

The multiple population genetic and demographic routes to islands of genomic divergence

Claudio S. Quilodrán¹, Kristen Ruegg^{1,2,3}, Ashley T. Sendell-Price¹, Eric Anderson⁴, Tim Coulson^{1†}, Sonya Clegg^{1†}

¹Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK. ²Center for Tropical Research, Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA, USA. ³Department of Biology, Colorado State University, Fort Collins, CO, USA. ⁴Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Santa Cruz, CA, USA. [†]These authors are joint senior authors.

Abstract

1. The way that organisms diverge into reproductively isolated species is a major question in biology. The recent accumulation of genomic data provides promising opportunities to understand the genomic landscape of divergence, which describes the distribution of differences across genomes. Genomic areas of unusually high differentiation have been called genomic islands of divergence. Their formation has been attributed to a variety of mechanisms, but a prominent hypothesis is that they result from divergent selection over a small portion of the genome, with surrounding areas homogenised by gene flow. Such islands have often been interpreted as being associated with divergence with gene flow. However other mechanisms related to genomic structure and population history can also contribute to the formation of genomic islands of divergence.
2. We currently lack a quantitative framework to examine the dynamics of genomic landscapes under the complex and nuanced conditions that are found in natural systems. Here, we develop an individual-based simulation to explore the dynamics of diverging genomes under various scenarios of gene flow, selection and genotype-phenotype maps.
3. Our modelling results are consistent with empirical observations demonstrating the formation of genomic islands under genetic isolation. Importantly, we have quantified the range of conditions that produce genomic islands. We demonstrate that the initial level of genetic diversity, drift, time since divergence, linkage disequilibrium, strength of selection and gene flow are all important factors that can influence the formation of genomic islands. Because the accumulation of genomic differentiation over time tends to erode the signal of genomic islands, genomic islands are more likely to be observed in recently divergent taxa, although not all recently diverged taxa will necessarily exhibit islands of genomic divergence. Gene flow primarily slows the swamping of islands of divergence with time.
4. By using this framework, further studies may explore the relative influence of particular suites of events that contribute to the emergence of genomic islands under sympatric, parapatric and allopatric conditions. This approach represents a novel tool to explore quantitative expectations of the speciation process, and should prove useful in elucidating past and projecting future genomic evolution of any taxa.

Keywords: Evolution, gls R package, genomic landscape, individual based model, island of genomic divergence.

Introduction

A major aim of evolutionary biology is to understand mechanisms associated with the divergence of organisms between populations and the emergence of new species. This motivated Charles Darwin and Alfred Wallace 160 years ago when they advanced the Theory of Natural Selection (Darwin & Wallace 1858). Since then, the increasing

49 accumulation of genetic, genomic and computational tools has allowed a better
50 understanding of the genetic basis of the speciation process, resulting in the rise of a
51 new era of evolutionary research (Hughes 2009; Chanderbali *et al.* 2016; Roux *et al.*
52 2016). Patterns of divergence at the level of the genome have been characterised for
53 an increasing number of taxa, but the extent to which observed patterns are
54 informative about evolutionary processes is actively debated (e.g. Ellegren *et al.*
55 2012; Renaut *et al.* 2013; Ruegg *et al.* 2014; Burri *et al.* 2015).

56 The genomic landscape of divergence describes the distribution of differences
57 across the genomes of diverging organisms. The genome of a diverging taxon does
58 not change uniformly, with some regions changing at higher rates than others
59 (Seehausen *et al.* 2014; Ravinet *et al.* 2017). If a single process uniquely generates a
60 particular divergence pattern, then identification of that pattern can confidently be
61 interpreted as representing a particular evolutionary history. In contrast, if multiple
62 processes can generate the same patterns of genomic divergence, then identification of
63 the pattern will not point to a specific process, though the suite of candidate processes
64 may be narrowed. In these cases, additional information beyond patterns of genomic
65 divergence, such as the ecological and evolutionary context of a given divergence,
66 will be required to understand patterns of evolutionary divergence.

67 Genomic islands of divergence - highly differentiated regions of the genome
68 that are surrounded by regions of low differentiation - are a particularly intriguing
69 pattern of genomic divergence (Turner, Hahn & Nuzhdin 2005; Harr 2006; Nosil,
70 Funk & Ortiz - Barrientos 2009). Genomic islands are often a focus of studies using
71 genomic scans to identify target genes of natural selection (Lotterhos & Whitlock
72 2014). Initially, genomic island formation was attributed to the action of divergent
73 natural selection on particular loci, which contribute to reproductive isolation via pre-
74 or post-zygotic barriers to a greater extent than elsewhere within the genome (Wu
75 2001; Wu & Ting 2004). This process is hypothesised to create elevated regions of
76 genomic differentiation, containing the selected loci and other physically linked loci,
77 surrounded by regions homogenised by gene flow (Nosil, Funk & Ortiz - Barrientos
78 2009). Several authors consequently interpreted presence of genomic islands of
79 divergence as a signal of divergence with gene flow (e.g. Feder *et al.* 2013),
80 concluding that the speciation process in sympatric or parapatric conditions may be
81 more common than previously thought (e.g. Nosil 2008; Fraïsse *et al.* 2014; Soria-

Carrasco *et al.* 2014). However, empirical studies have also proposed that genomic islands of divergence can arise in the absence of gene flow due to a variety of causes, such as the structure of the diverging genomes (e.g. variation in the recombination rate) and the action of genetic drift, background selection, and adaptation to local environmental conditions (Noor & Bennett 2009; Cruickshank & Hahn 2014; Campagna *et al.* 2015). Genomic regions of reduced recombination, such as areas close to centromeres or inside chromosome rearrangements that experience background or divergent selection have received particular attention (Noor & Bennett 2009). Selection acting in regions with low recombination rates is expected to reduce the effective population size of these genomic regions to a greater extent than elsewhere in the genome, which may result in the creation of genomic islands (Feder & Nosil 2009; Turner & Hahn 2010). Note that those regions of reduced recombination may still harbour important genes related to the divergence process (Samuk *et al.* 2017). Furthermore, regions of low divergence may occur because of incomplete lineage sorting rather than homogenisation by gene flow, with peaks of genetic divergence being an artefact of a loss of nucleotide diversity within populations after divergent selection (Cruickshank & Hahn 2014).

There have been some previous attempts to model the structure of genomic landscapes that can be applied to the formation of genomic islands of divergence. However, these models either simulate single bi-allelic selected loci (Charlesworth, Nordborg & Charlesworth 1997; Sedghifar, Brandvain & Ralph 2016) or consider a small number of simulated loci (Feder & Nosil 2009; Feder & Nosil 2010; Feder *et al.* 2012). They also provide a static view of the divergence process by summarizing selection as a single parameter (sometimes referred to as s). While informative, such models represent specific stages when populations have already achieved a given level of differentiation. Flaxman, Feder and Nosil (2013) used an individual-based model to project this dynamic forward in time, but their model was constrained to a uniform distribution of loci with constant recombination rates. Recently developed tools allow the simulation of large quantities of genomic information in an efficient coalescent framework (Kelleher, Etheridge & McVean 2016; Lohse 2017; Haller *et al.* 2019; Haller & Messer 2019). While simulations based on coalescence are particularly useful when exploring the dynamics of neutral evolution, they are less informative for small populations that experience selection and where reproduction is not panmictic (Curat *et al.* 2015). Nonetheless, a combination of recently developed

