets. Individuals with missing data >35% were removed from all datasets.

Population structure across the transect

We used Admixture version 1.3.0 (Alexander & Lange, 2011; Alexander, Novembre, & Lange, 2009) to evaluate the proportion of individual genome ancestry (*Q*) from each parental species on the complete dataset (the proportion of *P. carbonelli* ancestry, Q_C, is used in this study to describe the results; the proportion of *P. bocagei* ancestry is $Q_B = 1 - Q_C$). Admixture was run until the log- likelihood estimate increased less than 10^{-5}

between iterations. We first ran Admixture for K=2 because we were mainly interested in individual admixture between the two species. We also extended this analysis from K=2 to K=6 to detect possible sub-structure within species. The best K was estimated with a five-fold cross-validation error procedure, implemented in Admixture where the best value of K exhibits the lowest cross-validation error. The distributions of *Qc* scores in 10 classes of equal size (i.e., classes of 10%) for the contact zone are also reported.

Genomic variability among individuals was visualised by performing a principal component analysis (PCA) on the complete dataset with R version 4.0.3 (R Core Team 2020) using the adegenet R package version 2.0.1 (Jombart, 2008; Jombart & Ahmed, 2011). PCA requires that the dataset do not contain any missing data and thus the missing data values had to be replaced. Instead of using the commonly applied "mean method" to substitute the missing data (where missing data are replaced irrespective of the sample group, here the species), which increases similarity of samples when missing data are numerous and thus mimics the signal of introgression, we used the Breiman's regression random forest algorithm (Breiman, 2001) implemented in R package randomForest version 4.6-14 (Liaw & Wiener, 2002). The values of our missing data in each of the three datasets were predicted from 500 independently constructed regression trees and 50 bootstrap iterations with default bootstrap sample size.

To distinguish putative F1 hybrids in the hybrid zone from other hybrid classes (e.g., F2, backcrosses) and to evaluate the proportion of contemporary hybridization events, we used a triangle plot representing the relationship between interspecific heterozygosity (*He*) and the hybrid index (*HI*) as described in Fitzpatrick (2012). Further details are provided in Supplemental Information File 1 (Methods). Because in the literature the term "backcross" (BC) is used to refer to both first-generation BC and

later generation BC, we would like to clarify that throughout the article we use "backcross" in a general way to refer to both cases, otherwise each specific case will be indicated. We used the diagnostic dataset (1241 loci) to estimate the *He* and an *HI* ($HI_{DIAG-INT}$) for each individual in the two populations with signs of admixture (ESP and SIL; see Results), using the R package Introgress version 1.2.3 (Gompert & Buerkle 2010), setting *P. bocagei* parental individuals to have an $HI_{DIAG-INT}$ of 0 and *P. carbonelli* parental individuals to 1. The two populations in the extremes of the transect (POR and TOR) were used as the parental populations.

Mitochondrial genotyping

The mtDNA lineage of each sample was identified by sequencing a partial fragment of the 12S ribosomal RNA region (12S). Amplification and sequencing were performed as described in Pinho et al. (2006) using the primers published in Kocher et al. (1989). We used individuals from all populations sampled except MAD. Identification was performed by comparing the target sequences with our extensive database of reference sequences using tree building approaches like in Kaliontzopoulou et al. (2011). The identification of mtDNA lineage together with the sex identification of potential F1 and F2 hybrids and first-generation BCs may offer some clues about how F1s, F2s and first-generation BCs were formed, although the number of samples in these hybrid classes was very low (see Results), precluding any statistical analyses. Furthermore, mtDNA lineage was fit a to geographic cline model (*Geographic cline analyses* section) to compare with nuclear clines.

Simulations of genotypic composition within the contact zone

The low number of backcrosses detected (eight individuals) in our data compared to the number of F1 hybrids (three individuals) suggested poor fitness of F1 hybrids. However, our sample sizes of various hybridization classes were small, so random sampling effects were difficult to exclude. To test whether the observed relative proportion of backcrosses and F1 hybrids in the contact zone can be obtained by random sampling effects without counter-selection of F1 hybrids, we simulated 100 replicates of various scenarios combining “random mating” (RM; no restriction to mating between any individual) and “non-random mating” (NRM) with four levels of migration (0, 20%, 30% and 40% of individuals replaced per generation). In NRM scenarios the rate of heterospecific mating was set to the observed proportion of putative F1s to the total number of observed parental individuals in each generation. Once admixed individuals were formed there were no restrictions to gene flow between them and parental or other admixed individuals. A detailed description of simulations is provided in Supporting File 1.

Specifically, we tested if immigration and assortative mating between parental individuals in the contact zone are enough to explain the observed proportions of F1 hybrids and backcrosses without lower survival, lower fecundity, or lower matting success of F1 hybrids. Each distribution for each simulated scenario (combination of RM and NRM with the 0, 0.05, 0.2, 0.3 and 0.4 of immigrants) was compared to the real distribution using a series of pairwise Fisher’s exact tests and employing the Bonferroni correction for multiple comparisons computed with R package *rcompanion* version 2.0.0 (Mangiafico 2018). Last, we compared the average distribution for each scenario, i.e., the average of each Q score class of 10%, to the observed distribution with similar Fisher’s exact tests. We then tested for each scenario if the ratios of the F1s to backcrosses (F1/BC, where backcrosses are all admixed individuals except

F1s) were distinct from the observed data. When the proportion of simulations with higher ratios than the observed data was higher than 0.05, we considered that there was no significant difference between the observed data and the distribution generated by the simulations. All simulation scripts available at http://github.com/catpinho/admixture_simulation.

Geographic cline analyses

To perform analyses of geographic clines, we used the R package HZAR (Derryberry, Derryberry, Maley, & Brumfield, 2014). This program provides functions to fit allele frequency data to equilibrium geographic cline models (Barton & Gale, 1993; Gay, Crochet, Bell, & Lenormand, 2008; Szymura & Barton, 1986, 1991) using the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm. To each sampling location was assigned a distance along the transect that corresponds to the shortest distance to POR population along the coastline (i.e., avoiding a straight line over the ocean). Allele frequencies at each SNP locus included in the "80/20" dataset were fitted to 16 equilibrium geographic cline models using HZAR version 3.0.3 (Derryberry et al., 2014). We further used the Q_C values from the Admixture analysis as hybrid indices from the complete (*HI*) and diagnostic datasets ($HI_{DIAG\text{-}ADMIX}$) and also the mtDNA locus (*P. bocagei* or *P. carbonelli* maternal lineage) to fit to the same 16 models. Details about estimated parameters are provided in the Supplemental Information File 1 (Methods). We performed three independent runs of 1,000,000 MCMC iterations, a burn-in of 100,000 iterations and sampling every 10 iterations for each model and checked for convergence. For each SNP locus, HI, $HI_{DIAG\text{-}ADMIX}$, and mtDNA, the model with the lowest AIC score was selected as the best-fitting model. We then estimated the CI as the region delimited by the maximum and minimum values of the 95%

credible cline region (Derryberry et al. 2014). To test if each SNP locus significantly differed from the genome average, we compared its centre (*c*) and width (*w*) with the same parameters for the *HI* cline (c_{HI} and w_{HI}), which we took as a sort of a "genome-wide average". We considered each parameter for individual SNP clines to exhibit significant differences from the *HI* cline if their 95% CI did not overlap. We also searched for outlier loci based on the locus-specific estimates of *c* and *w*. Finally, to test if the number of SNP loci with centres shifted to north or south were similar or significantly different, we performed a binomial test. We did not try to estimate selection and dispersal using the approach of (Barton & Gale, 1993) as this assumes random mating within the hybrid zone, an assumption that seems unlikely given the low number of F1 hybrids in our sample (see below).

Genomic clines analyses

We quantified patterns of genomic introgression in admixed populations using the BGC software version 1.0 (Gompert & Buerkle, 2012). We performed preliminary analysis including the total number of loci and individuals, but to decrease the confounding effect of population structure (see Discussion), we report the results including only *P. bocagei* and *P. carbonelli* individuals occurring in syntopy, using the complete dataset. BGC uses a Bayesian genomic cline model (Gompert & Buerkle, 2011, 2012; Gompert et al., 2012a) to estimate ancestries of individuals summarized as a hybrid index (HI_{BGC}) and to quantify genome-wide variation in introgression among admixed populations. The model includes parameter α and β that describe, for every locus, the transition in allele frequency from the genomic background of origin (ancestry) to the genomic background of the other species, along the gradient of admixed genotype classes (i.e., HI_{BGC}) between the two parental species (Gompert & Buerkle, 2011).

Details about each parameter are provided in Supplemental Information File 1 (Methods). Despite the potential interest in estimating these parameters, simulation studies have shown that deviations in α relative to genomic average can be explained by underdominance when dispersal is low, directional selection or BDMI, whereas β can be affected by underdominance when dispersal is high or by BDMI (Gompert & Buerkle, 2011). However, BDMI have limited effect on genomic cline parameters (Gompert & Buerkle, 2011). As such, multiple selection types can result in similar patterns of α and β making its interpretation difficult and hence the interpretation of genomic clines should be made with caution.

A priori knowledge of pure parental individuals is required to estimate the proportion of the ancestry attributed to each parental species. We used the populations POR and MAD as *P. bocagei* parental individuals and ESM and TOR as *P. carbonelli*, while using ESP as the test (admixed) population. With the complete dataset we ran three independent MCMC chains for 300,000 steps each and recorded samples from the posterior distribution every 25th step following a 200,000-step burn-in to estimate marginal posterior probability distributions for HI_{BGC} and cline parameters α and β. We combined the output of the two chains after inspecting the MCMC output to assess convergence to the stationary distribution.

The significance of individual α and β values was assessed by checking whether their 95% CI included zero or not. These estimates provide a locus-specific view on the process of introgression between species, but they do not compare how genomic clines differ between loci. Therefore, in addition we searched for outlier loci on the basis of the genome-wide distribution of locus-specific estimates of α and β values by determining the cline parameter quantiles as recommended by Gompert & Buerkle (2011, 2012).

355

356 *Comparison between geographic and genomic clines*

357 To understand whether geographic and genomic cline parameters show
358 concordance in the patterns of reproductive isolation, we tested if the 2300 loci
359 common to "80-20" and complete datasets are correlated, using Person's correlation
360 tests implemented in R. Moreover, we checked whether the same loci showed
361 increased patterns of introgression isolation to compare the consistency of the results
362 obtained with both approaches.

363

364 *Distribution of cline parameters across the genome and "Large Z effect"*

365 To identify the position of each SNP on the genome we aligned each of the 6905
366 RADtag sequences from the complete dataset, including the 2300 "80/20" SNPs, with
367 the *Podarcis muralis* genome (Andrade et al., 2019). *Podarcis muralis* is the closest
368 relative to the Iberian and North African clade of *Podarcis* (the *Podarcis hispanicus*
369 complex) to which *P. bocagei* and *P. carbonelli* belong; they diverged around 10 million
370 years ago (Yang et al., 2021). The alignment was performed using the blastn algorithm
371 (Altschul et al., 1997) included in the BLAST+ package version 2.9.0 (Camacho et al.,
372 2009).

373 If many loci more or less equally distributed across the genome contribute to
374 species barrier, it is expected that levels of introgression are positively correlated with
375 recombination rate (Aeschbacher, Selby, Willis, & Coop, 2017; Martin, Davey, Salazar,
376 & Jiggins, 2019). Here we used chromosome length as a proxy to recombination rate
377 as chromosome length is negatively correlated with average recombination rate. We
378 tested if cline parameters are correlated with chromosome size using Person's

correlation tests on the “80-20” (for geographic cline parameters) and complete (for genomic cline parameters) datasets using the SNPs that aligned to *P. muralis* genome.

