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

Claudio S. Quilodrán[1], Kristen Ruegg[1,2,3], Ashley T. Sendell-Price[1], Eric Anderson[4], Tim Coulson[1†], Sonya Clegg[1†]

[1]Department of Zoology. University of Oxford. Oxford OX1 3PS, UK. [2]Center for Tropical Research, Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA, USA. [3]Department of Biology, Colorado State University, Fort Collins, CO, USA. [4]Fisheries 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 genetic architecture 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, 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 accumulation of genetic, genomic and computational tools has allowed a better

understanding of the genetic basis of the speciation process, resulting in the rise of a new era of evolutionary research (Hughes 2009; Chanderbali *et al.* 2016). Patterns of divergence at the level of the genome have been characterised for an increasing number of taxa, but the extent to which observed patterns are informative about evolutionary processes is actively debated (e.g. Ellegren *et al.* 2012; Renaut *et al.* 2013; Ruegg *et al.* 2014; Burri *et al.* 2015).

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

Genomic islands of divergence - highly differentiated regions of the genome that are surrounded by regions of low differentiation - are a particularly intriguing pattern of genomic divergence (Turner, Hahn & Nuzhdin 2005; Harr 2006; Nosil, Funk & Ortiz‐Barrientos 2009). Initially, their formation was attributed to the action of divergent natural selection on particular loci, creating elevated regions containing the selected loci and other physically linked loci, surrounded by regions homogenised by gene flow (Wu 2001; Wu & Ting 2004; Nosil, Funk & Ortiz‐Barrientos 2009). Several authors consequently interpreted presence of genomic islands of divergence as a signal of divergence with gene flow (e.g. Feder *et al.* 2013), concluding that the speciation process in sympatric or parapatric conditions may be 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 architecture of the diverging genomes (e.g. variation in 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).

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 after divergent selection (Cruickshank & Hahn 2014).

There have been some previous attempts to model the dynamic of the architecture 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. 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. A quantitative and more flexible framework than previous attempts is thus required to evaluate the dynamics of genomic landscapes, and increase the utility of accumulated 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. The model simulates any number of loci and allelic polymorphisms and can be theoretically motivated or parameterised using data. A major difference between our approach and previous simulations is the treatment of the fitness function, which is compatible with many structured ecological and evolutionary models. 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 architecture 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 genetic architectures, 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 is obtained by sampling from a distribution whose expected value is the expected demography. Once mating pairs are formed, the genotype of the young is determined by merging haploid gametes produced by each parent. Genetic variation of the offspring is determined by recombination and mutation.

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

Individual based framework

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

The model proceeds in discrete time steps representing generations. It is a forward simulation that includes reproduction and migration at each time step. Density dependence regulates the population growth rate, influencing the probability 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:

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

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:

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

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 *0* 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 genetic architecture 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 (θ) 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 vary depending on position of the chromosome, for example, through the use of a randomly distributed Poisson process determining crossover points. 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 (θ) between each pair of neighbouring loci. The probability of having a crossover (1) or not (0) is uniformly distributed at a rate defined by the value of θ between loci (i.e. positions with a probability smaller than θ recombine). The uniform distribution allows each position with the same values of θ to have an equal chance of crossover across all iterations. There is no recombination between homologous chromosomes when $\theta = 0$, both loci are completely linked (e.g. within an inversion or situated close to centromeres), while with a value of $\theta = 0.5$, 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 $\theta < 0.5$ means the loci are physically linked. In the second case, a single average recombination rate for the whole chromosome or part of the genotype of interest has to be supplied, and the crossover points are selected by following a random distribution (e.g. exponential). This last 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, $N_t^j m_{ji}$):

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

The model is implemented in *R* (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 script is available in the supporting information (appendix S1) and on GitHub (https://github.com/eriqande/gids), and is easily modifiable for further applications. 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 architecture of individuals was characterized by genotypes across 300 loci ($n_L = 300$), that were either strongly linked ($\theta = 0.0001$) or completely unlinked ($\theta = 0.5$). 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*.

Fifty additive loci were chosen to have non-zero 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. 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

313 genes that do not contribute to the phenotype. The first two categories of genes are
314 under selection, and the last two are not.
315 The number of mating pairs depends on the number of breeding females.
316 Female reproductive success was determined first, before male mates were assigned
317 to father each offspring. In this simulation we assumed random mating, although other
318 mating patterns are possible (e.g. Schindler *et al.* 2015). Offspring sex was assigned
319 randomly, with probability 0.5 (Table 1)
320
321 *Genomic divergence*
322 We measured pairwise F_{ST} at each locus to estimate genetic differentiation between
323 populations. *F*st is a widely used measure of heterogeneity across divergent genomes
324 in studies of genomic islands of divergence (e.g. Ellegren *et al.* 2012; Kusakabe *et al.*
325 2017). We computed F_{ST} at each simulated locus using the *R* package "*pegas*"
326 (Paradis 2010). Genetic differentiation averaged across multiple loci was calculated
327 using the approach of Nei (1973), as implemented in the *R* package "*mmod*" (Winter
328 2012).
329
330 *Simulations*
331 We conducted a number of simulation experiments using a wide suite of parameter
332 values. These were designed to examine how various scenarios of linkage between
333 loci, drift, selection, and time since divergence influence the formation of genomic
334 islands of divergence in both the presence and absence of gene flow. Parameter values
335 are presented in Table 1 and the supplementary information provides more details for
336 the choice of each parameter set (Table S1). The first simulations characterise the
337 effect of the founder population on the resulting genetic divergence (*F*st) at an early
338 stage of independent evolution (100 generations, $m_{ij} = 0$). We then explored in more
339 detail, the effect of linkage, gene flow and time since initial divergence. Drift is
340 included in all simulations through the group of genes that are not involved with the
341 phenotype trait value, and through the random selection of gametes at birth. We
342 assigned a name to each group of simulations and will briefly describe their structure.
343 1. Random sampling of founders and concordant selection: these simulations
344 were designed to examine how random sampling of the founder population influenced
345 genomic divergence. Both populations were exposed to neutral evolution and
346 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.