tools have helped to perform simulations beyond this original limitation of the coalescence theory (see Haller *et al.* 2019; Haller & Messer 2019). A flexible quantitative framework is desirable to help investigate the dynamics of genomic landscapes, thus further increasing the utility of existing genomic datasets (Feder *et al.* 2013; Seehausen *et al.* 2014). We develop a quantitative individual-based modelling approach to simulate the dynamic of a genomic landscape of divergence (R package “glads”: genomic landscape of divergence simulations). Our approach does not include a backward coalescent algorithm to reconstruct genetic diversity of a subsample of the population. Instead, all individuals in a population are simulated forward in time, allowing the exploration of complex demographic and genetic consequences of selection. The model simulates any number of loci and allelic polymorphisms and can be motivated or parameterised using data. There is a range of individual based frameworks that allow the simulation of all individuals in a population (e.g. Peng & Kimmel 2005; Currat *et al.* 2015). A major difference between our approach and these tools is the inclusion of a fitness function making our approach compatible with many structured ecological and evolutionary models (Nosil, Funk & Ortiz - Barrientos 2009; Hohenlohe *et al.* 2012). This allows us to explore the population and genomic dynamics associated with the process of divergence between populations before an equilibrium is reached, in a way that is often not possible when selection is fixed as a constant parameter. Basically, the fitness function allows to link selection (and hence evolution) to ecology (e.g. population dynamics) in a way that approaches based on parameters describing selection do not. The fitness function is therefore extending genetic variation right through to population dynamics (and potentially community dynamics) in way that has not been done before (see Coulson, Potter & Felmy 2019). Our approach is highly flexible and can be constructed for any genotype-phenotype map and any configuration of recombination rates between neighbouring loci, can be constructed for deterministic and stochastic environments, and incorporates any desired system of mating. The simulation method presented here provides a flexible framework to examine the dynamics of diverging genomic landscapes under various scenarios of gene flow and selection on single genes or networks of multiple interacting genes. Our approach represents a novel tool to evaluate quantitative expectations in genomic landscapes. It is useful to elucidate the influence of a range of demographic and evolutionary scenarios, including divergence

with or without gene flow, the divergence timeframe, and the structure of target genomes.

Methods

General description of the model

The purpose of our model is to provide insight into how a range of genetic and demographic processes can generate genomic signatures and patterns of genomic divergence between populations. Our primary motivation was to explore factors associated with the emergence of genomic islands of divergence, but our approach can be applied to many questions about the structure of genomes, genomic landscapes, and the evolution of divergent organisms.

The model is individual-based and consists of two populations that may or may not be linked by gene-flow. Our model is composed of three hierarchical levels: genotypes, phenotypic traits, and demographic rates. The dynamics of the populations, the distributions of genotypes at each locus and the phenotypic traits, are all emergent properties of the model. The model tracks the multivariate distribution of multi-locus genotypes and phenotypes. We simulate individuals that are characterised by sex and genetic identity (Fig. 1a). The genotype and the environment determine the phenotypic trait values of an individual via a genotype-phenotype map. The phenotypic trait values influence an individual's expected demography (i.e. survival, mate choice, and reproductive success). For example, assuming a per generation time step, the potential number of offspring produced by each individual depends on its phenotype $\omega = f(z)$, which in turn depends on the individual genotype and on the environment $z = g(G, E)$ (Fig. 1b). G is a numeric value determined by an individual's genotype, representing the genetic value of the genotype. In the case of an additive genetic map, the genetic value of a genotype will be a breeding value. E represents the effect of the environment on phenotypic expression, and this allows us to capture the effects of plasticity on phenotypic expression. The environmental effect is important when simulating real-life eco-evolutionary dynamics because it almost always interacts with the genotype to determine the expression of a phenotypic trait (Bradshaw 1965; Kokko *et al.* 2017). The realized demography, and consequently summary statistics such as the sex ratio and number of offspring, are obtained by sampling from a distribution whose expected value is the expected demography. Once

182 mating pairs are formed, the genotype of the young is determined by merging haploid
183 gametes produced by each parent. Genetic variation of the offspring is determined by
184 recombination and mutation.

185 We describe how the model is implemented in the next section. Our starting
186 point is the distribution of individuals classified by genotype, sex and population (Fig.
187 1a). First, we generate the phenotypic trait of each individual given its genotype (and
188 potentially the environment). Second, we calculate individual fitness given an
189 individual's phenotype, the population it is in, and potentially the environment.
190 Mating pairs are formed (randomly, assortatively, or disassortatively) based on these
191 individual fitness scores. Parental gametes are then produced given recombination and
192 mutation rates, before segregating within mating pairs to generate offspring
193 genotypes. The offspring can disperse to the neighbouring population with a given
194 probability. The loop is then repeated for the next offspring generation.

196 Individual based framework

197 We assume organisms are diploid and composed of males and females. Each
198 individual i is characterized by a two-dimensional array that represents a pair of
199 homologous chromosomes. Multiple pairs of arrays may also be constructed to allow
200 the characterisation of any number of chromosomes. Similarly, variation in the
201 number of dimensions of the arrays may be introduced to extend this framework to
202 haploid or polyploid organisms. Each element of the array is an integer defining the
203 copy of a given allele at a given locus. Individuals are also classified into populations.
204 We assume random mating within a population, although this assumption can easily
205 be relaxed (Schindler *et al.* 2015; Ellner, Childs & Rees 2016). Populations i and j are
206 linked by migration rates (m_{ij} and m_{ji}) describing movement from population i to j and
207 vice versa. We assume that individuals that migrate and reproduce successfully pass
208 their genes into the other population hence incorporating gene flow into the model.
209 The genotype-phenotype and phenotype-demography map can differ between
210 populations if required.

211 The model proceeds in discrete time steps representing generations. It is a
212 forward simulation that includes reproduction and migration at each time step.
213 Density dependence regulates the population growth rate, influencing the probability
214 of successful reproduction (Fig. 1a).

The fitness of individuals is associated with a phenotype (z). We only focus on an additive genetic genotype-phenotype map here, but maps including epistasis, pleiotropy and dominance are possible. In the additive case, the sum of values of alleles at each locus gives a breeding value (b_v) for each individual at that locus. The sum of breeding values across loci gives a breeding value for the phenotype. Therefore:

$$z = \sum_{v=1}^{n_a} b_v + \varepsilon_{env}(0, \sigma_{env}) \quad (1)$$

Where n_a is the number of additive loci. In our simulations, the environmental contribution (ε_{env}) is assumed to be stochastic and normally distributed, with mean 0 and standard variation σ_{env} . ε_{env} may also be dependent on population density or any other environmental driver (Coulson *et al.* 2017).

The fitness function (ω) defines the phenotype-fitness map and consequently the type of selection influencing the divergence between populations. Once a population has colonized a novel area, new phenotype-environment interactions appear on the phenotype-demography map, shifting the distribution of phenotypes that are expected to have higher fitness (i.e. phenotypic optima). The difference in phenotypic optima between the populations drives the strength of “divergent selection” (grey area, Fig. 1c). Populations exposed to equal phenotypic optima are considered to be under “concordant selection”. The fitness function we use has the form:

$$\omega = b_0 e^{-\frac{1}{2} \left(\frac{4z - b_1 n_a}{b_2 n_a} \right)^2} - b_3 N + \varepsilon_{dem}(0, \sigma_{dem}) \quad (2)$$

The first part on the right-hand side of equation (2) is based on a Gaussian-distribution determining the relation between the phenotypic trait value (z) and fitness (ω). The parameters b_0 , b_1 and b_2 define the maximum number of offspring produced, the phenotypic optima, and the variance of the Gaussian curve, respectively. The second part of equation (2) determines the intensity of density-dependence (b_3) on the fitness of individuals that are members of a population of size N . The final part of the

equation introduces a stochastic demographic variant with mean θ and standard variation σ_{dem} . The last two parts of the equation thus determine the increasing or decreasing variation of fitness due to fluctuations in population size and demographic stochasticity. Any other form of fitness function could be introduced to account for specific relationships between phenotypes (e.g. weight, height, bill size, colour pattern) and the expected number of offspring produced.