Furthermore, we tested if the distribution of values from each parameter used to describe introgression, from both geographic (*c* and *w*) or genomic (*α* and *β*) cline analyses, were similar between chromosomes and mtDNA, using non-parametric Kruskal-Wallis tests. If the test was significant for any of the parameters, we then performed pairwise Wilcoxon tests among chromosomes using a Bonferroni correction for multiple comparisons. These tests were performed with the R package *stats* version 4.0.3 (R Core Team, 2020). In a scenario of a “large Z-effect”, we expect introgression on Z-linked loci to be significantly lower than on autosome loci.

Results

SNP and mtDNA datasets

The final dataset consisted of 6905 SNPs (complete dataset) with an average depth of coverage of 28 and median of 26.8 (min = 12.5; max = 85.8) and 329 individuals with an average depth of coverage of 29 and median of 28.8 (min = 15.6; max = 48.6). Using this dataset, we assessed the replicability to be on average 99.87% for the tested samples. Additionally, we filtered the complete dataset to obtain two additional datasets, the “80/20” dataset containing of 2300 loci with allele frequencies higher than 0.8 in reference populations of one species and lower than 0.2 in the other, and the diagnostic dataset containing 1241 loci with fixed frequencies in each parental population. Finally, for mtDNA dataset, we successfully sequenced the 12S fragment from 336 individuals. Details about the final sample size for each dataset is provided in Supplemental Information File 1 (Table S1).

Population structure across the transect

As expected from previous works (see Caeiro-Dias et al., 2021; Pinho et al., 2009), the current contact zone between *P. bocagei* and *P. carbonelli* was located in ESP population and was mostly made of parental individuals (strongly bimodal hybrid zone; Figure 2): Admixture identified 147 individuals as parental from which 114 had $Qc \geq 0.9$ and 33 $Qc \leq 0.1$, including 66 with *Qc* of one and 21 with Q_C of zero (Figure 2a). The remaining 14 individuals captured in the contact zone showed signs of large proportion of admixture ($0.1 < Q_C < 0.9$). Actual syntopy was restricted to a ~450 meters wide stripe in ESP including 115 samples. Based on multilocus assignment results, this length was delimited by the most northern *P. carbonelli* with $Qc \geq 0.9$ and the most southern *P. bocagei* with $Qc \leq 0.1$. These proportions are commonly used as thresholds to distinguish parental from admixed genotypes. Within the syntopy area, 102 individuals were assigned to parental genotypes, 90 with a genomic background assigned to *P. carbonelli* and 12 to *P. bocagei*, including 52 with $Q_C = 1$ and three with $Q_C = 0$ (Figure 2b). The binomial test ($p<0.001$) shows that the number of samples in this area is biased towards *P. carbonelli*, suggesting a higher local density of this species since nothing indicated higher capturability. The other 13 individuals in the syntopy area showed signs of recent mixed ancestry. Summary results for *Qc* scores *per* population are presented in Supplemental Information File 1, Table S1. Accordingly, the first PCA axis (PC1) explains 57.1% of the variation in the data and clearly distinguishes *P. bocagei* and *P. carbonelli* populations (Figure 2c). PC2 explains 1.4% of the variation and corresponded to genetic structure within *P. carbonelli*, separating the ESP population from all other *P. carbonelli* populations. In fact, when we increased K with Admixture, K=3 (Figure 2a) was the best scenario to explain the variation in the SNP data. In concordance with the PCA, *P. carbonelli*

individuals from ESP were mostly assigned to a group apart from the other *P. carbonelli* populations showing sub-structuring within *P. carbonelli*.

The plot of *He* against *HI* (Figure 2d) showed results concordant with Admixture and PCA. One individual was close to the expectations for first-generation BC and three others were close to the expectations for F1 hybrids, although *He* was lower than 1. False homozygote call due to PCR duplicates is a known source of bias across RADseq protocols (for discussion on this see Andrews et al., 2014; Puritz et al., 2014) and would reduce *He* even in F1 hybrids. While the single-end reads ddRADseq protocol used in this study did not allow us to detect PCR duplicates (Andrews et al., 2014; Puritz et al., 2014), we performed four independent PCR for each sample, to decrease the proportion of PCR duplicates in the final library. Moreover, we were able to detect two F1s with interspecific *He* = 1 in a contact zone between *P. carbonelli* and *P. lusitanicus* using the same protocol (Caeiro-Dias et al., 2021). An alternative explanation is that those three potential F1s were a product of crosses between parental individuals from different species with residual levels of admixture, which is a plausible scenario since in the contact zone we observed a continuum of admixed proportions. While our results can be explained by one of these scenarios or by a combination of both, the observed *He* and *HI* values for these three individuals were closer to F1s than any other hybrid class and thus they were considered as F1 hybrids. Those three individuals were adult males, all with *P. bocagei* mtDNA. The potential first-generation BC was also identified as an adult male with mtDNA assigned to *P. carbonelli*.

An admixture signal was also evident in SIL population south to ESP since all individuals in SIL were identified as *P. carbonelli* on the basis of morphology and mtDNA but exhibited 0.02 to 0.05 of admixture from *P. bocagei* in SNP data (for K=2,

Fig. 2a); SIL was the only population away from the contact zone with evident signs of admixture.

Simulations of genotypic composition within the contact zone

Simulations suggested that most combinations of RM or NRM (in our model it was restricted to parental individuals) with various migration rates were unable to produce distributions of genotype classes matching our observed distribution. Only for NRM with 40% of parental genotypes migrating into the syntopy zone the simulations were not significantly different from our data (the corrected *p*-value = 0.461 was higher than 0.05; Supplemental Information File 1, Table S2). Without higher levels of migration, with either RM or NRM, F1 hybrid genotypes were no longer observed in a few generations.

The proportions of F1s relative to backcrosses for most scenarios with RM and NRM is lower than in observed data (Figure 3). Note that the relative proportions of F1s was the same in the observed data from the syntopy area and simulated data. The observed proportion was not significantly different in only three scenarios (RM0.3, RM0.4 and NRM0.4; Supplemental Information File 1, Table S3). Such results suggested less backcrosses than expected in the observed data given the number of F1 hybrids and hence a reduced fitness of F1 hybrids.

Geographic clines

The fixed/right geographic cline model (p_{min}/p_{max} fixed to zero and one with right tail fitted, i.e., estimated tail south of the hybrid zone towards *P. carbonelli*) was the best-fit model for the *HI* cline, with c_{HI} = 13.15km (CI = 11.98–14.04km) south of POR (ESP; Figure 4) and w_{HI} = 0.54km (CI = 0.1–3.18km) indicating a very narrow hybrid

zone. Tail parameters δ and τ were 0.43km and 0.238, respectively. The best-fitted models varied across the 2300 loci (see Supplemental Information File 1, Table S4) but most of the geographic clines for each SNP locus were highly concordant in both *c* and *w* (Figure 4a and Supplemental Information File 1, Table S5) and not significantly different from the *HI* estimates (Table 1). Only 3.2% of SNPs have a cline centre that is shifted north and 2.1% shifted south. The binomial test revealed that these proportions are significantly different (p<0.001) from random expectations. From those SNPs with *c* significantly shifted, we detected six outlier loci shifted to north but none to south (Figure 4a). Also, just 8.1% of the loci had significantly wider clines than the genomic average but none had a narrower cline. Detecting statistically significant lower *w* was not possible since w_{HI} was already very narrow and so was the lower margin of the confidence interval (w_{HI} = 0.54km, CI = 0.1–3.18km). The regression model estimated for *c* versus *w* suggested that there a significant linear correlation between these two parameters, i.e., clines with *c* shifted to north tended to have larger *w* (Figure 4a). Another aspect was that while for 59.4% of the loci the best fitted model did not include tail parameter estimations, for 28% the model included a right tail (to south), which was also reflected in the none/right model for the *HI*. From these loci, 98% were mapped to autosomes. On the other hand, the proportion of loci fitted to a model with left tail (to north) was only 3%, and from these, 95.7% were autosomal loci. For the $HI_{DIAG\text{-}ADMIX}$ the best-fitted model was the none/mirror (p_{min}/p_{max} not estimated and symmetric tails), with $c_{\text{HI-DIAG-ADMIX}}$ = 13.55km (CI = 13.18–13.92km) south of POR and $w_{\text{HI-DIAG-ADMIX}}$ = 1.99km (CI = 0.74–3.74km). Thus, both *HI* estimation did not produce significantly different cline centres. Finally, for the mtDNA cline the none/none (tails and p_{min}/p_{max} not estimated) was the best-fitted model, with c_{mtDNA} = 14.12km (CI = 13.06–14.14km) and w_{mtDNA} = 0.1km (CI = 0.01–3.74km). The mtDNA cline centre had

overlapping CIs with both *HI* estimates and hence the nuclear and the mtDNA clines were not statistically different.

Genomic clines

Preliminary results based on the total number of individuals sampled in the contact zone revealed the same overall pattern of α and β distribution. The main difference is that the parameter β shows more extreme positive values which conforms to the expectations in the presence of marked population structure. We thus focus our analysis in the syntropy area. The genomic clines within the syntopy area showed high heterogeneity across the 6905 loci ($-8.03<\alpha<5.43$; $-18.70<\beta<24.97$) with many loci significantly deviating from the expectations given the *HI* estimated from the complete dataset (Table 1, Figure 4b and Figure 5). We detected a significant excess of *P. bocagei* ancestry (lower bound of 95% CI for $\alpha > 0$) in 15.2% of the loci and an excess of *P. carbonelli* ancestry (upper bound of 95% CI for $\alpha < 0$) in 32.3%. These proportions were significantly different from chance expectations (binomial test $p<0.001$). A total of 52.5% of the loci did not present an excess of ancestry for any species given the *HI* (their 95% CI for α includes 0). We also detected 18.7% of loci with significantly steeper changes in allele frequencies (lower bound of 95% CI for $\beta > 0$), 16.6% with significantly shallower changes in allele frequencies (upper bound of 95% CI for $\beta < 0$) and 64.7% following the expectations given the *HI* (95% CI for β includes 0). From the 1053 loci with excess of *P. bocagei* ancestry, 19.5% follow the expectations given the HI_{BGC}, 8.8% exhibit steep changes in allele frequencies and 71.7% showed signs of increased introgression. For the 2227 loci with excess of *P. carbonelli* ancestry, 48% follow the expectations given the HI_{BGC}, 50.4% of the loci showed restricted

introgression and only 1.7% had shallower changes in allele frequencies. No outlier loci were detected for the observed distributions of *α* and *β*.

Comparison between geographic and genomic clines

By comparing geographic and genomic clines from the 2300 loci common to "80-20" and complete datasets, we detected a negative but week correlation between c and α (Figure 5a) and between w and β (Figure 5b). However, many loci have cline parameters that contrast with the significant correlations. For example, a considerable proportion of loci, despites increasing introgression in the syntopy zone (significantly negative β; 17.3%) or no differences from the genomic average (β non-significant; 29.6%), did not show significant w differences from the w_{HI}. Moreover, 18.9% of loci show significant negative β, i.e., increased introgression when compared to the rest of the genome, and the majority of those loci (17.4% of the total) had narrow w similar to w_{HI}. Only the remaining 1.6% of loci showed congruent patterns of significant introgression.

Additionally, we found inconsistent patterns of loci exhibiting increased introgression in the alternative genomic background between both approaches. While geographic clines showed 2.6% of loci with c shifted to north and w wider than w_{HI}, only 0.1% of the loci showed excess *P. carbonelli* ancestry with higher introgression, and neither set had loci in common. Conversely, 0.5% presented c shifted to south and wide clines, whereas 0.4% had excess *P. bocagei* ancestry and higher introgression but only 0.1% (three loci) were common to both sets. These results show that both approaches identify different sets of loci with alleles that might be under positive selection in the foreigner genomic background.

Distribution of cline parameters across the genome

We successfully aligned on the *P. muralis* genome 6027 RADtags (out of the 6905 from the complete dataset) and 2101 (out of the 2300 present in the "80/20" dataset), representing a density of 1 marker per 236Kb and 1 marker per 678Kb, respectively.

