The evolution of genomic islands by increased establishment probability of linked alleles

Sam Yeaman [*†‡•], Simon Aeschbacher[§•], and Reinhard Bürger [§§]

[†]Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada
[‡]Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada [§]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA [§§]Faculty of Mathematics, University of Vienna, Oskar–Morgenstern–Platz 1, A–1090 Vienna, Austria

*samuel.yeaman@ucalgary.ca
•*These authors contributed equally*

Key words: gene flow, local adaptation, linkage disequilibrium, recombination, genetic architecture, divergence hitchhiking

Abstract

Genomic islands are clusters of loci with elevated divergence that are commonly found in population genomic studies of local adaptation and speciation. One explanation for their evolution is that linkage between selected alleles confers a benefit, which increases the establishment probability of new mutations that are linked to existing locally adapted polymorphisms. Previous theory suggested there is only limited potential for the evolution of islands via this mechanism, but involved some simplifying assumptions that may limit the accuracy of this inference. Here we extend previous analytical approaches to study the effect of linkage on the establishment probability of new mutations, and identify parameter regimes that are most likely to lead to evolution of islands via this mechanism. We show how the interplay between migration and selection affects the establishment probability of linked vs. unlinked alleles, the expected maximum size of genomic islands, and the expected time required for their evolution. Our results agree with previous studies, suggesting that this mechanism alone is unlikely to be a general explanation for the evolution of genomic islands. However, this mechanism could occur more readily if there were other pre-adaptations to reduce local rates of recombination or increase the local density of mutational targets within the region of the island. We also show that island formation via erosion following secondary contact is much more rapid than island formation from *de novo* mutations, suggesting that this mechanism may be more likely.

Introduction

Studies of natural populations commonly find that genetic divergence is elevated in a number of genomically restricted regions along the chromosomes. These have been dubbed “genomic islands” of speciation (Turner *et al.* 2005), differentiation (Harr 2006), or divergence (Nosil *et al.* 2009), depending on the biological context and the way they are measured. Yet, other studies have not found such islands, although they might have been expected in their respective contexts (Nosil *et al.* 2009; Strasburg *et al.* 2012; Renaut *et al.* 2013). What can be learned about the evolutionary history of a species by studying these genomic islands, and what does their absence tell us? There are many different ways that the interplay between selection and demography can give rise to this characteristic pattern in the genome. Therefore, identifying ways to discount alternative explanations and evaluate the significance of presence, absence, and extent of genomic islands is critical to making inferences about evolutionary history.

Mechanistic explanations for the evolution of genomic islands can be categorized by whether clusters arise as: 1) purely the result of drift and demography, 2) neutral byproducts of linkage and natural selection with or without on-going gene flow, 3) a direct adaptive response, due to the advantage of physical linkage among loci experiencing divergent natural selection and gene flow (summarised in Table 1), or through some combination of these explanations. These three broad categories reflect differences in the number of selected loci in the island that affect the trait under divergent selection, with either none, one, or several causal loci per island, respectively. We emphasise that explanation 2 can lead to situations where multiple selected loci, each with their own cluster of neutral sites, are located in close proximity along the genome by random chance. This would lead to an empirical pattern very difficult to distinguish from explanation 3 in practice. For conceptual clarity, however, we categorize these mechanisms in terms of the number of causal loci and the type of linkage-mediated benefit among them.

Before reviewing these three mechanisms, it is important to situate our approach in the context of the many ways that genetic divergence can be quantified and used to identify genomic islands. The most widely used measures of divergence are F_{ST} (Wright 1931, 1943) and related statistics (Nei 1973; 1982; Slatkin 1991; Charlesworth 1998; Excoffier 2007), which express the variance in allele frequency among (partially) isolated populations, normalized by the genetic diversity in an appropriately defined panmictic population. As such, F_{ST}-like statistics are relative measures of divergence. They are sensitive to a variety of processes affecting diversity within and among populations, and therefore tend to be easily confounded. In contrast, absolute measures of divergence such as Dxy (Nei & Li 1979) are more specific, but not very sensitive with certain types of data. Here, we formulate our discussion of islands of divergence in terms of F_{ST}. Where necessary, we mention potential confounding factors, and in Table 1 we contrast F_{ST} to patterns including absolute measures that are diagnostic for various mechanisms. For a review of relative and absolute measures and their properties, we refer to Cruickshank & Hahn (2014).

Although the action of natural selection through either mechanism 2 or 3 may feel intuitively most parsimonious, it is important to discount neutral explanations in empirical studies (i.e. mechanism 1, Table 1). If a pattern of genomic islands is found along a single gradient or between a single pair of populations, it may be difficult to conclusively rule out the effect of demography. The distribution of coalescence times for neutral alleles becomes much broader under isolation-by-distance or hierarchical structure than under panmixia, so loci that appear to be outliers under an island model may in fact still be consistent with drift (Excoffier *et al.* 2009; Hermisson 2009). Linkage among loci in close proximity on a chromosome could then cause the appearance of an island of extreme F_{ST}. Additionally, rapid population expansion can result in 'allelic surfing' due to increased genetic drift at the wave front (e.g. Slatkin & Excoffier 2012). This could likely cause blocks of elevated divergence, especially in regions of low recombination, but has not been extensively studied. However, if the same genomic islands are found along several demographically independent gradients, it is much less likely that the pattern could have arisen purely by drift and demography.