The number of breeding events is regulated by the number of females present in the population. Males are randomly selected according to the number of breeding females. The genotypes of both parents participate in the genomic structure of their offspring by transmitting a haploid copy of genetic material. The offspring differs from the parents by carrying half of the genome of each parent and by specific rules defining the recombination rate (r) between homologous chromosomes. We do not explore the effect of new mutations here, because we are primarily interested in the emergence of genomic islands at relative early stages of evolutionary diversification. However, mutation can easily be incorporated by generating a novel polymorphism at a random locus at a given rate per generation (see example in Appendix S1).

The genetic variation of the new generation is determined by the recombination rate during the segregation of haploid gametes of each parent. Segregation starts with a randomly selected copy of a chromosome (i.e. one of the two dimensions of the individual array defining its genotypes). The recombination rate may either be a fixed value between neighbouring loci or may be specified as a per-base-pair rate with crossover points distributed on the physical chromosome via a Poisson process. In the first case, when a recombination map is available, a vector of $n_L - 1$ elements has to be supplied with the recombination rate (r) between each pair of neighbouring loci. Each time a gamete is segregated in the simulation, a recombination occurs (1) or not (0) at a position between loci by comparing the recombination rate (r) between the loci to a uniformly distributed random variable (U) on the unit interval. A separate U is simulated for each pair of neighbouring loci, and $U < r$ means a recombination occurs. There is no recombination between homologous chromosomes when $r = 0$, meaning both loci are completely linked (e.g. within an inversion or situated close to centromeres), while with a value of $r = 0.5$, the recombination rate is completely random (i.e. both loci are very distant on the same chromosome or are located on different chromosomes). A value of $r < 0.5$ means the loci are physically linked. In the second case, the physical positions of the markers on

the chromosome are specified, and then crossovers are simulated as a Poisson process along the physical distance of the chromosome. The simulator allows recombination to be a homogeneous Poisson process, with rate specified by the expected per-base-pair recombination rate (for example 1e-08, corresponding to 1 cM per megabase. This latter method may be preferred when trying to fit a large dataset of genomic information with an unknown recombination rate between neighbouring loci (e.g. Single Nucleotide Polymorphisms). Because we are primarily interested in the effects of various levels of linkage disequilibrium in the formation of islands of genomic divergence, we present results using the first approach, but an example with the second method is also shown in supporting information (Appendix S1).

The offspring represent individuals with the potential to reproduce in the next generation. We assume an equal sex ratio at birth and assign the sexes to offspring by sampling with replacement, with an equal probability of assignment to each sex. A weighted probability could be supplied when unequal sex ratios are considered in the simulations.

The final step is the migration of offspring to neighbouring populations. The probability of migration of each individual is obtained from a uniform distribution, so each individual has the same expected probability of migration. The final number of individuals of population i dispersing to population j is defined by the migration rate m_{ij} . Individuals of i having migration probability smaller than m_{ij} move to population j . A value of $m_{ij} = 0$ means no migration and thus no gene flow between populations, while a value of 0.5 means random migration (and hence random reproduction) between them.

The final number of individuals in population i at time $t+1$ can be estimated as the sum of fitness value of all females present in the population at time t ($N_t^{i,f}$) and the number of migrants from population j (males and females, $N_t^j m_{ji}$):

$$N_{t+1}^i = \sum_{l=1}^{N_t^{i,f}} \omega_{l,t} + N_t^j m_{ji} \quad (3)$$

The model is implemented in a *R* package called “glads” (R Development Core Team 2017), with some functions written in C++ and integrated to *R* by using the Rcpp package (Eddelbuettel *et al.* 2011). The package is available on GitHub (<https://github.com/eriqande/glads>), and is easily modifiable for further applications

simulating data based on allelic information, SNPs, or DNA sequences. An example is available in Appendix S1 and a user manual is available in Appendix S2. Below we describe a number of simulations with different parameterizations to explore how the signatures of genomic divergence are generated by various processes.

Initialization

We start by simulating how two populations of diploid individuals with equal intra-genomic variation at the beginning of the simulations diverge. The migration rate between the two populations was varied across different simulations to explore divergence without gene flow (i.e. $m_{ij} = 0$) and divergence with gene flow (i.e. $m_{ij} \neq 0$). The demographic and genetic parameter values were chosen to describe two fitness functions that can either have identical or contrasting phenotypic optima, but with a similar number of individuals in each population during the simulation (Fig 1c, Fig 1d, Table 1).

The mean population sizes of the two populations were always around 400 (Fig 1d). This is also the initial number of individuals at the beginning of the simulations. The genomic structure of individuals was characterized by genotypes across 300 loci ($n_L = 300$), that were either strongly linked ($r = 0.0001$) or completely unlinked ($r = 0.5$). This range of linkage allows us to explore the dynamic of genomic landscapes across more contiguous or distantly related loci. Because previous simulations on the formation of genomic islands of divergence were restricted to bi-allelic loci (e.g. Feder *et al.* 2012; Flaxman, Feder & Nosil 2013), we ran simulations with a higher number of alleles to allow for greater allelic variation (Table 1). The genomic identities of individuals were randomly assigned at the beginning of each simulation by setting the seed of the random number generator in R. This allow us to explore the same founder population under different evolutionary scenarios.

Fifty of the 300 simulated loci were chosen to have a non-zero and additive variation in allelic contributions to the phenotypic trait value. This fraction of loci is potentially subject to selection. By operating on the phenotype, selection changes the distribution of genotypes at each locus that contributes to the phenotype in the simulation. The remaining 250 loci not influencing phenotypes are neutral and were used to examine the effect of drift and linkage on the appearance of genomic islands. This allowed us to account for both adaptive and neutral evolution simultaneously. The phenotypes were always computed from 50 additive loci ($n_a = 50$), 10 of which

were always linked. These 50 additive loci contributed to the phenotypic trait values of individuals, with the additive value of each allele ranging between 0 and 1. The sum of additive values was then used to compute the phenotype, and then the fitness score, for each individual. However, further studies may expand this procedure to include any required genotype-phenotype map. In summary, we have four classes of genes: i) unlinked genes contributing to the phenotype; ii) linked genes where both loci contribute to the phenotype; iii) unlinked genes that do not contribute to the phenotype; and iv) linked genes that do not contribute to the phenotype. The first two categories of genes are under selection, and the last two are not.

The number of mating pairs depends on the number of breeding females. Female reproductive success was determined first, before male mates were assigned to father each offspring. In this simulation we assumed random mating, although other mating patterns are possible (e.g. Schindler *et al.* 2015). Offspring sex was assigned randomly, with probability 0.5 (Table 1)

Genomic divergence

We measured pairwise F_{ST} at each locus to estimate genetic differentiation between populations. F_{ST} is a widely used measure of differentiation across divergent genomes in studies of genomic islands of divergence (e.g. Ellegren *et al.* 2012; Kusakabe *et al.* 2017). We computed F_{ST} at each simulated locus using the *R* package “*pegas*” (Paradis 2010). Genetic differentiation averaged across multiple loci was calculated using the approach of Nei (1973), as implemented in the *R* package “*mmod*” (Winter 2012).

We identified significant genomic islands of divergence using the approach of Marques *et al.* (2016), modified from the original method of Hofer, Foll and Excoffier (2012). Marques *et al.* (2016) used Hidden Markov Models (HMM) to identify three hidden states: low, intermediate and high genomic differentiation. The regions of intermediate states are considered “genomic background” that arise randomly under a hierarchical island model. The regions of extremely high differentiation are considered to be genomic islands of divergence. The Baum-Welch algorithm is applied from 1000 different randomly chosen initial values for the parameters, and the final parameter estimates are taken to be those from the run producing the highest likelihood. Subsequently, using those estimates, the Viterbi algorithm is used to assign the most likely sequence of states across the genome (Hofer, Foll & Excoffier

2012). The significance of the clustering was assessed by randomly permuting islands 10,000 times across the whole simulated genomic area.