We did not identify significant correlations between any cline parameter and chromosome length. When testing using geographic cline parameters, for *c* R^2 = 0.01 and *p*-value = 0.68; for *w* R^2 = 0.21 and *p*-value = 0.05. When using genomic cline parameters, for *α* R^2 = 0.05 and *p*-value = 0.36; for *β* R^2 = 0.01 and *p*-value = 0.57. Thus, levels of introgression do not seem to be correlated with recombination rate.

The distributions of geographic and genomic cline parameters and the densities of values across each chromosome are represented in Figure 6. These plots, particularly the violin plots depicting density values (Figure 6b-e), show striking differences between the Z chromosome and the autosomes. Specifically, in the Z chromosome *c* and *w* are less variable, *w* is noticeably lower, *α* is skewed towards *P. carbonelli* ancestry (negative values), and *β* has a more marked bimodal distribution than the autosomes. The difference in the geographic cline parameter *c* was small but significant between the Z chromosome (mean = 13.84km) and all autosomes (mean = 13.37km; Table 2 and Supplemental Information File 1, Table S6), while the distribution of *w* was significantly different between the Z chromosome and all autosomes but one, the chromosome 18 (Table 2 and Supplemental Information File 1, Table S7). However, it is worth to note the lack of outliers in chromosome 18 (Figure 6c) and that the distribution of *w* in this autosome was not statistically different from the other autosomes. For the genomic cline parameter *α* we found significant differences only between Z and three other chromosomes (Table 2 and Supplemental Information File

1, Table S8), while for *β* all comparisons between the Z chromosome and all autosomes were significant, except with chromosomes 8 and 18 (Table 2 and Supplemental Information File 1, Table S9). Overall, these results suggest lower introgression in Z-linked loci.

Discussion

Genomic composition in the hybrid zone

As shown by previous works (Caeiro-Dias et al., 2021; Pinho et al., 2009), the hybrid zone between *P. bocagei* and *P. carbonelli* is very narrow and made mostly of parental genotypes. Nearly half of the individuals in the syntopy area did not show any evidence of admixture, although the proportions of parental individuals are largely biased towards *P. carbonelli*. However, the other half of the individuals presented a gradient of admixed proportions including three F1 hybrids, recent and later generation backcrosses, and residual levels of admixture. This continuum of assignment proportions across admixed genotypes shows that the two species hybridize and produce fertile offspring resulting in recurrent backcrosses over many generations, as previously demonstrated in Caeiro-Dias et al. (2020). Moreover, we did not find any mito-nuclear discordance outside of the contact zone.

Outside of the contact zone, substantial signs of admixture between both species were almost undetected, except in SIL population. The similar low frequencies of *P. bocagei* assignment across all individuals in SIL suggest that after gene flow between this population and ESP in the past, these populations may now be isolated, or gene flow has been highly restricted for several generations.

Within *P. carbonelli,* the individuals from ESP exhibit a genetic composition different from the other *P. carbonelli* populations across the transect, independently from *P. bocagei* admixture. This genetic structure within *P. carbonelli* suggests differentiation between ESP and southern populations. A comprehensive examination of population structure within the range of *P. carbonelli* would be needed to interpret this result.

Reduced hybrid fitness

In our simulation study, most of the simulated distributions of admixture classes do not match the data observed in the syntopy area. The simulated scenarios were similar to the observed data only when we implemented high migration levels (i.e., the replacement of 40% of the individuals in the contact zone by migrants in each generation). Given that the syntopy range is small (about 450 m) this scenario is possible but extreme. In other scenarios with less dispersal or without assortative mating (irrespective of dispersal), the effects of selection are necessary to explain our results. When compared to the simulated data, the frequency of BC relative to the number of F1s in the contact zone was low, suggesting a combination of lower fecundity of F1 and/or lower survival of BC individuals. Pinho et al. (2009) failed to detect abnormalities of the gonads or reduced number of spermatozoa in admixed individuals of this contact zone but we don't have direct estimates of fecundity or survival so we cannot conclude on the selective processes acting on hybrids there. Furthermore, although the three F1s are males with *P. bocagei* mtDNA, the BC male with *P. carbonelli* do not offer any clue suggesting that one sex may be less fit than the other since it was necessary a female F1 with *P. carbonelli* mtDNA to originate the BC. We recognize that assortative mating can indeed be stronger than what was

implemented in our model and may play a more important role than what our results suggest. In conclusion, strong pre- or post-mating barriers are necessary to explain the low number of F1 and BC we see in the contact zone, but we cannot yet decipher their relative contributions, although they both likely contribute to reproductive isolation in this system. To fully evaluate relative the importance of pre-mating isolation, low survival or low fertility of F1 and low survival of BC on reproductive isolation in this system, we would need larger sample sizes of various genotype classes combined with simulations incorporating mode complex and realistic assortative mating scenarios (e.g., Irwin, 2020).

Strong selection against introgression maintains a restricted hybrid zone

The geographic cline analyses and the estimation of selection in the centre of the contact zone provided further evidence of strong selection against gene flow between the two species. The geographic clines show a high proportion of loci (91.9%) with steep clines, indicating the presence of strong barriers to gene flow spread over the genome between populations of both species apart from the contact zone (Mallet et al., 1990), and most of the clines (89.7%) had approximately the same cline centre and width. These patterns are coherent with a tension zone model where multiple unlinked loci form concordant steep clines at the centre of the hybrid zone (Barton, 1983; Barton & Gale, 1993; Barton & Hewitt, 1985). In a stage where substantial barriers already exist, coupling of barrier effects can be achieved by processes like spatial movement of clines (Bierne et al., 2011) or cycles of population expansion and contraction (Hewitt, 1989), leading to concordant steep clines in the late stages of speciation. Both spatial movement of clines and population fluctuations can explain

current genetic variation across Iberian *Podarcis* (Pinho, Harris, & Ferrand, 2007; Sá-Sousa, 2001).

Differential introgression patterns across the genome detected within the contact zone

In contrast to the overall concordance among loci detected with geographic cline analysis, genomic clines show larger heterogeneity of introgression patterns among loci within the syntopy area, particularly for *β*. However, positive values of *β* may be inflated in the presence of population structure within the hybrid zone (Gompert & Buerkle, 2011, 2012; Gompert et al., 2012a). *Podarcis carbonelli* individuals used as reference (TOR) and those from the contact zone (ESP) were clustered in different groups with PCA and Admixture and thus true *β* values may be lower than our estimations. Nevertheless, genomic cline results indicate that most of the loci do not introgress more or less that what would be expected given the HI_{BGC}, which can be explained because most loci maintain a strong linkage with the other loci on the genomic background of origin (Gompert et al., 2012b), preventing intraspecific gene flow. Epistatic interactions such as BDMI are considered to be the most common intrinsic barriers to gene flow (Coyne & Orr, 2004; Qvarnström & Bailey, 2009) and in the early stages of speciation are expected to generate patterns of differential introgression across the genome because of stochasticity in the accumulation of such incompatibilities (Brennan, Hiscock, & Abbott, 2014; Fishman, Kelly, Morgan, & Willis, 2001; Scascitelli et al., 2010). However, in the late stages of speciation enough BDMI should have build-up across the genome to confer strong reproductive isolation. Across *P. hispanicus* complex allopatric divergence is likely the best explanation for climatic niche divergence (Caeiro-Dias et al., 2018) and thus the patterns of reproductive isolation across the genomes of *P. bocagei* and *P. carbonelli* may be related with

676 divergence in allopatry, where a large portion of the genome may have become
677 differentiated through selective or neutral processes (Fierst & Hansen, 2010; Gavrilets,
678 Hai, & Vose, 1998; Gompert et al., 2012b; Lynch & Force, 2000). The evolution of
679 current strong reproductive isolation likely involved different processes, interactions
680 between processes and potential new interactions arising along the evolution of
681 speciation (Kulmuni et al., 2020), More explicitly, intricated interactions between
682 genetic architecture of intrinsic barriers to gene flow, populations history, and selected
683 and neutral variation, might be contributing to the strong reproductive isolation (Bierne
684 et al., 2011; Butlin & Smadja, 2018; Kulmuni et al., 2020) but disentangling the effect
685 of each process is not easy. Whilst genomic clines are of limited usefulness to detect
686 pairwise epistatic loci (Gompert & Buerkle, 2011; Gompert & Buerkle, 2009), our
687 results show that in the late stages of speciation, most loci do not show signs of
688 increased introgression compared to genomic average, consistent with the hypothesis
689 that the genomic architecture of reproductive isolation is determined by many widely
690 dispersed genomic regions.
691
692 *Comparison of geographic and genomic clines*
693 A direct comparison of geographic and genomic cline parameters between loci
694 common to both analyses detected significant but week negative correlations.
695 Significant correlations between *c* and *α* were driven mostly because some loci with
696 excess of *P. carbonelli* ancestry have very high values of *c* (shifted to north). When
697 comparing *w* and *β*, significant results were obtained mostly because almost half of the
698 loci (44.9%) had steep clines (not significantly different from w_{HI}) and decreased
699 introgression in the syntopy zone (significantly positive *β*), showing a concordance in
700 the lack of introgression between both cline analyses. However, week correlations can

be explained because regardless of the variation observed in the genomic cline parameters, the geographic cline parameters were consistent for most loci denoting a strong reproductive isolation and a lack of substantial introgression outside of the contact zone. A particularly interesting result was that 17.3% of the loci showed wide genomic clines but narrow geographic width raising the question whether extrinsic selection might play a role constraining interspecific gene flow outside of the contact zone. While it is a plausible explanation because the two species exhibit significantly different climatic niches (Caeiro-Dias et al., 2018) it need further research.

Additionally, we found inconsistent patterns of loci potentially under positive selection between both approaches. Geographic and genomic clines identified, respectively, 2.6% and 0.1% of loci with *P. carbonelli* alleles with increased introgressing in *P. bocagei* genomic background, but none of the loci are common to both sets. Conversely, the same approaches identified, respectively, 0.5% and 0.4% of loci with *P. bocagei* alleles with increased introgressing in *P. carbonelli* genomic background but only 0.1% were loci identified by both methods. Loci identified by genomic clines in the syntopy zone exhibiting increased introgession are not expected to necessarily introgress away from the contact zone, but at least the loci with signs of increased introgression in populations apart from the contact zone should show consistent patterns of introgression with both approaches, regardless of what selective forces are contributing to the geographic patterns of introgression detected. While these results could be explained by *P. carbonelli* population structure, as it is expected to potentially inflate the signal of decreased introgression detected with genomic clines within the syntopy zone (Gompert & Buerkle, 2011, 2012; Gompert et al., 2012a), it is not clear in the available literature to date how to undoubtedly interpret such differences. In particular, because genomic cline analyses may be sensitive to many

distinct selective and demographic phenomena (Gompert & Buerkle, 2011) we cannot rule out the hypothesis that the differences suggested by both approaches are led by these confounding effects. We thus suggest that additional testing on simulated and empirical datasets using genomic cline approaches are needed to disentangle the causes of inconsistent results using different methods.

Genomic signatures of a moving hybrid zone and potential positive selection northwards

Results show that several loci introgress asymmetrically. About 28% of geographic clines were fitted to a model with an estimated tail to south while only 3% were fitted to a model with a tail estimated to north. In both cases, the large majority are autosome loci. These results are consistent with a scenario of a moving hybrid zone northwards in the direction of *P. bocagei*, since a moving hybrid zone is expected to leave a tail of clines of unlinked neutral markers resulting in asymmetrical cline variation (Buggs 2007; Teeter *et al.* 2010), essentially driven by autosomes. Although this pattern could alternatively be interpreted as unidirectional introgression away from a static hybrid zone (Moran, 1981), the consistency in the proportions of assignment to *P. bocagei* across SIL individuals suggests that the northward movement of the hybrid zone left *P. bocagei* alleles behind, and currently the frequencies are homogenized. The overall genomic clines pattern is also concordant with this interpretation; 13.7% of the loci with increased *P. bocagei* ancestry and increased introgression is considerably higher than the 4.4% of loci with restricted introgression, indicating that several loci with excess *P. bocagei* ancestry can introgress in the genomic background of *P. carbonelli*. On the other hand, the 1.2% of loci with increased *P. carbonelli* ancestry and increased introgression is much lower than 15.8%

with restricted introgression, showing that many loci with increased *P. carbonelli* ancestry do not introgress in the genomic background of *P. bocagei*.