4. Genomic linkage: Having characterised how initial conditions might influence results, we next examined the effect of 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 ($\theta = 0.0001$) to no linked loci ($\theta = 0.5$).

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. 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 (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 grey 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 (grey 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 black line in Fig. 2a represents a single founder population with a typical, heterogeneous genomic landscape that has formed over 100 generations. There are areas of higher or lower genomic divergence between the two populations, that appear seemingly randomly across the whole genome. 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.

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 (black line, Fig 2b). The range of variance 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 grey and black line, Fig. 2b). The same single founder population illustrated in Fig 2a and 2c (black lines) provides an example of where genomic islands form at some linked loci experiencing divergent selection. A single high island of genomic divergence did not emerge at unselected loci. Due to the large variation between simulations, divergent selection did not necessarily generate islands of genomic divergence at loci under selection (grey lines, Fig. 2c).

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. This result reveals that the appearance of a pattern of genomic islands at early stages of differentiation can be caused by the genetic variation at specific loci in the founder populations.

Genomic linkage

We ran simulations with the same founder population and parameter values used to generate the black line in Fig. 2c that resulted in an island of genomic divergence, except now we varied the recombination rate (θ) among linked selected genes. The average *F*st of those linked genes was much higher with nearly complete levels of linkage ($\theta < 0.02$), 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 genomic linkage therefore tends to

447 increase the *F*st value of genes under selection. The combined effect of divergent
448 selection and linkage is consequently important for the development of genomic
449 islands of larger sizes.
450
451 *Strong selection at a single, unlinked locus:*
452 The previous simulations revealed that divergent selection on linked selected loci
453 could sometime result in islands of genomic divergence. We therefore next considered
454 a founder population in which an island of genomic divergence formed (Fig. 2c), yet
455 altered the genotype-phenotype map such that one independent locus ($\theta = 0.5$)
456 contributed disproportionately to the phenotypic value. This locus resulted in a level
457 of *F*st of more than twice that observed elsewhere in the genome, including on linked
458 selected genes (Fig. 2f). This result reveals that patterns of genomic divergence are
459 not necessarily determined by strongly linked genes of similar effect, but can also
460 emerge when one gene of large effect is linked to other markers.
461
462 *Time since divergence with and without gene flow*
463 We extended our 50 previous simulations of "random sampling of founders and
464 divergent selection" by running them for longer (100 to 2,000 generations). Without
465 gene flow, the trend of higher genomic differentiation in selected loci, particularly in
466 genes that are linked, is more evident at early stages of divergence (100 generations,
467 Fig. 3a). The length of time that independent evolution has to act influences genome-
468 wide divergence, masking signals of genomic islands that arise from single or linked
469 loci. The pattern of heterogeneous genomic differentiation is therefore less evident
470 and tends to disappear as the numbers of generations since divergence without gene
471 flow increases (2,000 generations, Fig. 3a).
472 In our simulations, populations differentiate with time even in the absence of
473 divergent selection, when both populations have equal phenotypic optima under
474 concordant selection (Fig 2b). This is because there are many ways to generate the
475 same additive phenotypic trait value. The time since initial divergence increases the
476 likelihood of generating these different outcomes (Fig 3a), therefore with enough
477 generations of isolated reproduction, populations can still be highly differentiated
478 even when they are exposed to the same fitness peak.
479
480 *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 linkage ($\theta < 0.04$) 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, and time since divergence are summarised in Figure 4. The modelling approach used has provided 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, a long-held goal of genomic data analysis (Turner, Hahn & Nuzhdin 2005; Nosil 2008; Feder *et al.* 2013; Seehausen *et al.* 2014; Nosil *et al.* 2017).

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. Recent studies exploring the distribution of genomic islands have also advanced the idea that islands arise from neutral processes without a major contribution from divergent selection (Campagna *et al.* 2015; Wang *et al.* 2016) and the results of our model identify the scenarios where this is particularly likely to be the case. Some studies document a small number of very prominent islands (e.g. Turner, Hahn & Nuzhdin 2005; Wang *et al.* 2016), however finding multiple islands of low relief is also common (e.g. Ellegren *et al.* 2012; Ruegg *et al.* 2014; Soria-Carrasco *et al.* 2014; Feulner *et al.* 2015). Furthermore, comparisons often involve recently diverged populations (e.g. Nadeau *et al.* 2012; Via 2012; Ruegg *et al.* 2014), with some divergence timescales as short as 100 generations (e.g. Marques *et al.* 2016). Our modelling suggests that these patterns and types of comparisons could be explained without recourse to explanations that invoke selection.