Simulations

We conducted simulation experiments using a wide suite of parameter values (see Table S1, supporting information). These were designed to examine how various scenarios of linkage between loci, drift, selection, and time since divergence influence the formation of genomic islands of divergence in both the presence and absence of gene flow. Parameter values are presented in Table 1 and the supplementary information provides more details for the choice of each parameter set (Table S1). The first simulations characterise the effect of the founder population on the resulting genetic divergence (F_{st}) at an early stage of independent evolution (100 generations, $m_{ij} = 0$). We then explored in more detail, the effect of physical linkage, gene flow and time since initial divergence. Drift is included in all simulations through the group of genes that are not involved with the phenotype trait value, and through the random selection of gametes at birth. We assigned a name to each group of simulations and will briefly describe their structure.

1. Random sampling of founders and concordant selection: these simulations were designed to examine how random sampling of the founder population influenced genomic divergence. Both populations were exposed to neutral evolution and concordant selection, with identical phenotypic optima (equal to population 1, Table 1). We ran 50 simulations with different random initial founder genotypes. Founder genotypes were determined by sampling a uniform distribution with replacement.

2. Random sampling of founders and divergent selection: We considered the same 50 founder populations as before but added divergent selection. The selective pressures generating evolutionary divergence between populations were generated by their respective fitness functions. The amount of difference between phenotypic optima measures the strength of “divergent selection” (Fig 1c, Table 1).

3. Levels of heterozygosity in the founder population: Our third set of simulations was designed to explore how variable levels of heterozygosity among founder populations influenced the variance of genomic divergence at the end of the simulation. The level of heterozygosity in the founder population was varied by sampling alleles at a locus with variable frequencies of replacement (see Table S1 for more information). The variable frequency of replacement represented the weighted

probability of a random sampling with replacement among the 20 polymorphisms available for each locus. This ranged from 1 (an equal probability of allelic sampling and more heterozygous) to 100 (an unequal probability of allelic sampling and more homozygous). As this value becomes higher, it increases the probability for individuals to carry the same allele on both copies of their genes. Because linked loci are hypothesised to be more likely to be involved in the formation of genomic islands and we are analysing this factor separately, we excluded these loci in the final estimation of variability of genetic differentiation (an analysis without the exclusion of linked loci is included in supporting information).

4. Physical genomic linkage: Having characterised how initial conditions might influence results, we next examined the effect of physical linkage on the formation of islands of genomic divergence. We ran 100 simulations with equal founder populations, but changed the recombination rate between linked loci, ranging from nearly complete linkage ($r = 0.0001$) to no linked loci ($r = 0.5$).

5. Strong selection at a single, unlinked locus: We next explored the effect of strong selection on unlinked genes of large phenotypic effect. Fifty additive loci contributed to phenotypic expression, but one locus contributed 10 times more than the others. This means that rather than having multiple linked loci affecting the trait there is, in particular, one locus of very large effect that is unlinked to the other loci that influence phenotype. We considered the same founder population as in 4 (genomic linkage).

6. Time since divergence with and without gene flow: To explore how time since divergence influenced the formation of genomic islands, we ran simulations with the same 50 founder populations of our previous analysis “random sampling of founders and divergent selection”, but for different lengths of time: 100, 500, 1000 and 2000 generations. We repeated these simulations in the presence ($m = 0.01$) and absence ($m = 0$) of gene flow.

7. Physical linkage and gene flow: Finally, we explored how linkage and gene flow combined to influence the formation of genomic islands. We simulated various rates of migration and recombination, using the same 50 founder populations as in 4 (physical genomic linkage). We recorded average F_{st} at 10 linked loci affecting the expression of phenotypes (positions 150 to 159) and 10 independent loci not related to phenotype expression (positions 90 to 99). This allowed us to determine the

magnitude of differentiation between regions of linked divergent selection and the genomic background of neutral evolution.

Results

Random sampling of founders and concordant selection

Our first simulations explored the effect of initial conditions on divergence and the formation of genomic islands under equal selective pressures (i.e. concordant selection). The sky-blue lines in Fig. 2a show the resulting F_{st} values for all 50 random initial populations. When considering the average F_{st} values by loci across the 50 pairwise comparisons, linked genes that contribute to the phenotype have a slightly higher F_{st} than unlinked genes (dark blue line, Fig. 2b). However, independent of the type of loci (i.e. under selection or neutral), all positions have almost the same probability of becoming an area of higher or lower genomic divergence.

Different genotypes coding for identical phenotypes influence the dynamics of genetic differentiation with time (Fig. 2a). F_{st} values across the whole genome ranged between 0 and about 0.3. Interactions between the genotype-phenotype map and the phenotype-demographic map influence the development of genetic differentiation between populations. The dark blue line in Fig. 2a represents an example of a single founder population with a typical, heterogeneous genomic landscape that has formed over 100 generations. The Hidden Markov Model identified genomic islands in various, but not all of the 50 founder populations (Fig. S2, supporting information). These areas of higher genomic divergence appear seemingly randomly across the whole genome. Among all simulations that generated genomic islands (84% of simulations), 14% did not generate islands on selected linked loci affecting phenotypes. The variance in F_{st} we observed within and across loci reveal that the genotype-phenotype map of the founder populations influences the patterns of genomic divergence.

In our simulations, populations differentiate after 100 generations, even when both populations have equal phenotypic optima under concordant selection (Fig 2a). This is because there are many ways to generate the same additive phenotypic trait value. The time since initial divergence increases the likelihood of generating these different outcomes, therefore with enough generations of isolated reproduction,

populations can still be highly differentiated even when they are exposed to the same fitness peak.

Random sampling of founders and divergent selection

Genomic islands of divergence are, on average, more likely to be observed for genes that contribute to a phenotypic trait that experiences divergent selection across the two populations (orange line, Fig 2b). The range of change of F_{st} values was also higher under divergent selection than under concordant selection (Fig. 2c). The values of F_{st} across loci ranged between 0 and about 0.8, and this seemed to affect the average F_{st} across non-selected loci (compare orange and blue line, Fig. 2b). The same single founder population illustrated in Fig 2a and 2c (blue and orange lines) provides an example of where genomic islands usually form at some linked loci experiencing divergent selection. The Hidden Markov Model identified genomic islands in various, but not all of the 50 founder populations (Fig. S3, supporting information). A single high island of genomic divergence usually emerges at selected loci, but some islands appeared in neutral loci and other populations did not generate any pattern of genomic islands (Fig. S3, supporting information). Among all simulations that generated genomic islands (84% of simulations), only 2% did not generate islands on selected linked loci affecting phenotypes. In general, due to the large variation between simulations, divergent selection did not necessarily generate islands of genomic divergence at loci under selection (yellow lines, Fig. 2c and Fig. S3, supporting information).

Levels of heterozygosity in the founder population

As the variance of heterozygosity in the founder population increases, so too does the variance in F_{st} across the genome after 100 generations of independent evolution (Fig. 2d). This variance reflects an increase in F_{st} of loci not under selection. F_{st} at these loci can be as large as for genes under direct selection. A similar pattern is observed when physically linked selected loci are included in the analysis (Fig. S4, supporting information). This result reveals that the appearance of a pattern of genomic islands at early stages of differentiation can be caused by genetic variation at specific loci in the founder populations.

Physical genomic linkage

We ran simulations with the same founder population and parameter values used to generate the dark yellow line in Fig. 2c that resulted in an island of genomic divergence, except now we varied the recombination rate (r) among linked selected genes. The average F_{st} of those linked genes was much higher with nearly complete levels of linkage ($r < 0.02$), but tended to the average value of neutral genes when the recombination rate was higher, even when they were still physically linked (compare Fig. 2e and Fig. 2b). These results show that strong linkage may facilitate the appearance of genomic islands when those genes are affected by divergent selection, even in the absence of gene flow. Extreme physical linkage therefore tends to increase the F_{st} value of genes under selection. The combined effect of divergent selection and linkage is consequently important for the development of genomic islands of larger sizes.