Interestingly, a restricted group of loci exhibit an opposite pattern: the 2% of loci have *c* shifted to south on average 1.65km from the *HI* cline centre, while 3.2% of clines with *c* shifted to north on average 5.03km. These proportions are statistically different. Also, clines with centre shifted to north tend to be wider than other clines (both parameters are positively correlated), wider clines than the *HI* cline, and include outlier loci. Loci with *c* shifted from the majority of other clines have been interpreted as being subject to positive selection on one side of the hybrid zone (Baldassarre, White, Karubian, & Webster, 2014; Ballard & Whitlock, 2004; Barton & Hewitt, 1985; Teeter & Payseur, 2008). It is reasonable to hypothesise that loci with *c* shifted to north are good candidates for harbouring genes with *P. carbonelli* alleles under positive selection that confer advantage in the genomic background or habitat of *P. bocagei*. Note, however, that genetic drift in hybrid zones can increase false positives when testing for adaptive introgression (Jofre & Rosenthal, 2021) and affect clines shape (Polechová & Barton, 2011). Drift is higher in smaller demes with lower migration rates among demes. In our species, dispersal is probably highly restricted among local populations (small sized terrestrial nonflying vertebrate in fragmented environment), but local populations are apparently quite large, which should prevent strong drift. Nevertheless, we cannot entirely discard at this stage that drift rather than selection generated variation in cline centre among loci. Simulations or direct tests of selection would be necessary to undoubtably rule out the effect of drift.

Distribution of cline parameters across the genome and "Large Z effect"

We did not find any correlation between chromosome length and cline parameters, suggesting that barrier loci are distributed across the genome and show little variation in introgression levels. In the late stages of speciation, it is expected that loci are in such strong linkage with the genomic background of origin that even a high recombination rate is not enough to restore linkage in admixed genotypes.

The *w* of loci on the Z chromosome was significantly smaller on average than that of most autosomes (the only exception being chromosome 18). Similar results were obtained for *β*, that was significantly smaller in the Z than in most of autosomes, except chromosome 8 and 18. Despite non-significant differences between Z and chromosomes 8 and 18, these two chromosomes do not show consistent significant differences when compared to other autosomes. We interpret these results as a consistent signature of lower introgression in the Z chromosome compared to the autosomes. Such differences are expected when the sexual chromosomes have a distinct role in the reproductive isolation compared to autosomes and are concordant with the “large X effect” (or “large Z effect” in the case of the ZW sex determination system) which is a long-standing principle of speciation and has been identified to play a role in introgression patterns and reproductive isolation across many species (e.g., Dufresnes et al., 2021; Guénet, et al., 1990; Jiggins et al., 2001; Llopart, 2012; Orr, 1987; Storchová et al., 2010).

However, differences in effective population sizes between autosomes and Z/X chromosomes are also expected to generate steeper clines (higher slope, lower width) for Z/X compared to autosomes as smaller effective size result in stronger effect of drift (Polechová & Barton, 2011) even if there is no difference in selection. Differences in slopes due to drift are difficult to predict, as they depend on selection, dispersal, and effective population size, with strong selection, strong dispersal and large population

size resulting in low drift and similar slopes for autosomes and sex chromosomes. In our hybrid zone, we inferred strong selection, and neither dispersal nor population size appear to be restricted. We thus discard drift as the main explanation for the steeper Z clines compared to autosomes.

Lasne et al. (2017) reported that clines generated by local adaptation are steeper for loci on the X or Z chromosomes than for loci on autosomes, but in this case the cause of the systematically higher slope for the X/Z is that the X/Z is more responsive to selection than the autosomes are (T. Lenormand and T. Connallon, pers. com.). This effect only applies to primary clines generated by selection of alternative alleles across an environmental change in space and does not affect tension zones maintained by a balance between dispersal and selection again hybrids/intergrades, as the slope of clines in tension zones only depends on selection and dispersal (Barton & Gale 1993), both of which are independent of effective population size.

Another possible explanation for steeper Z clines would be female-biased dispersal, as a higher dispersal of the heterogametic sex tend to increase the difference in slopes between Z/X and autosomes. However, in squamates, males tend to disperse more (see Ferchaud et al., 2015 and references therein), including an example in the *Podarcis* genus (Vignoli, Vuerich, & Bologna, 2012). Moreover, mtDNA geographical cline parameters *c* and *w* were not statistically different from the nuclear *HI* cline parameters, suggesting that dispersion is not female-biased. Even if there is no information on sex biases in dispersal in our focal species, we can thus probably discard sex-biased dispersal as an explanation here for the steeper Z clines compared to the neutral expectation relative to the autosomes. Unequal sex-ratio and differences in the variance of reproductive success can also lead to deviations from neutral expectations for Z and autosome cline slope ratio. While sex ratio in our dataset is not

different from 1:1 (observed 1 female to 1.2 males; X^2 = 0.703; *p*-value = 0.402), differences in reproductive success are not possible to exclude.

Based on our results, steeper Z clines are thus likely the result of "large Z-effect". Furthermore, we detected a discordance in *c* between Z and autosomes. The geographical cline centres of Z-linked loci are in general displaced to the south (essentially matching with the area of syntopy) from autosome cline centres (generally matching with the c_{HI}, slightly to the north of the syntopy area). Although this difference is only of a few hundred meters, it is nevertheless significant. In the context of a moving hybrid zone towards north (previously discussed), the displacement of Z cline centres to the south is consistent with the scenario of "large Z effect" where Z cline centres essentially match the centre of the contact zone and do not show signs of introgression.

Concluding remarks

The strong bimodality of the hybrid zone between *P. bocagei* and *P. carbonelli* and the patterns of strong reproductive isolation between these two deeply divergent species confirm that these species have reached the late stages of speciation. Accordingly, we found multiple lines of evidence of selection against hybrids. However, even in this late stage of speciation, a relatively small proportion of the loci can still introgress more than the genomic average, but such introgression is mostly asymmetric and suggests the hybrid zone is moving towards *P. bocagei* distribution range. Furthermore, we found evidence that the Z chromosome plays a disproportionally large role in reproductive isolation. Overall, our results support a model of increasingly numerous barriers as speciation proceeds resulting in a genomic architecture of reproductive isolation in the late stages of speciation that relies on cohesion across the genome against interspecific gene flow determined by multiple interacting barriers.

Acknowledgments

We thank E. Pérez, M. A. Carretero, N. Sillero, A. Kaliontzopoulou, A. Perera, A. Crotini and A. Salas for helping with sample collection and Thomas Lenormand, Luis-Miguel Chevin, Tim Connallon and Clémentine Lasne for their help with cline analyses. Molecular Ecology's subject editor and two reviewers made highly valuable comments that significantly improved the manuscript. This study benefited from the Montpellier Bioinformatics Biodiversity platform supported by the LabEx CeMEB, an ANR *Investissements d'avenir* program (ANR-10-LABX-04-01). GCD was supported by a PhD grant (SFRH/BD/89750/2012), and CP by an IF contract (IF/01597/2014/CP1256/CT0009), under the *Programa Operacional Potencial Humano—Quadro de Referência Estratégico Nacional* funds from the European Social Fund and Portuguese *Ministério da Educação e Ciência* and PAC by the ANR grant ANR-19-CE02-0011 – IntroSpec. Support was also provided by national funds through FCT projects: PTDC/BIA-BEC/102179/2008 – FCOMP-01-0124-FEDER-007062, under FEDER COMPETE funds; PTDC/BIA-EVL/28090/2017 – POCI-01-0145-FEDER-028090 and PTDC/BIA-EVL/30288/2017 – NORTE-01-0145-FEDER-30288 co-funded by NORTE2020 through Portugal 2020 and FEDER Funds. Specimens for this study were captured and handled under permit numbers 486 to 490/2014/CAPT by *Instituto da Conservação da Natureza e das Florestas* (ICNF, Portugal).

References

Aeschbacher, S., Selby, J. P., Willis, J. H., & Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences*, *114*(27), 7061-7066.

875 Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm
876 for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246.
877 Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of
878 ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. doi:
879 10.1101/gr.094052.109.vidual
880 Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., &
881 Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
882 protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.
883 Andrade, P., Pinho, C., i de Lanuza, G. P., Afonso, S., Brejcha, J., Rubin, C.-J., …
884 Bellati, A. (2019). Regulatory changes in pterin and carotenoid genes underlie
885 balanced color polymorphisms in the wall lizard. *Proceedings of the National*
886 *Academy of Sciences*, *116*(12), 5633–5642.
887 Andrew, R. L., & Rieseberg, L. H. (2013). Divergence is focused on few genomic
888 regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*,
889 *67*(9), 2468–2482.
890 Andrews, K. R., Hohenlohe, P. A., Miller, M. R., Hand, B. K., Seeb, J. E., & Luikart,
891 G. (2014). Trade-offs and utility of alternative RADseq methods: reply to Puritz et
892 al. *Molecular Ecology*, 23, 5943-5946.
893 Baldassarre, D. T., White, T. A., Karubian, J., & Webster, M. S. (2014). Genomic and
894 morphological analysis of a semipermeable avian hybrid zone suggests
895 asymmetrical introgression of a sexual signal. *Evolution*, *68*(9), 2644–2657. doi:
896 10.1111/evo.12457
897 Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of
898 mitochondria. *Molecular Ecology*, *13*(4), 729–744.
899 Barton, N. H. (1983). Multilocus clines. *Evolution*, *37*(3), 454–471.

Barton, N. H, & Gale, K. S. (1993). Genetic analysis of hybrid zones. In Harrison, R. G. (Ed.). *Hybrid Zones and the Evolutionary Process*, 13–45. Oxford University Press.

Barton, Nicholas H, & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, *16*(1), 113–148.

Bateson, W. (1909). Heredity and variation in modern lights. In: Seward, A.C. (Ed.), *Darwin and Modern Science*. Cambridge University Press, Cambridge, 85-101.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brelsford, A., Dufresnes, C., & Perrin, N. (2016). High-density sex-specific linkage maps of a European tree frog (*Hyla arborea*) identify the sex chromosome without information on offspring sex. *Heredity*, *116*(2), 177.

Brennan, A. C., Hiscock, S. J., & Abbott, R. J. (2014). Interspecific crossing and genetic mapping reveal intrinsic genomic incompatibility between two *Senecio* species that form a hybrid zone on Mount Etna, Sicily. *Heredity*, *113*(3), 195.

Caeiro-Dias, G., Brelsford, A., Kaliontzopoulou, A., Meneses-Ribeiro, M., Crochet, P.-A., & Pinho, C. (2021). Variable levels of introgression between the endangered *Podarcis carbonelli* and highly divergent congeneric species. *Heredity*, 126, 463-476.

Caeiro-Dias, G., Luís, C., Pinho, C., Crochet, P. A., Sillero, N., & Kaliontzopoulou, A. (2018). Lack of congruence of genetic and niche divergence in *Podarcis hispanicus* complex. *Journal of Zoological Systematics and Evolutionary Research*. doi: 10.1111/jzs.12219

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 1–9.