Another scenario that may be particularly prone to stochastic effects is where one of the diverging populations experiences a geographic expansion. Klopfstein,

549 Currat and Excoffier (2006) suggest that the effect of drift is stronger in expanding
550 populations because of "allelic surfing", where alleles that happen to be at the
551 expansion front may incidentally increase in frequency (see Hofer, Foll & Excoffier
552 2012; Excoffier, Quilodrán & Currat 2014). This, in turn, impacts genetic
553 composition, and if occurring very early during the divergence process, the
554 combination of early differences in genetic composition and drift could generate
555 highly stochastic patterns of islands of divergence.
556 We have shown that populations diverge through time even under the equal
557 selective pressures of concordant selection. Indeed, highly polygenic traits may also
558 express divergence based on which alleles of the genes under selection in the founder
559 population end up increasing in frequency. Selection will tend to create shorter
560 coalescence times around those selected loci, meaning a lower effective population
561 size and hence greater drift (Nordborg 1997). However, it should be noted that the
562 specification of the additive genotype-phenotype map we use means there are
563 multiple genotypes that will produce the same phenotypic value. This explains why
564 populations with identical selection regimes can diverge, with some developing
565 islands of genomic divergence, and others not. The nature of our genotype-phenotype
566 map in the model could also underpin the influence of initial genetic composition on
567 our results. Future work will explore whether the same conclusions hold with
568 genotype-phenotype maps that do not assume small additive contributions to the
569 phenotype from genotypes at multiple loci. However, the genotype-phenotype map
570 we use is widely assumed in quantitative genetics, and given that many traits are
571 highly polygenic, is an appropriate initial map to assume in simulations.
572
573 *Linkage and divergent selection generate islands of divergence independent of gene*
574 *flow*
575 Extreme linkage in combination with divergent selection was necessary, though not
576 sufficient, for the development of the most prominent genomic islands, regardless of
577 whether gene flow occurred or not. These findings are consistent with observations of
578 prominent genomic islands between populations presumed to be under strong
579 divergent selection, and not connected by gene flow (Burri *et al.* 2015; Zhang *et al.*
580 2017). The occurrence of candidate genes, hypothesised to be under natural selection,
581 associated with genomic islands of divergence also supports the role of selection
582 (Sousa & Hey 2013; Kusakabe *et al.* 2017). However, empirical results also provide

583 examples where SNPs under selection are not associated with islands of divergence
584 (e.g. Ruegg *et al.* 2014; Han *et al.* 2017; Riesch *et al.* 2017).
585 The importance of linkage in the appearance of genomic islands has been
586 highlighted in both theoretical and empirical studies (Feder & Nosil 2010; Renaut *et*
587 *al.* 2013; Flaxman *et al.* 2014), with extreme linkage, such as that found near
588 centromeres or within genomic inversions, often associated with the most prominent
589 genomic islands of divergence (Feder & Nosil 2009; Ellegren *et al.* 2012; Kawakami
590 *et al.* 2014). Selection acting in these zones of low rates of recombination (i.e. linked
591 selection) reduces the effective population size of these genomic regions to a greater
592 degree than in the rest of the genomes, generating genomic islands (Feder & Nosil
593 2009; Turner & Hahn 2010).
594 We did not explore the effect of linkage on deleterious variants (i.e.
595 background selection) in our simulations. However, previous studies have shown that
596 selection on both adaptive and deleterious mutations has a similar effect of reducing
597 within population diversity (Nordborg, Charlesworth & Charlesworth 1996; Slatkin &
598 Wiehe 1998), and influencing the formation of genomic islands (Cruickshank & Hahn
599 2014).
600
601 *The effect of gene flow on generation and persistence of genomic islands*
602 The idea that genomic islands of divergence were generated primarily by antagonistic
603 effects of divergent selection and gene flow was a favoured explanation until recently
604 (Turner, Hahn & Nuzhdin 2005; Nosil 2008; Feder, Egan & Nosil 2012). According
605 to this mechanism, genomic islands form around selected loci involved with the
606 divergence process, and genes physically linked to them, while adjacent neutral or
607 weakly selected regions are homogenised by gene flow (Turner & Hahn 2010;
608 Flaxman, Feder & Nosil 2013; Kawakami *et al.* 2014). Our modelling provides
609 further support that the presence of gene flow is not an essential condition, however,
610 an additional insight is that when gene flow does occur, it can lengthen the time that
611 genomic islands are visible. Verbal models of changing genomic landscapes over time
612 predicted that genomic islands would disappear with the accumulation of genome-
613 wide divergence over time (Wu & Ting 2004; Nosil 2012; Nosil & Feder 2012).
614 Empirical support that this is indeed the case is provided from studies where genomic
615 islands are more frequently documented in recently diverged versus distantly related
616 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.

Limitations

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). For non-model organisms, applications may adapt the parameter values from sister species for which information is available. 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).