Strong selection at a single, unlinked locus:

The previous simulations revealed that divergent selection on linked selected loci could sometime result in islands of genomic divergence. We therefore next considered a founder population in which an island of genomic divergence formed (Fig. 2c), yet altered the genotype-phenotype map such that one independent locus ($r = 0.5$) contributed disproportionately to the phenotypic value. This locus resulted in a level of F_{st} of more than twice that observed elsewhere in the genome, including on linked selected genes (Fig. 2f). This result reveals that patterns of genomic divergence are not necessarily determined by strongly linked genes of similar effect, but can also emerge when one gene of large effect is linked to other markers.

Time since divergence with and without gene flow

We extended our 50 previous simulations of “random sampling of founders and divergent selection” by running them for longer (100 to 2,000 generations). Without gene flow, the trend of higher genomic differentiation in selected loci, particularly in genes that are linked, is more evident at early stages of divergence (100 generations, Fig. 3a). The length of time that independent evolution has to act influences genome-wide divergence, masking signals of genomic islands that arise from single or linked loci. The pattern of heterogeneous genomic differentiation is therefore less evident and tends to disappear as the numbers of generations since divergence without gene flow increases (2,000 generations, Fig. 3a).

Physical linkage and gene flow

All previous simulations were performed in the absence of gene flow. Gene flow increases the number of generations over which genomic islands of divergence are apparent. The genomic islands of higher F_{st} are still present after 2,000 generations, when performing the same simulations as in figure 3a, but allowing a level of migration between populations ($m = 0.01$, Fig. 3b). However, as the level of gene flow increases, the prevalence of islands of genomic divergence decreases (see the zero values in Fig. 3c).

We performed the simulations of divergent selection using the same 50 founder populations (Fig. 2a and 2c), while varying both the migration rate between populations and the recombination rate among linked loci. The numbers inside the grey squares in Fig. 3c indicate the magnitude of difference between independent neutral loci (i.e. genomic background) and linked selected loci (i.e. genomic islands). The largest differences were present under conditions with extreme physical linkage ($r < 0.04$) and a low migration rate ($m < 0.02$), and ranged from 0.1 to 0.2. Those differences are negligible under concordant selection (Fig S2). Overall, these results show that gene flow may influence the persistence of genomic islands but is not the only factor determining their emergence.

Discussion

Genomic islands of divergence

The application of our quantitative framework to model the generation of genomic islands of divergence has revealed that while there are several routes that can result in genomic islands, the conditions required to generate islands are relatively narrow, and importantly, there is no single set of circumstances that guarantee their emergence. For instance, formation of large genomic islands requires a combination of divergent selection and strong linkage, regardless of the gene flow scenario. In contrast smaller genomic islands can form via drift in the early stages of divergence in particular. However, in both cases, genomic islands can also fail to form even when these conditions are met, because outcomes are highly dependent on the initial genetic composition of the diverging populations. Our simulations suggest that genomic islands are most obvious during the early stages of divergence, and tend to disappear with the accumulation of genome-wide divergence over time. If present, gene flow

can slow this loss up to a point, however including gene flow is not necessary to explain genomic island formation. The importance of evolutionary processes that were modelled (divergent selection, drift, gene flow), along with influencing factors of initial genetic composition, degree of genetic linkage, time since divergence and their interactions are summarised in Figure 4. In addition to these factors, previous simulation studies have also highlighted the challenge of identifying genomic islands due to sampling design (Lotterhos & Whitlock 2014) or the statistic used to identify divergence (Matthey-Doret & Whitlock 2018). Our modelling approach provides a nuanced understanding of how genomic islands arise, yet it is not possible to confidently interpret a particular process from genomic data on its own (Turner, Hahn & Nuzhdin 2005; Nosil 2008; Feder *et al.* 2013; Seehausen *et al.* 2014; Jiggins & Martin 2017; Nosil *et al.* 2017). This is particularly true in the case of divergence in the presence of gene flow and divergent selection, which is not a necessary condition for the emergence of genomic islands, as concluded by initial studies in the field (Nosil 2008). Recent simulation studies have also highlighted the importance of multiple processes, in addition to gene flow and selection, such as drift and linkage, in generating genomic islands (Yeaman, Aeschbacher & Bürger 2016). While this complexity is already known, our contribution is to show the relationship among those factors, highlighting that its presence does not guarantee the emergence of this pattern, because the time since divergence and the genetic composition at the beginning of the divergence process are also important.

Initial genetic composition and drift influence the generation of genomic islands

Our model revealed two ways that random effects can influence the formation of genomic islands of divergence. First, a previously unappreciated but critical factor influencing their generation was the genetic composition of the initial populations. This was evident from comparison of simulations with identical parameter values but different starting populations i.e. different genetic composition. In some simulations, genomic islands were generated and in others they failed to form. Furthermore, the starting values influenced island appearance even in regions of the genome that were not influenced by selection, linkage or gene flow – all of which are thought to be important in genomic island formation as discussed below (Feder *et al.* 2013; Flaxman, Feder & Nosil 2013). Second, drift alone could generate a pattern of numerous islands of small size particularly in the early stages of divergence. Lineage

616 sorting is a slow process when explained only by drift (Cruickshank & Hahn 2014), in
617 which case ancestral polymorphisms remain shared between recently divergent taxa,
618 reducing genetic diversity and potentially inducing a pattern of genomic islands (Ma
619 *et al.* 2018). Recent studies exploring the distribution of genomic islands have also
620 advanced the idea that islands arise from neutral processes without a major
621 contribution from divergent selection (Campagna *et al.* 2015; Wang *et al.* 2016) and
622 the results of our model identify the scenarios where this is particularly likely to be
623 the case. Some studies document a small number of very prominent islands (e.g.
624 Turner, Hahn & Nuzhdin 2005; Wang *et al.* 2016), however finding multiple islands
625 of low relief is also common (e.g. Ellegren *et al.* 2012; Ruegg *et al.* 2014; Soria-
626 Carrasco *et al.* 2014; Feulner *et al.* 2015). Furthermore, comparisons often involve
627 recently diverged populations (e.g. Nadeau *et al.* 2012; Via 2012; Ruegg *et al.* 2014),
628 with some divergence timescales as short as 100 generations (e.g. Marques *et al.*
629 2016). Our modelling suggests that these patterns and types of comparisons could be
630 explained without recourse to explanations that invoke selection, but initial diversity
631 of the founder populations and drift.

632 Another scenario that may be particularly prone to stochastic effects is where
633 one of the diverging populations experiences a geographic expansion. Klopstein,
634 Currat and Excoffier (2006) suggest that the effect of drift is stronger in expanding
635 populations because of “allelic surfing”, where alleles that happen to be at the
636 expansion front may incidentally increase in frequency (see Hofer, Foll & Excoffier
637 2012; Excoffier, Quilodrán & Currat 2014). This, in turn, impacts genetic
638 composition and variation (i.e. Arenas *et al.* 2012; Branco & Arenas 2018; Branco *et al.*
639 2018), and if occurring very early during the divergence process, the combination
640 of early differences in genetic composition and drift could generate highly stochastic
641 patterns of islands of divergence.

642 We have shown that populations diverge through time even under the equal
643 selective pressures of concordant selection. Indeed, highly polygenic traits may also
644 express divergence based on which alleles of the genes under selection in the founder
645 population end up increasing in frequency. Selection will tend to create shorter
646 coalescence times around those selected loci, meaning a lower effective population
647 size and hence greater drift (Nordborg 1997). However, it should be noted that the
648 specification of the additive genotype-phenotype map we use means there are
649 multiple genotypes that will produce the same phenotypic value. This explains why

populations with identical selection regimes can diverge, with some developing islands of genomic divergence, and others not. The nature of our genotype-phenotype map in the model could also underpin the influence of initial genetic composition on our results. Future work will explore whether the same conclusions hold with genotype-phenotype maps that do not assume small additive contributions to the phenotype from genotypes at multiple loci. However, the genotype-phenotype map we use is widely assumed in quantitative genetics, and given that many traits are highly polygenic, is an appropriate initial map to assume in simulations.