925 Carling, M. D., & Brumfield, R. T. (2008). Haldane's rule in an avian system: using
926 cline theory and divergence population genetics to test for differential
927 introgression of mitochondrial, autosomal, and sex-linked loci across the
928 *Passerina* bunting hybrid zone. *Evolution*, *62*(10), 2600–2615.
929 Catchen, J. M. (2013). Stacks: an analysis tool set for population genomics.
930 *Molecular Ecology*, *22*(11), 3124–3140. doi: 10.1111/mec.12354.Stacks
931 Coyne, J. A., & Orr, H. A. (2004). *Speciation. Sunderland, MA*. Sinauer Associates,
932 Inc.
933 Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). HZAR:
934 hybrid zone analysis using an R software package. *Molecular Ecology*
935 *Resources*, *14*(3), 652–663.
936 Dobzhansky, T. G. (1937). *Genetics and the Origin of Species* (Vol. 11). Columbia
937 University Press.
938 Dufresnes, C., Brelsford, A., Jeffries, D. L., Mazepa, G., Suchan, T., Canestrelli, D.,
939 … Martínez-Solano, I. (2021). Mass of genes rather than master genes underlie
940 the genomic architecture of amphibian speciation. *Proceedings of the National*
941 *Academy of Sciences*, *118*(36).
942 Egan, S. P., Ragland, G. J., Assour, L., Powell, T. H. Q., Hood, G. R., Emrich, S., …
943 Feder, J. L. (2015). Experimental evidence of genome-wide impact of ecological
944 selection during early stages of speciation-with-gene-flow. *Ecology Letters*,
945 *18*(8), 817–825.
946 Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., …
947 Qvarnström, A. (2012). The genomic landscape of species divergence in
948 *Ficedula* flycatchers. *Nature*, *491*(7426), 756.
949 Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-

950 flow. *Trends in Genetics*, *28*(7), 342–350.
951 Feder, J. L., Gejji, R., Yeaman, S., & Nosil, P. (2012). Establishment of new
952 mutations under divergence and genome hitchhiking. *Philosophical Transactions*
953 *of the Royal Society B: Biological Sciences*, *367*(1587), 461–474.
954 Feder, J. L., & Nosil, P. (2010). The efficacy of divergence hitchhiking in generating
955 genomic islands during ecological speciation. *Evolution: International Journal of*
956 *Organic Evolution*, *64*(6), 1729–1747.
957 Ferchaud, A., Eudeline, R., Arnal, V., Cheylan, M., Pottier, G., Leblois, R., & Crochet,
958 P. (2015). Congruent signals of population history but radically different patterns
959 of genetic diversity between mitochondrial and nuclear markers in a mountain
960 lizard. *Molecular Ecology*, *24*(1), 192–207.
961 Fierst, J. L., & Hansen, T. F. (2010). Genetic architecture and postzygotic
962 reproductive isolation: evolution of Bateson–Dobzhansky–Muller incompatibilities
963 in a polygenic model. *Evolution: International Journal of Organic Evolution*,
964 *64*(3), 675–693.
965 Fishman, L., Kelly, A. J., Morgan, E., & Willis, J. H. (2001). A genetic map in the
966 *Mimulus guttatus* species complex reveals transmission ratio distortion due to
967 heterospecific interactions. *Genetics*, *159*(4), 1701–1716.
968 Fitzpatrick, B. M. (2012). Estimating ancestry and heterozygosity of hybrids using
969 molecular markers. *BMC Evolutionary Biology*, *12*(1), 131.
970 Gavrilets, S., Hai, L., & Vose, M. D. (1998). Rapid parapatric speciation on holey
971 adaptive landscapes. *Proceedings of the Royal Society of London B: Biological*
972 *Sciences*, *265*(1405), 1483–1489.
973 Gay, L., Crochet, P. A., Bell, D. A., & Lenormand, T. (2008). Comparing clines on
974 molecular and phenotypic traits in hybrid zones: A window on tension zone

models. *Evolution*, *62*(11), 2789–2806. doi: 10.1111/j.1558-5646.2008.00491.x

Gompert, Z., & Buerkle, C. A. (2009). A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, *18*(6), 1207–1224. doi: 10.1111/j.1365-294X.2009.04098.x

Gompert, Z., & Buerkle, C. A. (2010). INTROGRESS: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, *10*(2), 378–384.

Gompert, Z., & Buerkle, C. A. (2011). Bayesian estimation of genomic clines. *Molecular Ecology*, *20*, 2111–2127. doi: 10.1111/1755-0998.12009.x

Gompert, Z., & Buerkle, C. A. (2012). bgc: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, *12*, 1168–1176. doi: doi: 10.1111/1755-0998.12009.x

Gompert, Z., Parchman, T. L., & Buerkle, C. A. (2012a). Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 439–450. doi: 10.1098/rstb.2011.0196

Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Williamson, R. J., & Buerkle, C. A. (2010). Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, *19*(12), 2455–2473. doi: 10.1111/j.1365-294X.2010.04666.x

Gompert, Z. Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., & Buerkle, C. A. (2012b). Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, *66*(7), 2167–2181.

Gourbiere, S., & Mallet, J. (2010). Are species real? The shape of the species boundary with exponential failure, reinforcement, and the “missing snowball.”

Evolution, *64*(1), 1–24.

Gowen, F. C., Maley, J. M., Cicero, C., Peterson, A. T., Faircloth, B. C., Warr, T. C., & McCormack, J. E. (2014). Speciation in Western Scrub-Jays, Haldane's rule, and genetic clines in secondary contact. *BMC Evolutionary Biology*, *14*(1), 1–15.

Guénet, J.-L., Nagamine, C., Simon-Chazottes, D., Montagutelli, X., & Bonhomme, F. (1990). Hst-3: an X-linked hybrid sterility gene. *Genetics Research*, *56*(2–3), 163–165.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, *12*(2), 101–109.

Janoušek, V., Wang, L., Luzynski, K., Dufková, P., Vyskočilová, M. M., Nachman, M. W., … Tucker, P. K. (2012). Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. *Molecular Ecology*, *21*(12), 3032–3047.

Jiggins, C. D., Linares, M., Naisbit, R. E., Salazar, C., Yang, Z. H., & Mallet, J. (2001). Sex-linked hybrid sterility in a butterfly. *Evolution*, *55*(8), 1631–1638.

Jofre, G. I., & Rosenthal, G. G. (2021). A narrow window for geographic cline analysis using genomic data: Effects of age, drift, and migration on error rates. *Molecular Ecology Resources*, *21*(7), 2278–2287.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071.

Kaliontzopoulou, A., Pinho, C., Harris, D. J., & Carretero, M. A. (2011). When cryptic diversity blurs the picture: A cautionary tale from Iberian and North African *Podarcis* wall lizards. *Biological Journal of the Linnean Society*, *103*(4), 779–800.

doi: 10.1111/j.1095-8312.2011.01703.x

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V, Pääbo, S., Villablanca, F. X., & Wilson, a C. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(16), 6196–6200. doi: 10.1073/pnas.86.16.6196

Larson, E. L., White, T. A., Ross, C. L., & Harrison, R. G. (2014). Gene flow and the maintenance of species boundaries. *Molecular Ecology*, *23*(7), 1668–1678. doi: 10.1111/mec.12601

Lasne, C., Sgrò, C. M., & Connallon, T. (2017). The relative contributions of the X chromosome and autosomes to local adaptation. *Genetics*, *205*(3), 1285–1304.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Llopart, A. (2012). The rapid evolution of X-linked male-biased gene expression and the large-X effect in *Drosophila yakuba*, *D. santomea*, and their hybrids. *Molecular Biology and Evolution*, *29*(12), 3873–3886.

Luttikhuizen, P. C., Drent, J., Peijnenburg, K., Van der Veer, H. W., & Johannesson, K. (2012). Genetic architecture in a marine hybrid zone: comparing outlier detection and genomic clines analysis in the bivalve *Macoma balthica*. *Molecular Ecology*, *21*(12), 3048–3061.

Lynch, M., & Force, A. G. (2000). The origin of interspecific genomic incompatibility via gene duplication. *The American Naturalist*, *156*(6), 590–605.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, *20*(5), 229–237. doi: 10.1016/j.tree.2005.02.010

Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M., & Eeley, H. (1990).

Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics*, *124*(4), 921–936.

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS biology*, *17*(2), e2006288.

Masly, J. P., & Presgraves, D. C. (2007). High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biology*, *5*(9), e243.

Matute, D. R., Butler, I. A., Turissini, D. A., & Coyne, J. A. (2010). A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science*, *329*(5998), 1518–1521.

Melo-Ferreira, J., Alves, P. C., Freitas, H., Ferrand, N., & Boursot, P. (2009). The genomic legacy from the extinct *Lepus timidus* to the three hare species of Iberia: Contrast between mtDNA, sex chromosomes and autosomes. *Molecular Ecology*, *18*(12), 2643–2658. doi: 10.1111/j.1365-294X.2009.04221.x

Moran, C. (1981). Genetic demarcation of geographical distribution by hybrid zones. *Proc. Ecol. Soc. Aust*, *11*, 67–73.

Mullen, S. P., Dopman, E. B., & Harrison, R. G. (2008). Hybrid zone origins, species boundaries, and the evolution of wing-pattern diversity in a polytypic species complex of North American admiral butterflies (Nymphalidae: *Limenitis*). *Evolution*, *62*(6), 1400–1417. doi: 10.1111/j.1558-5646.2008.00366.x

Muller, H. (1942). Isolating mechanisms, evolution, and temperature. *Biol. Symp.*, *6*, 71–125.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587): 332-342.

1075 Nosil, P., Harmon, L. J., & Seehausen, O. (2009). Ecological explanations for
1076 (incomplete) speciation. *Trends in Ecology & Evolution*, *24*(3), 145–156.
1077 Orr, H. A. (1987). Genetics of male and female sterility in hybrids of *Drosophila*
1078 *pseudoobscura* and *D. persimilis*. *Genetics*, *116*(4), 555–563.
1079 Parchman, T. L., Gompert, Z., Braun, M. J., Brumfield, R. T., McDonald, D. B., Uy, J.
1080 A. C., … Buerkle, C. A. (2013). The genomic consequences of adaptive
1081 divergence and reproductive isolation between species of manakins. *Molecular*
1082 *Ecology*, *22*(12), 3304–3317. doi: 10.1111/mec.12201
1083 Parchman, Thomas L, Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., &
1084 Buerkle, C. (2012). Genome-wide association genetics of an adaptive trait in
1085 lodgepole pine. *Molecular Ecology*, *21*(12), 2991–3005.
1086 Payseur, B. A., Krenz, J. G., & Nachman, M. W. (2004). Differential patterns of
1087 introgression across the X chromosome in a hybrid zone between two species of
1088 house mice. *Evolution*, *58*(9), 2064–2078.
1089 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012).
1090 Double digest RADseq: an inexpensive method for de novo SNP discovery and
1091 genotyping in model and non-model species. *PloS One*, *7*(5), e37135.
1092 Pinho, C., Kaliontzopoulou, A., Carretero, M. A., Harris, D. J., & Ferrand, N. (2009).
1093 Genetic admixture between the Iberian endemic lizards *Podarcis bocagei* and
1094 *Podarcis carbonelli*: Evidence for limited natural hybridization and a bimodal
1095 hybrid zone. *Journal of Zoological Systematics and Evolutionary Research*,
1096 *47*(4), 368–377. doi: 10.1111/j.1439-0469.2009.00532.x
1097 Pinho, Catarina, Ferrand, N., & Harris, D. J. (2006). Reexamination of the Iberian and
1098 North African *Podarcis* (Squamata: Lacertidae) phylogeny based on increased
1099 mitochondrial DNA sequencing. *Molecular Phylogenetics and Evolution*, *38*(1),

266–273. doi: 10.1016/j.ympev.2005.06.012

Pinho, Catarina, Harris, D. J., & Ferrand, N. (2007). Contrasting patterns of population subdivision and historical demography in three western Mediterranean lizard species inferred from mitochondrial DNA variation. *Molecular Ecology*, *16*(6), 1191–1205. doi: 10.1111/j.1365-294X.2007.03230.x

Pinho, Catarina, & Hey, J. (2010). Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 215–230.