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. 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

651 exercises that examine the processes that have the potential to generate such a pattern.
652 The methods described here provide a modelling framework which helps to depict
653 such signatures of past evolution, as well as potential routes of future evolution for
654 any divergent taxa.
655
656
657 **Literature Cited**
658
659 Akerman, A. & Bürger, R. (2014) The consequences of gene flow for local adaptation and
660 differentiation: a two-locus two-deme model. *Journal of mathematical biology*, **68,** 1135-
661 1198.
662 Bradshaw, A.D. (1965) Evolutionary significance of phenotypic plasticity in plants. *Advances in*
663 *genetics*, **13,** 115-155.
664 Burri, R., Nater, A., Kawakami, T., Mugal, C.F., Olason, P.I., Smeds, L., Suh, A., Dutoit, L., Bureš, S.
665 & Garamszegi, L.Z. (2015) Linked selection and recombination rate variation drive the
666 evolution of the genomic landscape of differentiation across the speciation continuum of
667 Ficedula flycatchers. *Genome research*, **25,** 1656-1665.
668 Campagna, L., Gronau, I., Silveira, L.F., Siepel, A. & Lovette, I.J. (2015) Distinguishing noise from
669 signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular*
670 *Ecology*, **24,** 4238-4251.
671 Chanderbali, A.S., Berger, B.A., Howarth, D.G., Soltis, P.S. & Soltis, D.E. (2016) Evolving ideas on
672 the origin and evolution of flowers: new perspectives in the genomic era. *Genetics*, **202,** 1255-
673 1265.
674 Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997) The effects of local selection, balanced
675 polymorphism and background selection on equilibrium patterns of genetic diversity in
676 subdivided populations. *Genetical research*, **70,** 155-174.
677 Coulson, T., Kendall, B.E., Barthold, J., Plard, F., Schindler, S., Ozgul, A. & Gaillard, J.-M. (2017)
678 Modeling adaptive and nonadaptive responses of populations to environmental change. *The*
679 *American Naturalist*, **190,** 313-336.
680 Cruickshank, T.E. & Hahn, M.W. (2014) Reanalysis suggests that genomic islands of speciation are
681 due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23,** 3133-3157.
682 Darwin, C. & Wallace, A. (1858) On the tendency of species to form varieties; and on the perpetuation
683 of varieties and species by natural means of selection. *Zoological Journal of the Linnean*
684 *Society*, **3,** 45-62.
685 Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J. & Bates, D.
686 (2011) Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40,** 1-18.
687 Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends in*
688 *Ecology & Evolution*, **29,** 51-63.
689 Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen,
690 H., Nadachowska-Brzyska, K. & Qvarnström, A. (2012) The genomic landscape of species
691 divergence in Ficedula flycatchers. *Nature*, **491,** 756-760.
692 Ellner, S.P., Childs, D.Z. & Rees, M. (2016) *Data-driven modelling of structured populations*.
693 Springer.
694 Excoffier, L., Quilodrán, C.S. & Currat, M. (2014) Models of hybridization during range expansions
695 and their application to recent human evolution. *Cultural Developments in the Eurasian*
696 *Paleolithic and the Origin of Anatomically Modern Humans* (eds A. Derevianko & M.
697 Shunkov), pp. 122-137. Department of the Institute of Archaeology and Ethnography SB
698 RAS, Novosibirsk, Russia.
699 Feder, J.L., Egan, S.P. & Nosil, P. (2012) The genomics of speciation-with-gene-flow. *Trends in*
700 *Genetics*, **28,** 342-350.
701 Feder, J.L., Flaxman, S.M., Egan, S.P., Comeault, A.A. & Nosil, P. (2013) Geographic mode of
702 speciation and genomic divergence. *Annual Review of Ecology, Evolution, and Systematics*,
703 **44,** 73-97.
704 Feder, J.L., Gejji, R., Yeaman, S. & Nosil, P. (2012) Establishment of new mutations under divergence
705 and genome hitchhiking. *Phil. Trans. R. Soc. B*, **367,** 461-474.