Linkage and divergent selection generate islands of divergence independent of gene flow

Extreme linkage in combination with divergent selection was necessary, though not sufficient, for the development of the most prominent genomic islands, regardless of whether gene flow occurred or not. These findings are consistent with observations of prominent genomic islands between populations presumed to be under strong divergent selection, and not connected by gene flow (Burri *et al.* 2015; Zhang *et al.* 2017). The occurrence of candidate genes, hypothesised to be under natural selection, associated with genomic islands of divergence also supports the role of selection (Sousa & Hey 2013; Kusakabe *et al.* 2017). However, empirical results also provide examples where SNPs under selection are not associated with islands of divergence (e.g. Ruegg *et al.* 2014; Han *et al.* 2017; Riesch *et al.* 2017).

The importance of linkage in the appearance of genomic islands has been highlighted in both theoretical and empirical studies (Feder & Nosil 2010; Renaut *et al.* 2013; Flaxman *et al.* 2014), with extreme linkage, such as that found near centromeres or within genomic inversions, often associated with the most prominent genomic islands of divergence (Feder & Nosil 2009; Ellegren *et al.* 2012; Kawakami *et al.* 2014). Selection acting in these zones of low rates of recombination (i.e. linked selection) reduces the effective population size of these genomic regions to a greater degree than in the rest of the genomes, generating genomic islands (Feder & Nosil 2009; Turner & Hahn 2010).

We did not explore the effect of linkage on deleterious variants (i.e. background selection) in our simulations. However, previous studies have shown that selection on both adaptive and deleterious mutations has a similar effect of reducing within population diversity (Nordborg, Charlesworth & Charlesworth 1996; Slatkin &

Wiehe 1998; Comeron 2014), and influencing the formation of genomic islands (Cruickshank & Hahn 2014).

The effect of gene flow on generation and persistence of genomic islands

The idea that genomic islands of divergence were generated primarily by antagonistic effects of divergent selection and gene flow was a favoured explanation until recently (Turner, Hahn & Nuzhdin 2005; Nosil 2008; Feder, Egan & Nosil 2012). According to this mechanism, genomic islands form around selected loci involved with the divergence process, and genes physically linked to them, while adjacent neutral or weakly selected regions are homogenised by gene flow (Turner & Hahn 2010; Flaxman, Feder & Nosil 2013; Kawakami *et al.* 2014). Our modelling provides further support that the presence of gene flow is not an essential condition, however, an additional insight is that when gene flow does occur, it can lengthen the time that genomic islands are visible. Verbal models of changing genomic landscapes over time predicted that genomic islands would disappear with the accumulation of genome-wide divergence over time (Wu & Ting 2004; Nosil 2012; Nosil & Feder 2012). Empirical support that this is indeed the case is provided from studies where genomic islands are more frequently documented in recently diverged versus distantly related taxa (e.g. Nadeau *et al.* 2012; Via 2012; Marques *et al.* 2016). Our results reveal that this dynamic can be moderated by gene flow, where a limited amount of gene flow serves to slow down the swamping of genomic islands over time, whereas large amounts of gene flow tend to erase the pattern of islands of divergence altogether.

Advantage and limitations of our framework

Our framework is designed to identify how various processes can generate patterns of divergence between populations. Discriminating whether divergence has occurred in the presence or absence of gene flow can be challenging (Feder *et al.* 2013), and our simulation tool provide a useful way to discriminate between the two scenarios. There is one other package available in R to simulate the genetics and demography of populations, but it is restricted to one locus (Andrello & Manel 2015). However, in other programming languages, there is a range of individual based tools simulating the evolution of large genetic datasets forward in time (e.g. Peng & Kimmel 2005; Currat *et al.* 2015; Haller & Messer 2019), but none of these allow simulation via a fitness function as presented here, which is important to explore the range of

ecological and evolutionary dynamics associated with the divergence between populations. We are confident our approach works, as it gave results consistent with SimuPOP, a tool widely used in ecology and evolution (Fig. S6, supporting information) (Peng & Kimmel 2005; Peng & Amos 2008).

The main limitation of our simulation approach lies in the amount of genetic and ecological information required to parameterize it for a field system. Empirical information is needed to identify fitness functions, specify the genotype-phenotype map or estimate rates of migration. Model organisms with short generation time that have been extensively studied in the past represent a source of data for potential application (e.g. Mackay 2014). Adapting parameter values from sister species for which information is available may be of use for non-model organisms. This limitation is expected to become less important in the future as the rapid accumulation of freely available ecological and genomic datasets grow (Jones *et al.* 2008; Ellegren 2014). However, in the absence of sufficient information to parametrize a fitness function, this framework is still useful to elucidate neutral evolution, which can be simulated in the framework by replacing the fitness function with a random distribution (e.g. Poisson) in order to generate the next generation of offspring (see example in Appendix S1). While mutation is not explored here, as we were mostly interested in divergence at relatively early stages of evolution, its incorporation would not likely change any of the patterns observed in this study (see example in Appendix S1).

The elapsed time of simulations may be a potential limitation for studies simulating a huge number of individuals and genetic information. The time depends on the number of individuals, the number of simulated generations and the amount of genetic information in the analysis. For instance, simulation of a single population of 100 individuals, with 100 SNPs, over 100 generations take approximately 0.2 seconds, while the same analysis takes one second for 1000 SNPs and approximately 10 seconds when simulating 1000 SNPs over 1000 generations (see Fig. S7, supporting information). However, the computation time may be largely reduced by parallelising the number of required simulations by the number of available cores in the computer. For instance, a single simulation of divergence between two populations, composed of 100 individuals with 1000 SNPs, as the one presented in figure S6 (supporting information), takes around 2.58 seconds, while 8 simulations using 8 cores parallelised with the library “parallel” of R takes 5.9 seconds (0.7

seconds by simulation). All analyses of computation time were performed by using the graphical version of R, in a MAC OS X 10.13 64 bits with 3.1GHZ CPU.

Conclusions

We have developed a quantitative framework to explore the dynamics of genomic landscapes and identify how various processes can generate patterns of divergence between populations. Our work builds on previous insights (Charlesworth, Nordborg & Charlesworth 1997; Feder & Nosil 2009; Flaxman, Feder & Nosil 2013; Akerman & Bürger 2014; Sedghifar, Brandvain & Ralph 2016). We have been able to demonstrate that the formation of genomic islands of divergence is not a deterministic phenomenon, but that they can arise via a number of routes. Those routes are not mutually exclusive and further studies may explore how the interaction of multiple processes at once may influence the observation of a pattern of genomic islands. We urge extreme caution in inferring a particular ecological or evolutionary process when a particular genomic pattern is observed. Narrowing down the potential cause of a particular signature will likely require ancillary information beyond the genome sequence or modelling exercises that examine the processes that have the potential to generate such a pattern. The methods described here provide a modelling framework which helps to depict such signatures of past evolution, as well as potential routes of future evolution for any divergent taxa.

Acknowledgements: This study was financed by a fellowship from the Swiss National Science Foundation (n°P2GEP3_168973 and P400PB_183930 to CSQ). We thank Laurent Excoffier and David Marques for their support on the Hidden Markov Model method. We also thank Jose Manuel Nunes for his suggestions regarding the R package code. Finally, we are also grateful to two anonymous reviewers for their comments and suggestions on an earlier version of this manuscript.