Polechová, J., & Barton, N. (2011). Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, *189*(1), 227–235.

Purcell, J., Brelsford, A., Wurm, Y., Perrin, N., & Chapuisat, M. (2014). Convergent genetic architecture underlies social organization in ants. *Current Biology*, *24*(22), 2728–2732.

Puritz, J. B., Matz, M. V, Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, 23, 5937-5942.

Qvarnström, A., & Bailey, R. I. (2009). Speciation through evolution of sex-linked genes. *Heredity*, *102*(1), 4.

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., … Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477.

Rieseberg, L. H. (2009). Evolution: replacing genes and traits through hybridization. *Current Biology*, *19*(3), R119–R122. doi: https://doi.org/10.1016/j.cub.2008.12.016

Sá-Sousa, P. (2001). Comparative chorology between *Podarcis bocagei* and *P. carbonellae* (Sauria: Lacertidae) in Portugal. *Revista Española de Herpetologia*,

15, 85–97.

Scascitelli, M., Whitney, K. D., Randell, R. A., King, M., Buerkle, C. A., & Rieseberg, L. H. (2010). Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H. debilis*) reveals asymmetric patterns of introgression and small islands of genomic differentiation. *Molecular Ecology*, *19*(3), 521–541.

Storchová, R., Reif, J., & Nachman, M. W. (2010). Female heterogamety and speciation: reduced introgression of the Z chromosome between two species of nightingales. *Evolution*, *64*(2), 456–471. doi: 10.1111/j.1558-5646.2009.00841.x

Szymura, J. M., & Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution*, *40*(6), 1141–1159.

Szymura, J. M., & Barton, N. H. (1991). The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, *45*(2), 237–261.

Tarroso, P., Pereira, R. J., Martínez-Freiría, F., Godinho, R., & Brito, J. C. (2014). Hybridization at an ecotone: Ecological and genetic barriers between three Iberian vipers. *Molecular Ecology*, *23*(5), 1108–1123. doi: 10.1111/mec.12671

Teeter, K. C., Thibodeau, L. M., Gompert, Z., Buerkle, C. A., Nachman, M. W., & Tucker, P. K. (2010). The variable genomic architecture of isolation between hybridizing species of house mice. *Evolution: International Journal of Organic Evolution*, *64*(2), 472–485.

Teeter, K., & Payseur, B. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, *18*, 67–76. doi: 10.1101/gr.6757907.1

Templeton, A. R. (1992). The meaning of species and speciation: a genetic perspective. *The Units of Evolution*, 159–185.

Tucker, P. K., Sage, R. D., Warner, J., Wilson, A. C., & Eicher, E. M. (1992). Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution*, *46*(4), 1146–1163.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, *3*(9), e285.

van Schooten, B., Meléndez-Rosa, J., Van Belleghem, S. M., Jiggins, C. D., Tan, J. D., McMillan, W. O., & Papa, R. (2020). Divergence of chemosensing during the early stages of speciation. *Proceedings of the National Academy of Sciences*, *117*(28), 16438–16447.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1587), 451–460.

Vignoli, L., Vuerich, V., & Bologna, M. A. (2012). Experimental study of dispersal behaviour in a wall lizard species (*Podarcis sicula*) (Sauria Lacertidae). *Ethology Ecology & Evolution*, *24*(3), 244–256.

Walsh, J., Shriver, W. G., Olsen, B. J., & Kovach, A. I. (2016). Differential introgression and the maintenance of species boundaries in an advanced generation avian hybrid zone. *BMC Evolutionary Biology*, *16*(1), 1–18.

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. doi: 10.1046/j.1420-9101.2001.00335.x

Yang, W., Feiner, N., Pinho, C., While, G. M., Kaliontzopoulou, A., Harris, D. J., … Uller, T. (2021). Extensive introgression and mosaic genomes of Mediterranean endemic lizards. *Nature Communications*, *12*(1), 1–8.

Data Accessibility Statement

Custom Python script used to select one SNP per locus is available at https://github.com/catpinho/filter_RADseq_data and the scripts used to perform the simulations, format conversions and data set manipulations are available at https://github.com/catpinho/admixture_simulation. Raw sequence reads from ddRAD-seq are deposited in the NCBI Sequence Read Archive (SRA; BioProject PRJNA665746). BioSample numbers for each individual are provided in Supplemental Information File 2. Genotype data in VCF format is available on Dryad (DOI https://doi.org/10.5061/dryad.n8pk0p31k). Sequences of mtDNA are deposited in GeneBank and accession numbers for each individual are provided in Supplemental Information File 2.

Author Contributions

GCD was responsible for field and laboratory work, bioinformatics, data analysis, and writing the manuscript. AB was responsible for laboratory work and manuscript revision. MMR was responsible for field and laboratory work and manuscript revision. PAC was responsible for project conception, design and coordination, manuscript revision and obtaining funding. CP was responsible for project conception, design and coordination, bioinformatics, data analysis, manuscript writing and revision, and obtaining funding.

Figures

Figure 1 – a) Map of Iberian Peninsula with distribution of *Podarcis bocagei* (in red) and *P. carbonelli* (in blue); **b)** The highlighted area in a) showing the locations of sampled populations in the transect across the hybrid zone; **c)** zoom in of the area in b) where individuals of both parental species were found in syntopy. Each triangle represents an individual assigned to each parental species (Q < 0.1 or Q > 0.9) and each star represents an individual with Q between 0.1 and 0.9 based on Admixture analysis.

Figure 2 – Population structure across the hybrid zone between *P. bocagei* and *P. carbonelli*. **a)** Results from individual multilocus genotype clustering for the complete dataset (329 individuals and 6905 loci). Each individual is represented by a vertical line proportionally partitioned into the K=2 and K=3 coloured segments. Population acronyms and colours as in Figure 1. **b)** Distribution of Qc scores in 10 classes of equal size within the syntopy area, derived from the multilocus genotype assignment analysis with the complete dataset. **c)** Principal Component Analysis (PCA) for the complete dataset. Circles represent individuals from populations north of the contact zone, triangles correspond to the individuals from the contact zone and squares identify the individuals from populations south of the contact zone. Red represents individual identified as *P. bocagei* and blue as *P. carbonelli* based on morphological identification in the field. **d)** Triangle plot showing the distribution of individual hybrid index ($HI_{DIAG\text{-}INT}$) and interspecific heterozygosity (He) in the contact zones based on diagnostic dataset (1241 loci) between reference populations (POR and TOR) in the contact zone. Colours and shapes as in panel **b**. *Podarcis bocagei* reference individuals were set to have a HI of zero, and *P. carbonelli* was set to one.

Figure 3 – Box plots representing the ratios of simulated distributions of F1 hybrids by backcrosses (F1/BC) with different migration rates, i.e., the proportions of genotypes replaced each generation by parental genotypes (*HI* = 0 or 1) in the same proportion as they appear in the observed dataset. **a)** Distributions based on simulations with random mating and **b)** based on non-random mating. In both panels horizontal doted lines represent the F1/BC ratio in the observed data.

Figure 4 – Geographic and genomic cline variation across SNPs. **a)** Geographic cline variation of the 2300 loci ("80/20" dataset) described by *c* and *w*, estimated with HZAR. Gray circles are the loci without significant differences in *c* when compared to c_{HI} (non-sig.), red circles are the loci with *c* significantly shifted to North (sign. to North) and blue circles are the loci with *c* significantly shifted to South (sign. to South). Yellow filled circles are the loci with *w* significantly higher than w_{HI} (sign. larger). The circles with an asterisk denote loci with outlier *c*. Horizontal dotted line represent the c_{HI} (13.15km) and the vertical represents the w_{HI} (0.54km). Solid line represents the regression line between *c* and *w* values. Regression significance is displayed on top-right corner. **b)** Genomic cline variation in the contact zone (ESP population) described by α and β, estimated with BGC. Each point represents one of the 6905 loci (complete dataset), and contour lines show regions of the $\alpha-\beta$ parameter space with similar density of loci.

Figure 5 – Concordance of analogous parameters from geographic and genomic clines using the 2300 loci common to both complete and "80/20" datasets. Regression lines, it's respective r^2 values and their significance are shown in each plot. **a)** Concordance of *α* and *c*. Circles filled in red are loci with *α* significantly associated *P. bocagei* genomic background and circles filled in blue are loci with *α* significantly associated *P. carbonelli* genomic background. Grey filled circles are loci that *α* do not significantly differ from neutral expectations. Red circles represent loci with *c* significantly shifted from c_{HI} to North and blue circles to South. Grey circles are loci that *c* do not significantly differ from c_{HI}. **b)** Concordance of *β* and *w*. Yellow filled circles represent loci with increased introgression rate (significant negative values) and green with decreased introgression rate (significant positive values), relatively to neutral expectations. Grey filled circles are loci that *β* do not significantly differ from neutral expectations. Yellow circles are loci with *w* significantly higher from w_{HI} (loci with w significantly lower from w_{HI} were not detected) and grey circles are loci that *w* do not significantly differ from w_{HI}.

Figure 6 – Distribution of geographic and genomic cline parameters across the loci mapped on the *P. muralis* genome. **a)** Genomic distribution of loci and it's respective geographic and genomic cline parameters estimated with HZAR and BGC, respectively. Each grey bar in the bottom represents one chromosome and it is proportional to its size. Each coloured dot represents statistically significant loci for each parameter. Colours as in Figure 5. **b, c)** Violin plots representing the density of values for geographic cline centre (c) and width (w) estimated with HZAR for the 2101 mapped loci contained in the "80/20" dataset. Due to extreme outlier values in both panels, y-axis is not continuous to facilitate visualization; plots with complete set of outliers are shown in Supplemental Information File 1, Figure S1. **d, e)** Violin plots representing the density of values for genomic cline centre (α) and introgression rate (β) estimated with BGC for the 6027 mapped loci contained in the complete dataset. The horizontal solid line within each box plot represents the median.

Tables

Table 1 – Summary results of geographic cline analysis performed with HZAR using 2300 loci ("80/20" dataset, upper section) and genomic cline analysis performed with BGC using 6905 loci (complete dataset, lower section). For each section are represented the number of loci (N) and correspondent percentage falling in each class of geographic clines centre (c)/cline width (w) and locus specific ancestry (α)/introgression rate (β) of genomic cline.

Table 2 – Non-parametric global Kruskal-Wallis tests to assess if the distribution of each parameter value used to describe introgression, from both geographic (*c* and *w*) or genomic (*α* and *β*) cline analyses, were similar between chromosomes. Significant differences ($p < 0.05$) are marked with an asterisk (*).