706 Feder, J.L. & Nosil, P. (2009) Chromosomal inversions and species differences: when are genes
707 affecting adaptive divergence and reproductive isolation expected to reside within inversions?
708 *Evolution*, **63,** 3061-3075.
709 Feder, J.L. & Nosil, P. (2010) The efficacy of divergence hitchhiking in generating genomic islands
710 during ecological speciation. *Evolution*, **64,** 1729-1747.
711 Feulner, P.G., Chain, F.J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., Lenz, T.L., Samonte,
712 I.E., Stoll, M. & Bornberg-Bauer, E. (2015) Genomics of divergence along a continuum of
713 parapatric population differentiation. *PLoS genetics*, **11,** e1004966.
714 Flaxman, S.M., Feder, J.L. & Nosil, P. (2013) Genetic hitchhiking and the dynamic buildup of genomic
715 divergence during speciation with gene flow. *Evolution*, **67,** 2577-2591.
716 Flaxman, S.M., Wacholder, A.C., Feder, J.L. & Nosil, P. (2014) Theoretical models of the influence of
717 genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23,** 4074-4088.
718 Fraïsse, C., Roux, C., Welch, J.J. & Bierne, N. (2014) Gene-flow in a mosaic hybrid zone: is local
719 introgression adaptive? *Genetics*, **197,** 939-951.
720 Han, F., Lamichhaney, S., Grant, B.R., Grant, P.R., Andersson, L. & Webster, M.T. (2017) Gene flow,
721 ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence
722 among Darwin's finches. *Genome research*.
723 Harr, B. (2006) Genomic islands of differentiation between house mouse subspecies. *Genome*
724 *research*, **16,** 730-737.
725 Hofer, T., Foll, M. & Excoffier, L. (2012) Evolutionary forces shaping genomic islands of population
726 differentiation in humans. *BMC genomics*, **13,** 107.
727 Hughes, A.L. (2009) Evolution in the post-genome era. *Perspectives in biology and medicine*, **52,** 332-
728 337.
729 Jones, O.R., Clutton‐Brock, T., Coulson, T. & Godfray, H.C.J. (2008) A web resource for the UK's
730 long‐term individual‐based time‐series (LITS) data. *Journal of Animal Ecology*, **77,** 612-
731 615.
732 Kawakami, T., Backström, N., Burri, R., Husby, A., Olason, P., Rice, A.M., Ålund, M., Qvarnström,
733 A. & Ellegren, H. (2014) Estimation of linkage disequilibrium and interspecific gene flow in
734 Ficedula flycatchers by a newly developed 50k single‐nucleotide polymorphism array.
735 *Molecular ecology resources*, **14,** 1248-1260.
736 Klopfstein, S., Currat, M. & Excoffier, L. (2006) The fate of mutations surfing on the wave of a range
737 expansion. *Molecular Biology and Evolution*, **23,** 482-490.
738 Kokko, H., Chaturvedi, A., Croll, D., Fischer, M.C., Guillaume, F., Karrenberg, S., Kerr, B.,
739 Rolshausen, G. & Stapley, J. (2017) Can Evolution Supply What Ecology Demands? *Trends*
740 *in Ecology & Evolution*.
741 Kusakabe, M., Ishikawa, A., Ravinet, M., Yoshida, K., Makino, T., Toyoda, A., Fujiyama, A. &
742 Kitano, J. (2017) Genetic basis for variation in salinity tolerance between stickleback
743 ecotypes. *Molecular ecology*, **26,** 304-319.
744 Mackay, T.F. (2014) Epistasis and quantitative traits: using model organisms to study gene–gene
745 interactions. *Nature Reviews Genetics*, **15,** 22.
746 Marques, D.A., Lucek, K., Meier, J.I., Mwaiko, S., Wagner, C.E., Excoffier, L. & Seehausen, O.
747 (2016) Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS*
748 *Genet*, **12,** e1005887.
749 Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A.,
750 Joron, M., Blaxter, M.L. & Mallet, J. (2012) Genomic islands of divergence in hybridizing
751 Heliconius butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B*,
752 **367,** 343-353.
753 Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National*
754 *Academy of Sciences*, **70,** 3321-3323.
755 Noor, M.A. & Bennett, S.M. (2009) Islands of speciation or mirages in the desert? Examining the role
756 of restricted recombination in maintaining species. *Heredity*, **103,** 439.
757 Nordborg, M. (1997) Structured coalescent processes on different time scales. *Genetics*, **146,** 1501-
758 1514.
759 Nordborg, M., Charlesworth, B. & Charlesworth, D. (1996) The effect of recombination on
760 background selection. *Genetics Research*, **67,** 159-174.
761 Nosil, P. (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17,** 2103-2106.
762 Nosil, P. (2012) Ecological speciation: Oxford series in ecology and evolution. Oxford University
763 Press Oxford:.
764 Nosil, P. & Feder, J.L. (2012) Genomic divergence during speciation: causes and consequences. The
765 Royal Society.