Authors' contributions: S.C, T.C., E.A, K.R., A.T.S.P and C.S.Q. conceived the original idea and designed the experiments. C.S.Q., E.A. and T.C. coded the R package.

C.S.Q. performed the experiments and wrote the first version of the manuscript. All authors analysed the data and edited the manuscript.

Data Availability: All data necessary to repeat the simulations are described in the main text and supporting information. The R package code is available on GitHub: <https://github.com/eriqande/glads>.

Literature Cited

- Akerman, A. & Bürger, R. (2014) The consequences of gene flow for local adaptation and differentiation: a two-locus two-deme model. *Journal of mathematical biology*, **68**, 1135-1198.
- Andrello, M. & Manel, S. (2015) MetaPopGen: an r package to simulate population genetics in large size metapopulations. *Molecular Ecology Resources*, **15**, 1153-1162.
- Arenas, M., François, O., Currat, M., Ray, N. & Excoffier, L. (2012) Influence of admixture and paleolithic range contractions on current European diversity gradients. *Molecular Biology and Evolution*, **30**, 57-61.
- Bradshaw, A.D. (1965) Evolutionary significance of phenotypic plasticity in plants. *Advances in genetics*, **13**, 115-155.
- Branco, C. & Arenas, M. (2018) Selecting among Alternative Scenarios of Human Evolution by Simulated Genetic Gradients. *Genes*, **9**, 506.
- Branco, C., Velasco, M., Benguigui, M., Currat, M., Ray, N. & Arenas, M. (2018) Consequences of diverse evolutionary processes on american genetic gradients of modern humans. *Heredity*, **121**, 548.
- Burri, R., Nater, A., Kawakami, T., Mugal, C.F., Olason, P.I., Smeds, L., Suh, A., Dutoit, L., Bureš, S. & Garamszegi, L.Z. (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome research*, **25**, 1656-1665.
- Campagna, L., Gronau, I., Silveira, L.F., Siepel, A. & Lovette, I.J. (2015) Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology*, **24**, 4238-4251.
- Chanderbali, A.S., Berger, B.A., Howarth, D.G., Soltis, P.S. & Soltis, D.E. (2016) Evolving ideas on the origin and evolution of flowers: new perspectives in the genomic era. *Genetics*, **202**, 1255-1265.
- Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical research*, **70**, 155-174.
- Comeron, J.M. (2014) Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS genetics*, **10**, e1004434.
- Coulson, T., Kendall, B.E., Barthold, J., Plard, F., Schindler, S., Ozgul, A. & Gaillard, J.-M. (2017) Modeling adaptive and nonadaptive responses of populations to environmental change. *The American Naturalist*, **190**, 313-336.
- Coulson, T., Potter, T. & Felmy, A. (2019) Fitness functions, genetic and non-genetic inheritance, and why ecological dynamics and evolution are inevitably linked. *bioRxiv*, 762658.
- Cruickshank, T.E. & Hahn, M.W. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133-3157.
- Currat, M., Gerbault, P., Di, D., Nunes, J.M. & Sanchez-Mazas, A. (2015) Forward-in-Time, Spatially Explicit Modeling Software to Simulate Genetic Lineages Under Selection. *Evolutionary bioinformatics*, **11**, 27-30.

833 Darwin, C. & Wallace, A. (1858) On the tendency of species to form varieties; and on the perpetuation
 834 of varieties and species by natural means of selection. *Zoological Journal of the Linnean*
 835 *Society*, **3**, 45-62.
 836 Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J. & Bates, D.
 837 (2011) Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1-18.
 838 Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends in*
 839 *Ecology & Evolution*, **29**, 51-63.
 840 Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen,
 841 H., Nadachowska-Brzyska, K. & Qvarnström, A. (2012) The genomic landscape of species
 842 divergence in *Ficedula* flycatchers. *Nature*, **491**, 756-760.
 843 Ellner, S.P., Childs, D.Z. & Rees, M. (2016) *Data-driven modelling of structured populations*.
 844 Springer.
 845 Excoffier, L., Quilodrán, C.S. & Currat, M. (2014) Models of hybridization during range expansions
 846 and their application to recent human evolution. *Cultural Developments in the Eurasian*
 847 *Paleolithic and the Origin of Anatomically Modern Humans* (eds A. Derevianko & M.
 848 Shunkov), pp. 122-137. Department of the Institute of Archaeology and Ethnography SB
 849 RAS, Novosibirsk, Russia.
 850 Feder, J.L., Egan, S.P. & Nosil, P. (2012) The genomics of speciation-with-gene-flow. *Trends in*
 851 *Genetics*, **28**, 342-350.
 852 Feder, J.L., Flaxman, S.M., Egan, S.P., Comeault, A.A. & Nosil, P. (2013) Geographic mode of
 853 speciation and genomic divergence. *Annual Review of Ecology, Evolution, and Systematics*,
 854 **44**, 73-97.
 855 Feder, J.L., Gejji, R., Yeaman, S. & Nosil, P. (2012) Establishment of new mutations under divergence
 856 and genome hitchhiking. *Phil. Trans. R. Soc. B*, **367**, 461-474.
 857 Feder, J.L. & Nosil, P. (2009) Chromosomal inversions and species differences: when are genes
 858 affecting adaptive divergence and reproductive isolation expected to reside within inversions?
 859 *Evolution*, **63**, 3061-3075.
 860 Feder, J.L. & Nosil, P. (2010) The efficacy of divergence hitchhiking in generating genomic islands
 861 during ecological speciation. *Evolution*, **64**, 1729-1747.
 862 Feulner, P.G., Chain, F.J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., Lenz, T.L., Samonte,
 863 I.E., Stoll, M. & Bornberg-Bauer, E. (2015) Genomics of divergence along a continuum of
 864 parapatric population differentiation. *PLoS genetics*, **11**, e1004966.
 865 Flaxman, S.M., Feder, J.L. & Nosil, P. (2013) Genetic hitchhiking and the dynamic buildup of genomic
 866 divergence during speciation with gene flow. *Evolution*, **67**, 2577-2591.
 867 Flaxman, S.M., Wacholder, A.C., Feder, J.L. & Nosil, P. (2014) Theoretical models of the influence of
 868 genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074-4088.
 869 Fraïsse, C., Roux, C., Welch, J.J. & Bierne, N. (2014) Gene-flow in a mosaic hybrid zone: is local
 870 introgression adaptive? *Genetics*, **197**, 939-951.
 871 Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W. & Ralph, P.L. (2019) Tree - sequence recording
 872 in SLiM opens new horizons for forward - time simulation of whole genomes. *Molecular*
 873 *Ecology Resources*, **19**, 552-566.
 874 Haller, B.C. & Messer, P.W. (2019) SLiM 3: Forward genetic simulations beyond the Wright-Fisher
 875 model. *Molecular Biology and Evolution*, **36**, 632-637.
 876 Han, F., Lamichhaney, S., Grant, B.R., Grant, P.R., Andersson, L. & Webster, M.T. (2017) Gene flow,
 877 ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence
 878 among Darwin's finches. *Genome research*.
 879 Harr, B. (2006) Genomic islands of differentiation between house mouse subspecies. *Genome*
 880 *research*, **16**, 730-737.
 881 Hofer, T., Foll, M. & Excoffier, L. (2012) Evolutionary forces shaping genomic islands of population
 882 differentiation in humans. *BMC genomics*, **13**, 107.
 883 Hohenlohe, P.A., Bassham, S., Currey, M. & Cresko, W.A. (2012) Extensive linkage disequilibrium
 884 and parallel adaptive divergence across threespine stickleback genomes. *Phil. Trans. R. Soc.*
 885 *B*, **367**, 395-408.
 886 Hughes, A.L. (2009) Evolution in the post-genome era. *Perspectives in biology and medicine*, **52**, 332-
 887 337.
 888 Jiggins, C. & Martin, S. (2017) Glittering gold and the quest for Isla de Muerta. *Journal of*
 889 *Evolutionary Biology*, **30**, 1509-1511.