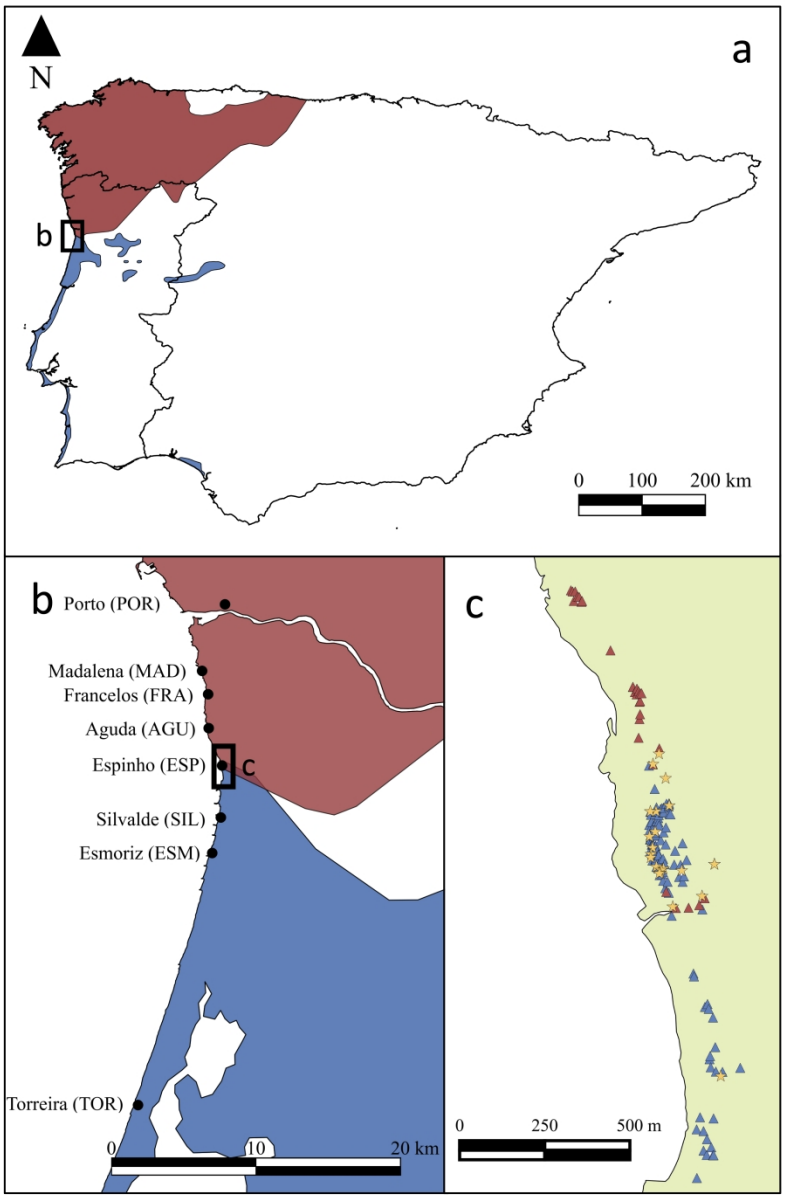

Figure 1 – a) Map of Iberian Peninsula with distribution of *Podarcis bocagei* (in red) and *P. carbonelli* (in blue); b) The highlighted area in a) showing the locations of sampled populations in the transect across the hybrid zone; c) zoom in of the area in b) where individuals of both parental species were found in syntopy. Each triangle represents an individual assigned to each parental species (Q < 0.1 or Q > 0.9) and each star represents an individual with Q between 0.1 and 0.9 based on Admixture analysis.

1276x1937mm (72 x 72 DPI)

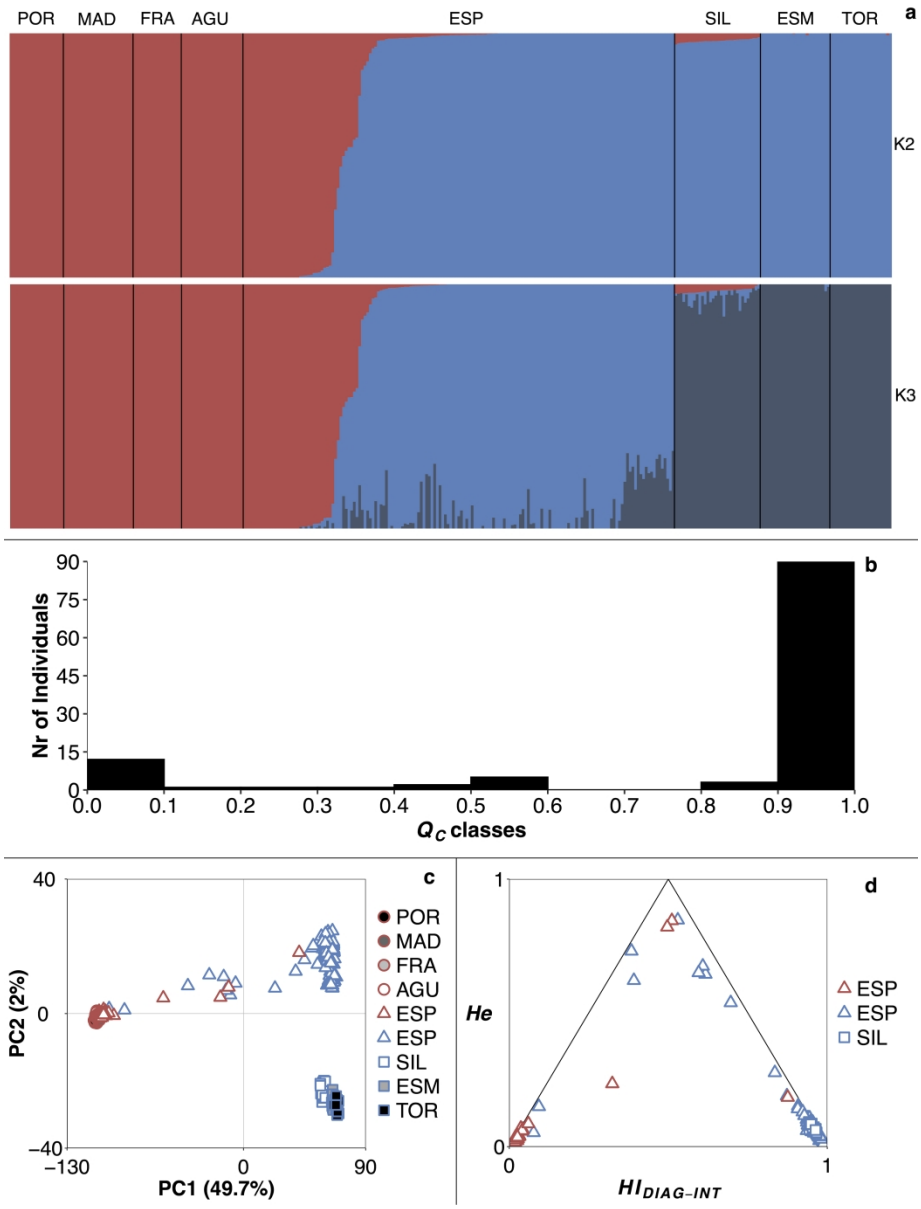

Figure 2 – Population structure across the hybrid zone between *P. bocagei* and *P. carbonelli*. a) Results from individual multilocus genotype clustering for the complete dataset (329 individuals and 6905 loci). Each individual is represented by a vertical line proportionally partitioned into the K=2 and K=3 coloured segments. Population acronyms and colours as in Figure 1. b) Distribution of Qc scores in 10 classes of equal size within the syntopy area, derived from the multilocus genotype assignment analysis with the complete dataset. c) Principal Component Analysis (PCA) for the complete dataset. Circles represent individuals from populations north of the contact zone, triangles correspond to the individuals from the contact zone and squares identify the individuals from populations south of the contact zone. Red represents individual identified as *P. bocagei* and blue as *P. carbonelli* based on morphological identification in the field. d) Triangle plot showing the distribution of individual hybrid index (HIDIAG-INT) and interspecific heterozygosity (He) in the contact zones based on diagnostic dataset (1241 loci) between reference populations (POR and TOR) in the contact zone. Colours and shapes as in panel b. *Podarcis bocagei* reference individuals were set to have a HI of zero, and *P. carbonelli* was set to one.

2114x2774mm (65 x 65 DPI)

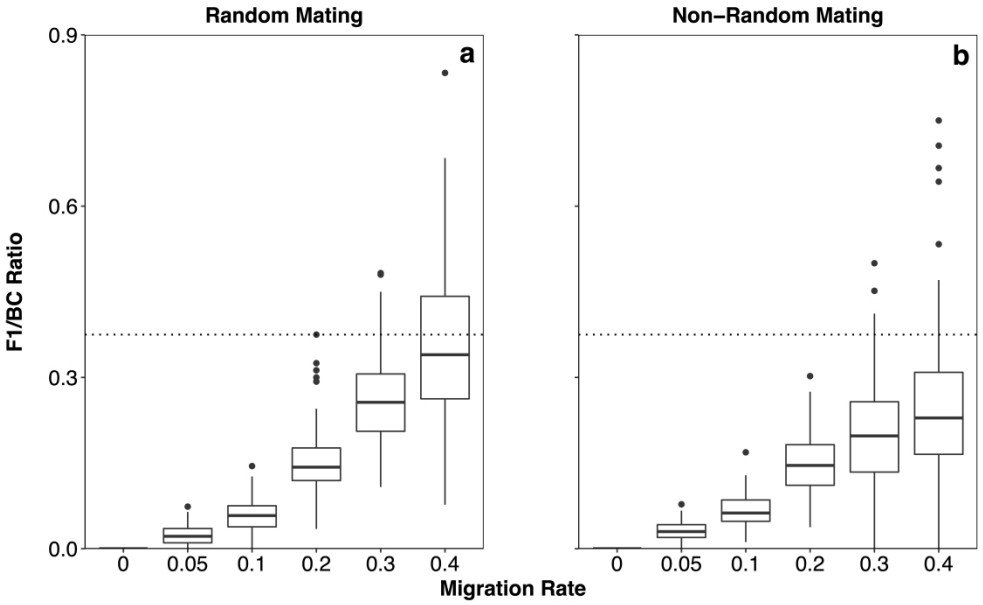

Figure 3 – Box plots representing the ratios of simulated distributions of F1 hybrids by backcrosses (F1/BC) with different migration rates, i.e., the proportions of genotypes replaced each generation by parental genotypes (*HI* = 0 or 1) in the same proportion as they appear in the observed dataset. a) Distributions based on simulations with random mating and b) based on non-random mating. In both panels horizontal doted lines represent the F1/BC ratio in the observed data.

2363x1439mm (72 x 72 DPI)