766 Nosil, P., Feder, J.L., Flaxman, S.M. & Gompert, Z. (2017) Tipping points in the dynamics of
767 speciation. *Nature ecology & evolution*, **1,** 0001.
768 Nosil, P., Funk, D.J. & Ortiz‐Barrientos, D. (2009) Divergent selection and heterogeneous genomic
769 divergence. *Molecular Ecology*, **18,** 375-402.
770 Paradis, E. (2010) pegas: an R package for population genetics with an integrated–modular approach.
771 *Bioinformatics*, **26,** 419-420.
772 R Development Core Team (2017) R: A Language and Environment for Statistical Computing. R
773 Foundation for Statistical Computing, Vienna, Austria.
774 Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlović, M., Noor, M., Mehlig, B. &
775 Westram, A. (2017) Interpreting the genomic landscape of speciation: a road map for finding
776 barriers to gene flow. *Journal of Evolutionary Biology*, **30,** 1450-1477.
777 Renaut, S., Grassa, C., Yeaman, S., Moyers, B., Lai, Z., Kane, N., Bowers, J., Burke, J. & Rieseberg,
778 L. (2013) Genomic islands of divergence are not affected by geography of speciation in
779 sunflowers. *Nature Communications*, **4,** 1827.
780 Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A.A., Farkas, T.E., Lucek, K.,
781 Hellen, E., Soria-Carrasco, V. & Dennis, S.R. (2017) Transitions between phases of genomic
782 differentiation during stick-insect speciation. *Nature ecology & evolution*, **1,** 0082.
783 Ruegg, K., Anderson, E.C., Boone, J., Pouls, J. & Smith, T.B. (2014) A role for migration‐linked
784 genes and genomic islands in divergence of a songbird. *Molecular Ecology*, **23,** 4757-4769.
785 Schindler, S., Gaillard, J.M., Grüning, A., Neuhaus, P., Traill, L.W., Tuljapurkar, S. & Coulson, T.
786 (2015) Sex‐specific demography and generalization of the Trivers–Willard theory. *Nature*,
787 **526,** 249.
788 Sedghifar, A., Brandvain, Y. & Ralph, P. (2016) Beyond clines: lineages and haplotype blocks in
789 hybrid zones. *Molecular ecology*, **25,** 2559-2576.
790 Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L.,
791 Saetre, G.-P., Bank, C. & Brännström, Å. (2014) Genomics and the origin of species. *Nature*
792 *Reviews. Genetics*, **15,** 176.
793 Slatkin, M. & Wiehe, T. (1998) Genetic hitch-hiking in a subdivided population. *Genetics Research*,
794 **71,** 155-160.
795 Soria-Carrasco, V., Gompert, Z., Comeault, A.A., Farkas, T.E., Parchman, T.L., Johnston, J.S.,
796 Buerkle, C.A., Feder, J.L., Bast, J. & Schwander, T. (2014) Stick insect genomes reveal
797 natural selection's role in parallel speciation. *Science*, **344,** 738-742.
798 Sousa, V. & Hey, J. (2013) Understanding the origin of species with genome-scale data: modelling
799 gene flow. *Nature Reviews Genetics*, **14,** 404-414.
800 Turner, T.L. & Hahn, M.W. (2010) Genomic islands of speciation or genomic islands and speciation?
801 *Molecular Ecology*, **19,** 848-850.
802 Turner, T.L., Hahn, M.W. & Nuzhdin, S.V. (2005) Genomic islands of speciation in Anopheles
803 gambiae. *Plos Biology*, **3,** e285.
804 Via, S. (2012) Divergence hitchhiking and the spread of genomic isolation during ecological
805 speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B:*
806 *Biological Sciences*, **367,** 451-460.
807 Wang, L., Wan, Z.Y., Lim, H.S. & Yue, G.H. (2016) Genetic variability, local selection and
808 demographic history: genomic evidence of evolving towards allopatric speciation in Asian
809 seabass. *Molecular Ecology*, **25,** 3605-3621.
810 Winter, D.J. (2012) MMOD: an R library for the calculation of population differentiation statistics.
811 *Molecular Ecology Resources*, **12,** 1158-1160.
812 Wu, C.-I. & Ting, C.-T. (2004) Genes and speciation. *Nature Reviews Genetics*, **5,** 114.
813 Wu, C.I. (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14,** 851-
814 865.
815 Zhang, D., Song, G., Gao, B., Cheng, Y., Qu, Y., Wu, S., Shao, S., Wu, Y., Alström, P. & Lei, F.
816 (2017) Genomic differentiation and patterns of gene flow between two long‐tailed tit species
817 (Aegithalos). *Molecular Ecology*.
818
819

820 **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]
θ_{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

821 [†]P1and P2 refer to the value for population 1 and 2, respectively.