- Jones, O.R., Clutton - Brock, T., Coulson, T. & Godfray, H.C.J. (2008) A web resource for the UK's long - term individual - based time - series (LITS) data. *Journal of Animal Ecology*, **77**, 612-615.
- Kawakami, T., Backström, N., Burri, R., Husby, A., Olason, P., Rice, A.M., Ålund, M., Qvarnström, A. & Ellegren, H. (2014) Estimation of linkage disequilibrium and interspecific gene flow in *Ficedula* flycatchers by a newly developed 50k single - nucleotide polymorphism array. *Molecular ecology resources*, **14**, 1248-1260.
- Kelleher, J., Etheridge, A.M. & McVean, G. (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, **12**, e1004842.
- Klopfstein, S., Currat, M. & Excoffier, L. (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482-490.
- Kokko, H., Chaturvedi, A., Croll, D., Fischer, M.C., Guillaume, F., Karrenberg, S., Kerr, B., Rolshausen, G. & Stapley, J. (2017) Can Evolution Supply What Ecology Demands? *Trends in Ecology & Evolution*.
- Kusakabe, M., Ishikawa, A., Ravinet, M., Yoshida, K., Makino, T., Toyoda, A., Fujiyama, A. & Kitano, J. (2017) Genetic basis for variation in salinity tolerance between stickleback ecotypes. *Molecular ecology*, **26**, 304-319.
- Lohse, K. (2017) Come on feel the noise—from metaphors to null models. *J. Evol. Biol.*, **30**, 1506-1508.
- Lotterhos, K.E. & Whitlock, M.C. (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178-2192.
- Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., Feng, J., Jiang, D., Ahani, H. & Abbott, R.J. (2018) Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proceedings of the National Academy of Sciences*, **115**, E236-E243.
- Mackay, T.F. (2014) Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, **15**, 22.
- Marques, D.A., Lucek, K., Meier, J.I., Mwaiko, S., Wagner, C.E., Excoffier, L. & Seehausen, O. (2016) Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS Genet*, **12**, e1005887.
- Matthey-Doret, R. & Whitlock, M.C. (2018) Background selection and the statistics of population differentiation: consequences for detecting local adaptation. *bioRxiv*, 326256.
- Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A., Joron, M., Blaxter, M.L. & Mallet, J. (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B*, **367**, 343-353.
- Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, **70**, 3321-3323.
- Noor, M.A. & Bennett, S.M. (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439.
- Nordborg, M. (1997) Structured coalescent processes on different time scales. *Genetics*, **146**, 1501-1514.
- Nordborg, M., Charlesworth, B. & Charlesworth, D. (1996) The effect of recombination on background selection. *Genetics Research*, **67**, 159-174.
- Nosil, P. (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103-2106.
- Nosil, P. (2012) Ecological speciation: Oxford series in ecology and evolution. Oxford University Press Oxford.
- Nosil, P. & Feder, J.L. (2012) Genomic divergence during speciation: causes and consequences. The Royal Society.
- Nosil, P., Feder, J.L., Flaxman, S.M. & Gompert, Z. (2017) Tipping points in the dynamics of speciation. *Nature ecology & evolution*, **1**, 0001.
- Nosil, P., Funk, D.J. & Ortiz - Barrientos, D. (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375-402.
- Paradis, E. (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419-420.
- Peng, B. & Amos, C.I. (2008) Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics*, **24**, 1408-1409.

- Peng, B. & Kimmel, M. (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**, 3686-3687.
- R Development Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlović, M., Noor, M., Mehlig, B. & Westram, A. (2017) Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, **30**, 1450-1477.
- Renaut, S., Grassa, C., Yeaman, S., Moyers, B., Lai, Z., Kane, N., Bowers, J., Burke, J. & Rieseberg, L. (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.
- Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A.A., Farkas, T.E., Lucek, K., Hellen, E., Soria-Carrasco, V. & Dennis, S.R. (2017) Transitions between phases of genomic differentiation during stick-insect speciation. *Nature ecology & evolution*, **1**, 0082.
- Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N. & Bierne, N. (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. *Plos Biology*, **14**, e2000234.
- Ruegg, K., Anderson, E.C., Boone, J., Pouls, J. & Smith, T.B. (2014) A role for migration - linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, **23**, 4757-4769.
- Samuk, K., Owens, G.L., Delmore, K.E., Miller, S.E., Rennison, D.J. & Schluter, D. (2017) Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology*, **26**, 4378-4390.
- Schindler, S., Gaillard, J.M., Grüning, A., Neuhaus, P., Traill, L.W., Tuljapurkar, S. & Coulson, T. (2015) Sex - specific demography and generalization of the Trivers-Willard theory. *Nature*, **526**, 249.
- Sedghifar, A., Brandvain, Y. & Ralph, P. (2016) Beyond clines: lineages and haplotype blocks in hybrid zones. *Molecular ecology*, **25**, 2559-2576.
- Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C. & Brännström, Å. (2014) Genomics and the origin of species. *Nature Reviews. Genetics*, **15**, 176.
- Slatkin, M. & Wiehe, T. (1998) Genetic hitch-hiking in a subdivided population. *Genetics Research*, **71**, 155-160.
- Soria-Carrasco, V., Gompert, Z., Comeault, A.A., Farkas, T.E., Parchman, T.L., Johnston, J.S., Buerkle, C.A., Feder, J.L., Bast, J. & Schwander, T. (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738-742.
- Sousa, V. & Hey, J. (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404-414.
- Turner, T.L. & Hahn, M.W. (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, **19**, 848-850.
- Turner, T.L., Hahn, M.W. & Nuzhdin, S.V. (2005) Genomic islands of speciation in *Anopheles gambiae*. *Plos Biology*, **3**, e285.
- 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**, 451-460.
- Wang, L., Wan, Z.Y., Lim, H.S. & Yue, G.H. (2016) Genetic variability, local selection and demographic history: genomic evidence of evolving towards allopatric speciation in Asian seabass. *Molecular Ecology*, **25**, 3605-3621.
- Winter, D.J. (2012) MMOD: an R library for the calculation of population differentiation statistics. *Molecular Ecology Resources*, **12**, 1158-1160.
- Wu, C.-I. & Ting, C.-T. (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114.
- Wu, C.I. (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851-865.
- Yeaman, S., Aeschbacher, S. & Bürger, R. (2016) The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular Ecology*, **25**, 2542-2558.
- Zhang, D., Song, G., Gao, B., Cheng, Y., Qu, Y., Wu, S., Shao, S., Wu, Y., Alström, P. & Lei, F. (2017) Genomic differentiation and patterns of gene flow between two long - tailed tit species (*Aegithalos*). *Molecular Ecology*.

1005 **Table 1.** List of parameters of the model with default values

Symbol	Definition	Value [†]
N_i	Number of individuals in population i	Initial size: $N_1=N_2=400$
n_L	Number of loci	300
n_a	Number of additive loci	50
A_p	Number of alleles at locus p	20
B_v	Breeding values of additive loci	[0,1]
m_{ij}	Migration rate of population i to population j	[0,0.5]
r_{pq}	Recombination rate between loci p and q	[0,0.5] [‡]
b_0	Maximum generated offspring	$P_1 = P_2 = 6$
b_1	Phenotypic optima	$P_1 = 0.25$; $P_2 = 0.75$
b_2	Variance of the fitness curve	$P_1 = P_2 = 0.5$
b_3	Density-dependent demographic effect	$P_1=0.01$; $P_2 = 0.005$
σ_{env}	Stochastic environmental variant	0.01
σ_{dem}	Stochastic demographic variant	1

1006 [†]P1 and P2 refer to the value for population 1 and 2, respectively.

1007 [‡]It may also represent a single average value for the whole chromosome (see methods)

1006
1007
1008
1009