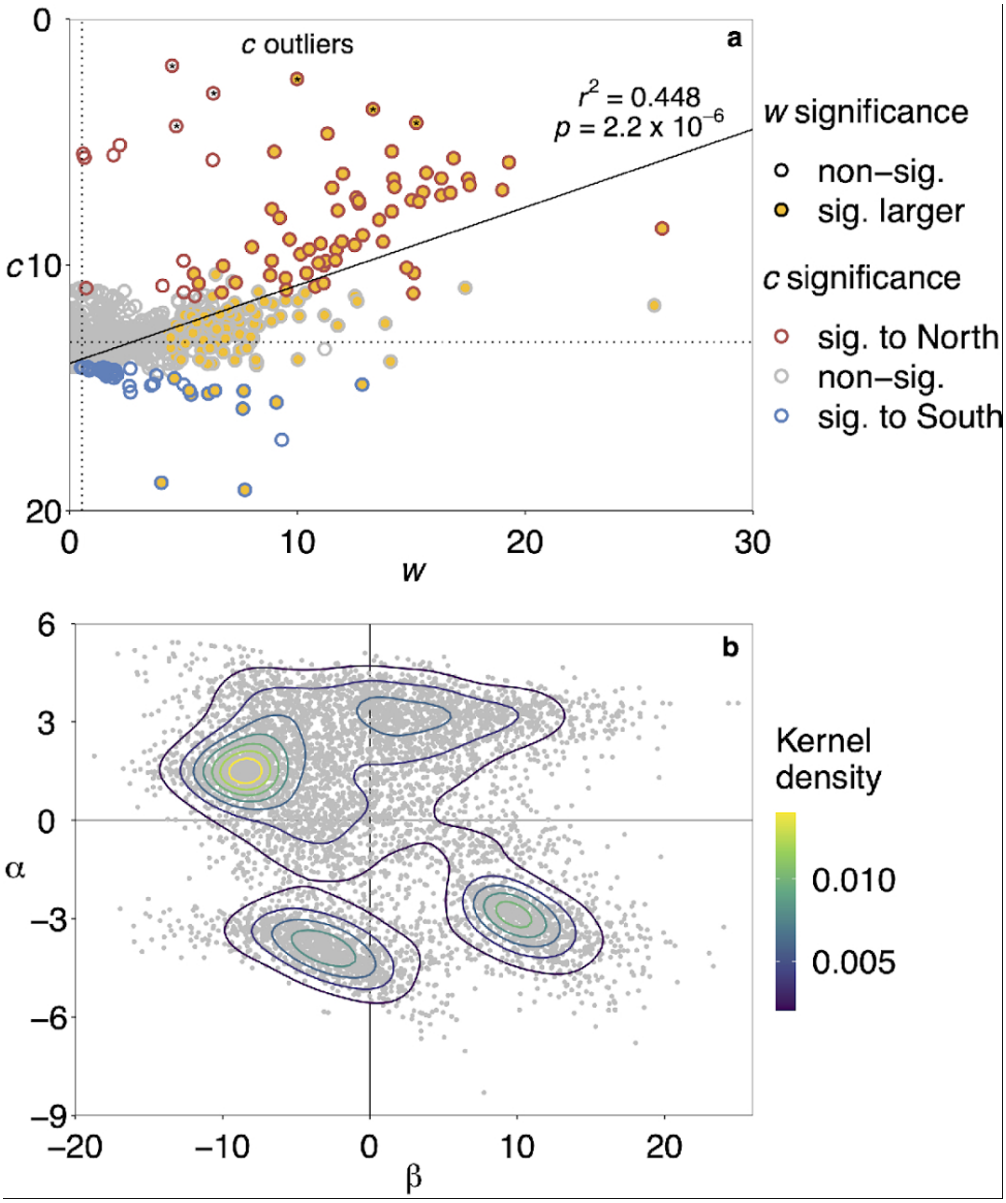

Figure 4 – Geographic and genomic cline variation across SNPs. a) Geographic cline variation of the 2300 loci ("80/20" dataset) described by *c* and *w*, estimated with HZAR. Gray circles are the loci without significant differences in *c* when compared to c_{HI} (non-sig.), red circles are the loci with *c* significantly shifted to North (sign. to North) and blue circles are the loci with *c* significantly shifted to South (sign. to South). Yellow filled circles are the loci with *w* significantly higher than w_{HI} (sign. larger). The circles with an asterisk denote loci with outlier *c*. Horizontal dotted line represent the c_{HI} (13.15km) and the vertical represents the w_{HI} (0.54km). Solid line represents the regression line between *c* and *w* values. Regression significance is displayed on top-right corner. b) Genomic cline variation in the contact zone (ESP population) described by α and β, estimated with BGC. Each point represents one of the 6905 loci (complete dataset), and contour lines show regions of the α–β parameter space with similar density of loci.

189x226mm (144 x 144 DPI)

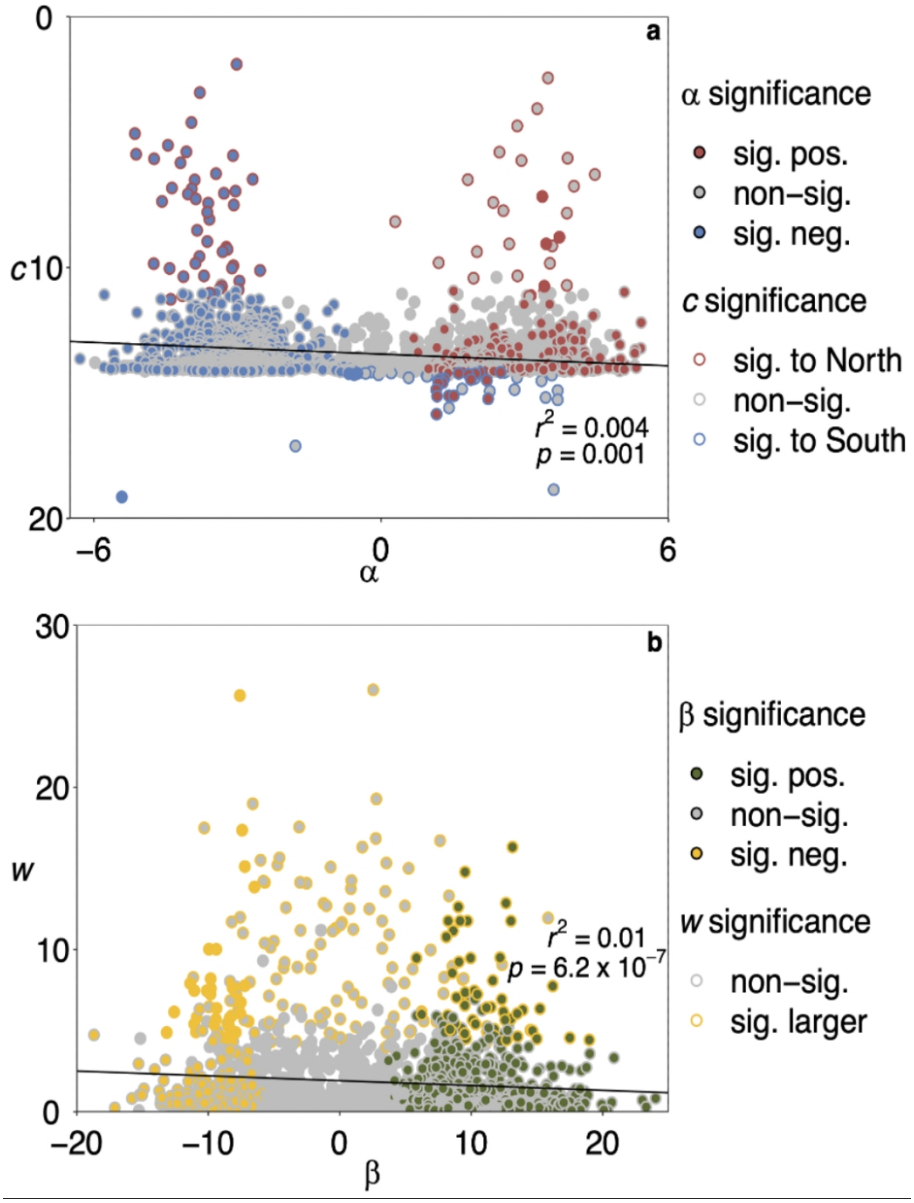

Figure 5 – Concordance of analogous parameters from geographic and genomic clines using the 2300 loci common to both complete and "80/20" datasets. Regression lines, it's respective r^2 values and their significance are shown in each plot. a) Concordance of *α* and *c*. Circles filled in red are loci with *α* significantly associated *P. bocagei* genomic background and circles filled in blue are loci with *α* significantly associated *P. carbonelli* genomic background. Grey filled circles are loci that *α* do not significantly differ from neutral expectations. Red circles represent loci with *c* significantly shifted from c_{HI} to North and blue circles to South. Grey circles are loci that *c* do not significantly differ from c_{HI}. b) Concordance of *β* and *w*. Yellow filled circles represent loci with increased introgression rate (significant negative values) and green with decreased introgression rate (significant positive values), relatively to neutral expectations. Grey filled circles are loci that *β* do not significantly differ from neutral expectations. Yellow circles are loci with *w* significantly higher from w_{HI} (loci with *w* significantly lower from w_{HI} were not detected) and grey circles are loci that *w* do not significantly differ from w_{HI}.

188x248mm (144 x 144 DPI)

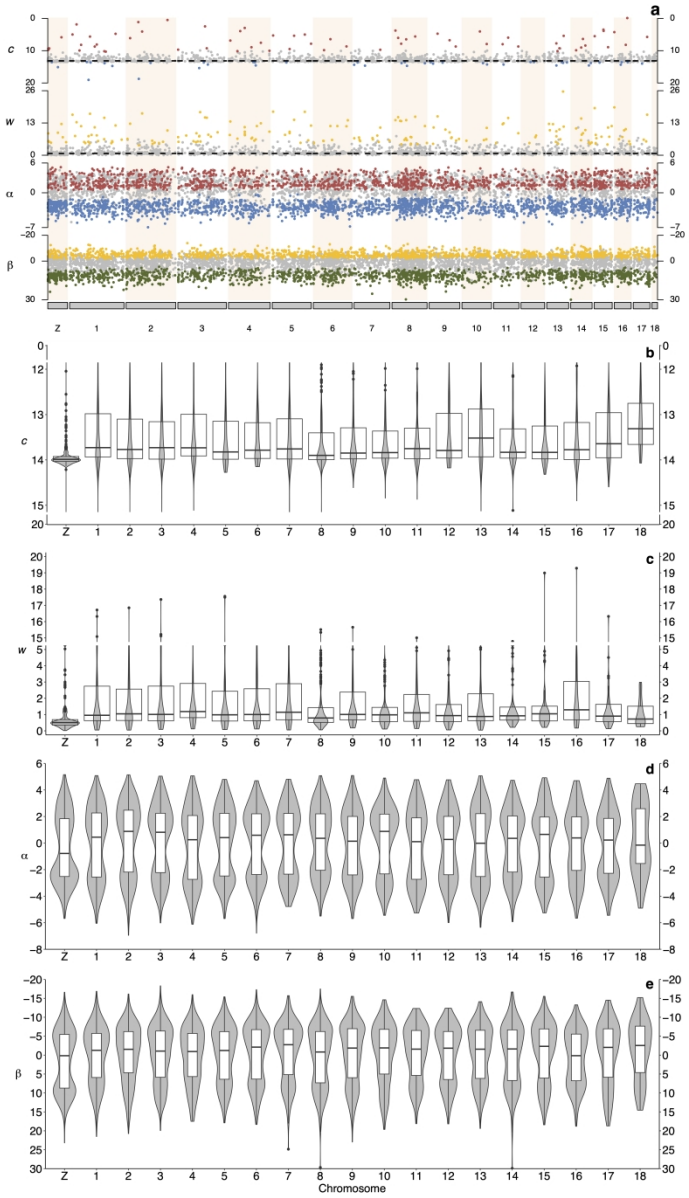

Figure 6 – Distribution of geographic and genomic cline parameters across the loci mapped on the *P. muralis* genome. a) Genomic distribution of loci and it's respective geographic and genomic cline parameters estimated with HZAR and BGC, respectively. Each grey bar in the bottom represents one chromosome and it is proportional to its size. Each coloured dot represents statistically significant loci for each parameter. Colours as in Figure 5. b, c) Violin plots representing the density of values for geographic cline centre (*c*) and width (*w*) estimated with HZAR for the 2101 mapped loci contained in the "80/20" dataset. Due to extreme outlier values in both panels, y-axis is not continuous to facilitate visualization; plots with complete set of outliers are shown in Supplemental Information File 1, Figure S1. d, e) Violin plots representing the density of values for genomic cline centre (α) and introgression rate (β) estimated with BGC for the 6027 mapped loci contained in the complete dataset. The horizontal solid line within each box plot represents the median.

1625x2810mm (72 x 72 DPI)

Geographic clines		*w*							
		Similar to *HI*		Smaller than *HI*		Higher than *HI*		Total	
		N	%	N	%	N	%	N	%
c	Not shifted	2064	89.7	0	0	116	5	2180	94.8
	Shifted North	13	0.6	0	0	60	2.6	73	3.2
	Shifted South	36	1.6	0	0	11	0.5	47	2
	Total	2113	91.9	0	0	187	8.1	2300	100

Genomic clines		*β*							
		Zero		Negative		Positive		Total	
		N	%	N	%	N	%	N	%
α	Zero	3194	46.3	354	5.1	77	1.1	3625	52.5
	Positive	205	2.9	755	10.9	93	1.3	1053	15.2
	Negative	1068	15.5	37	0.5	1122	16.2	2227	32.3
	Total	4467	64.7	1146	16.7	1292	18.7	6905	100

Parameter		Kruskal-Wallis chi-squared value	*p*-value
Geographic clines	*c*	169.4568103	1.16×10^{-28} *
	w	186.3242633	5.33×10^{-30} *
Genomic clines	*α*	22.10320156	0.009 *
	β	33.67921046	8.43×10^{-08} *