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

823

824

825

826

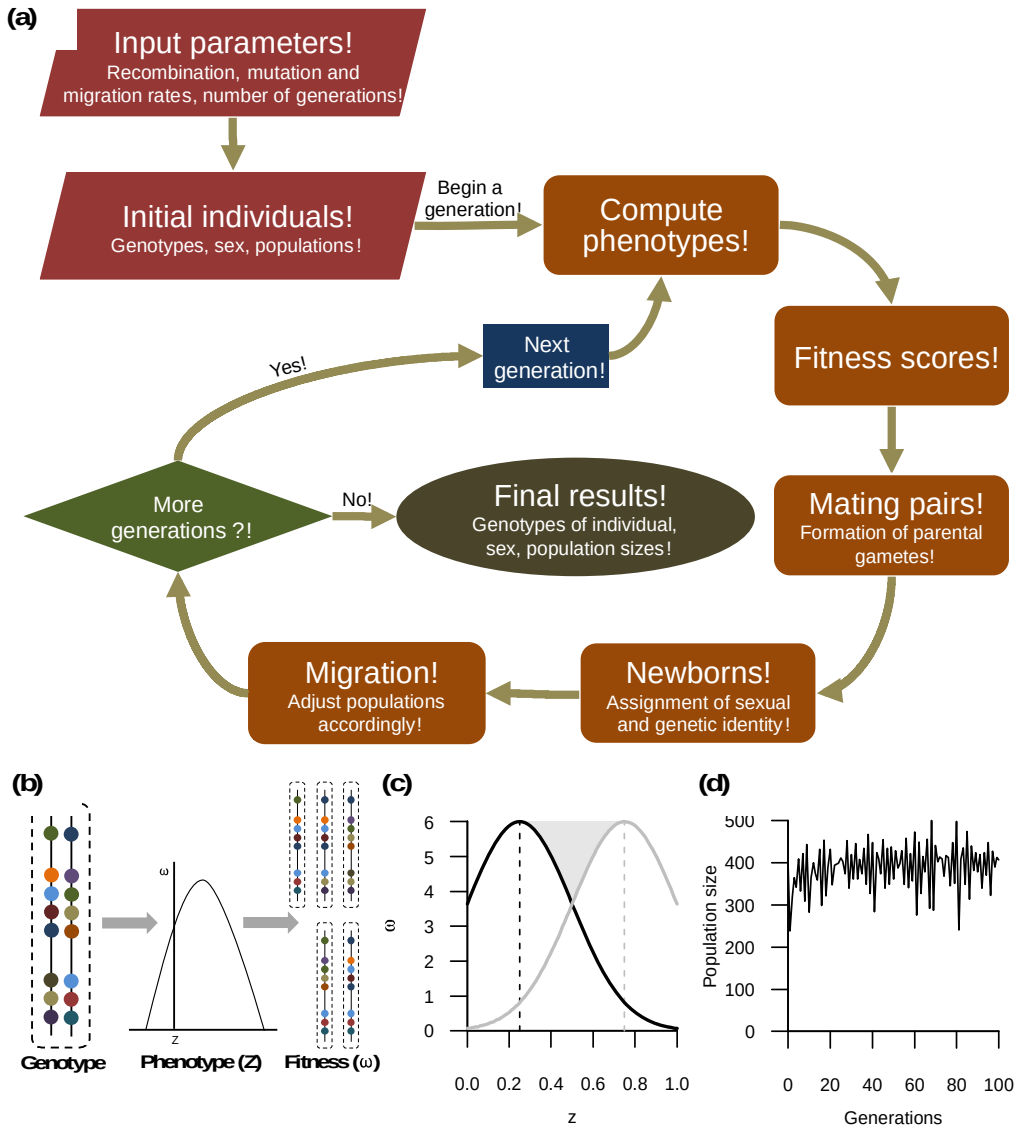

Fig. 1. General description of the simulation framework. A) Main steps of the general modelling approach. The red polygons represent the starting conditions. The orange squares are the different computing steps on each generation. The green polygon is a condition variable stating either the running of a next generation (blue square) or the end of simulations (olive green circle). B) Relationship between genotypes, phenotypes, and fitness. The genetic variation is represented in different colours. The space between points represents unequal centiMorgan distances. C) Two different fitness functions with different phenotypic optima. D) Example of population size across time.

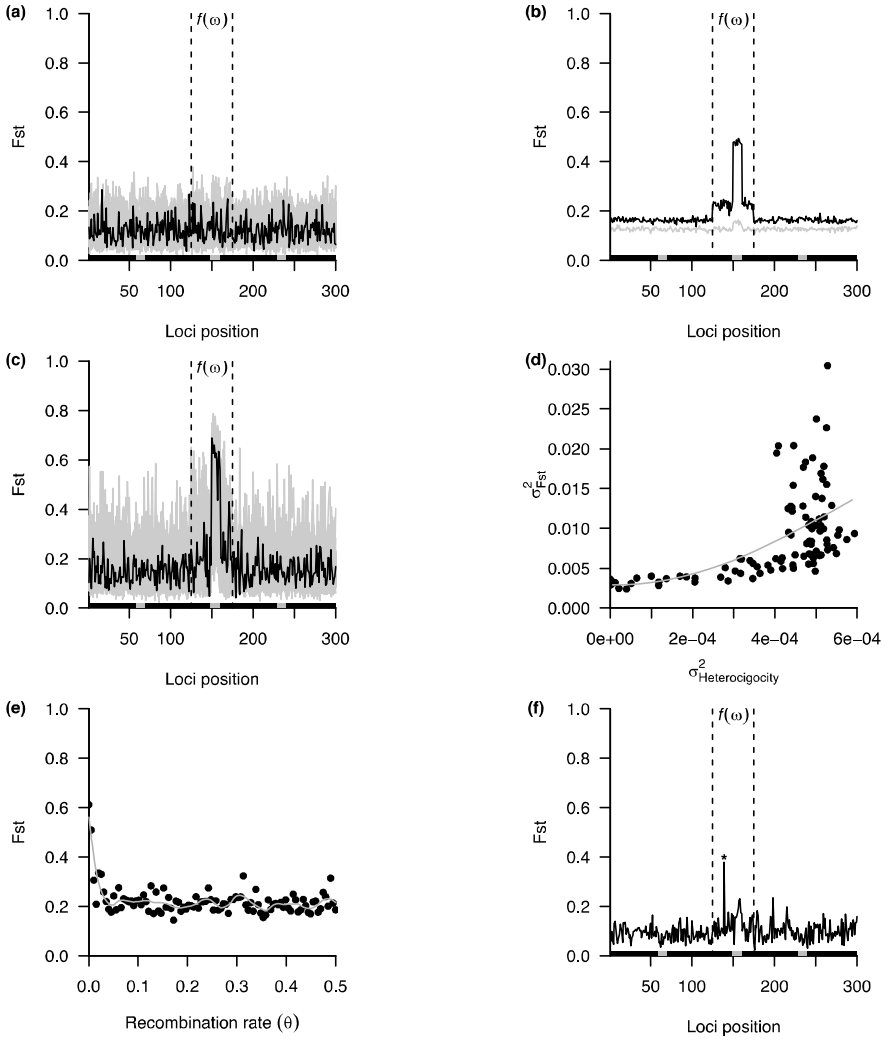

Fig. 2. Simulations of genetic differentiation between two simulated populations without gene flow. a) Random sampling of founders and concordant selection. The grey lines represent the resulting genetic differentiation (*F*st) on 50 comparisons with different random sampling of founders. The black line illustrates the resulting values for a single founder population. The grey squares on the horizontal axis represent linked loci (θ = 0.0001) and the black rectangles independent ones (θ = 0.5). The dotted vertical lines delimit loci participating in the computation of phenotypes. b) Mean divergence by loci on the 50 founder populations. The black and the grey lines represent divergent selection and concordant selection, respectively. c) Random sampling of founders and divergent selection. The grey lines and the black line represent equal founder populations as in figure 2a, but adding divergent selection to the analysis. d) Levels of heterozygosity in the founder population. Influence of the heterozygosity variance at the beginning of the simulations on the variance of *F*st at the end of the simulations. e) Genomic linkage. Effect of the strength of linkage on the formation of a genomic island. *F*st values are averaged over the 10 linked loci influencing the computation of phenotypes and using the same starting conditions as the black line in Fig 2c. f) Strong selection at a single, unlinked locus. A single independent locus with a stronger additive effect on the computation of phenotypes (*). All data are presented after 100 generations of independent evolution.

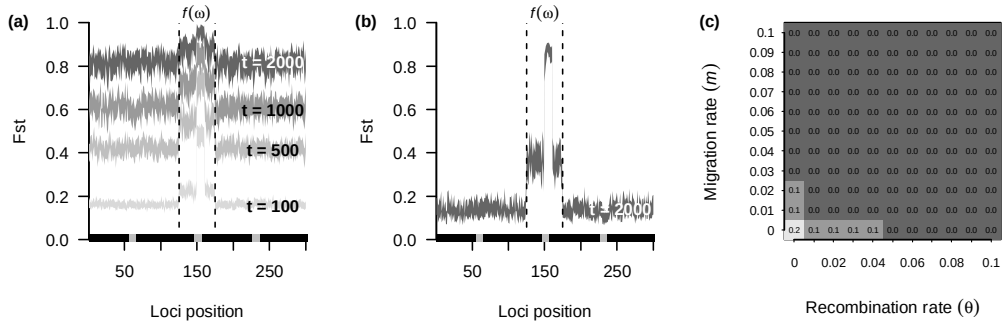

Fig. 3. Time since divergence and gene flow. a) Divergence without gene flow. The grey squares on the horizontal axis represent linked loci (θ = 0.0001) and the black rectangles unlinked ones (θ = 0.5). The dotted vertical lines delimit the loci participating in the computation of phenotypes. The coloured areas represent a confidence interval at 95% of *F*st values, estimated over 50 simulations with randomly assigned genetic identity of individuals at the beginning of the divergence (t = generations). b) Divergence with gene flow. This is similar to the previous figure, but allows for gene flow between populations (m = 0.01). c) Linkage and gene flow. Combined effect of migration rate and recombination rate on the magnitude of a genomic island. The numbers inside the squares represent the difference between mean *F*st estimated at the 10 linked loci influencing the computation of phenotypes (i.e. genomic island, positions 150 to 159) and 10 loci not related to fitness and independent (i.e. genomic background, positions 90 to 99). These numbers represent the average difference over the same starting conditions used to estimate the confidence interval of Fig 3a. The data in this last figure are presented after 100 generations of divergent evolution.

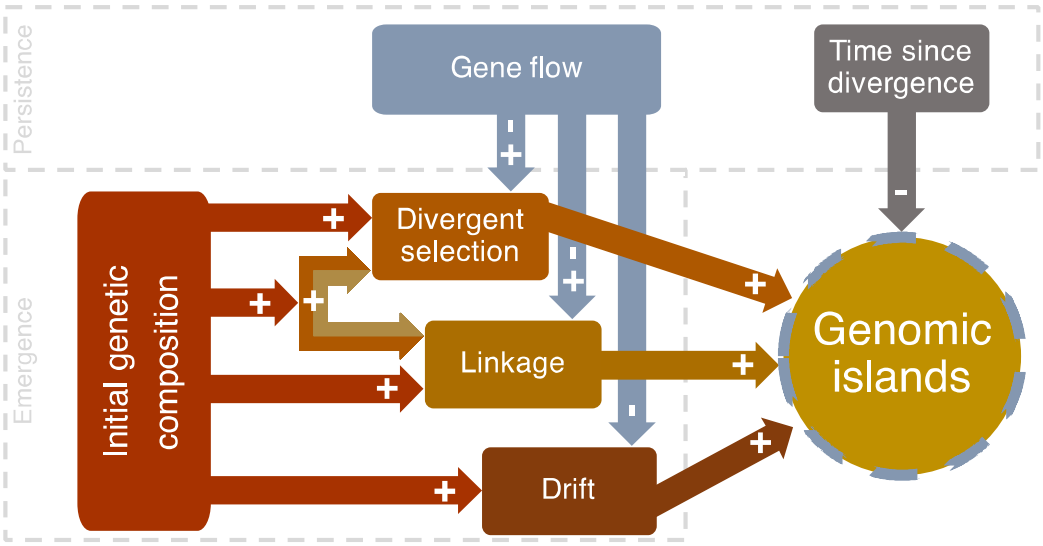

Fig 4. Factors influencing the patterns of genomic islands of divergence. Genomic islands may emerge under the influence of linkage, divergent selection, an interaction between these two factors, or drift depending upon the initial genetic composition of the starting populations (positive effects). Gene flow and time since divergence have an effect on the persistence of islands once formed. Gene flow has an indirect effect by interacting with factors influencing the emergence of this pattern. At early stages of divergence, gene flow can lengthen the time that genomic islands are visible (positive effect), but too high a level of gene flow can erase genomic island patterns (negative effect). The time since divergence has a negative effect on genomic islands, which are more visible under earlier rather than later stages of genomic differentiation.