Within the second category of explanation, there are several different ways that linkage of neutral loci to an allele under selection can result in elevated divergence along a chromosome. Recurrent bouts of either purifying or positive and spatially homogenous selection can result in reduced genetic variation within populations at regions linked to the loci under selection (Maynard Smith & Haigh 1974; Charlesworth *et al.* 1993), which can inflate relative metrics of divergence, such as F_{ST} (sub-category 2A in Table 1; Nei 1973; Charlesworth *et al.* 1997; Nordborg 1997; Cruickshank & Hahn 2014). It is unclear whether alternative, absolute metrics of divergence that are insensitive to variation within populations, such as D_{xy}, are sensitive enough to address this problem, but if genomic islands do not co-occur with locally reduced heterozygosity, then the explanations of purifying selection and global sweeps can be discounted. Alternatively, spatially heterogeneous selection can lead to increased divergence at linked neutral regions. If this occurs as a single selective sweep within a restricted region of the species range (without gene flow), then hitchhiking can decrease heterozygosity and increase divergence at linked sites, but this signature will decay over time (sub-category 2B in Table 1; Maynard Smith and Haigh 1974; Gillespie 2000). When divergent selection operates in the face of sufficiently low gene flow, such local sweeps are never completed and an equilibrium is reached where the change in allele frequency due to selection is opposed by gene flow. In populations of finite size, this recurrent selection against maladaptive gene flow generates a persistent pattern of increased divergence and reduced heterozygosity at linked sites (sub-category 2C in Table 1; Charlesworth *et al.* 1997; Nordborg 1997; Via & West 2008; Feder & Nosil 2010). Finally, if locally adapted and previously allopatric populations come into secondary contact, erosion of divergence will occur most rapidly at neutral loci that are less tightly linked to a given adapted locus (sub-category 2D in Table 1; Petry 1983; Barton and Bengtsson 1986; Barton and Hewitt 1989).

By the third category of explanation, genomic islands evolve as an adaptive consequence of the way recombination mediates the tension between divergent selection and gene flow. If two or more locally beneficial mutations are tightly physically linked, recombination is less likely to break up this favourable combination than when linkage is weak (Fisher 1930; Charlesworth 1979; Lenormand & Otto 2000). This advantage due to linkage can lead to clustered genetic architectures under divergent selection in the face of gene flow, through one of four potential evolutionary routes (category 3 in Table 1): A) increased probability of establishment of linked alleles during initial divergence (Yeaman & Whitlock 2011; Feder *et al.* 2012; Aeschbacher & Bürger 2014); B) competition among genetic architectures as new linked mutations displace older unlinked mutations of equal overall effect (Yeaman and Whitlock 2011); C) competition among genomic architectures, which favours the fixation of rearrangements moving loci into close proximity on a chromosome (Yeaman 2013) or the establishment of segregating inversions that reduce recombination (Noor *et al.* 2001; Rieseberg 2001; Kirkpatrick 2006); D) increased persistence time for linked and selected loci under secondary contact following local adaptation in allopatry (Petry 1983; Barton & Bengtsson 1986; Barton & Hewitt 1989). In all of the above cases in category 3, selection at the causal loci would also cause elevated divergence at linked neutral loci.

One way to differentiate between explanations of category 2 and 3 is to identify whether, in a single genomic island, there are multiple sites harbouring alleles that contribute to fitness differences between populations. This is expected to occur under the latter but not the former explanation. While it is very difficult to identify which mutations are functionally important, it may be possible to infer their presence from observations about multiple peaks within an island, or through inferences about the maximum expected width of an island under explanations of category 2 vs. 3. Furthermore, we could learn much about a species' evolutionary history by discriminating between the four possible explanations within category 3. In practice, this is difficult because some of the expected statistical signatures from these routes to clustering may be very similar at equilibrium. For instance, at equilibrium, both scenarios 3.A) and 3.D) lead to an expected pattern of F_{ST} peaks surrounded by regions of low F_{ST}. That said, the trajectories to this equilibrium pattern may be quite different. In 3.A), peaks of divergence could be seen as mountains arising against a more or less constant background, whereas in 3.D) regions of reduced divergence are better viewed as valleys eroding between mountains of a (roughly) fixed height. As such, it is helpful to use theoretical methods to investigate the expected properties of these various explanations and examine their relative likelihood.

Theoretical arguments have been made suggesting that the higher establishment probability (EP) of linked mutations (explanation 3.A) is unlikely to result in a strong signal of clustering, because the increase in EP is small and extends over a small region of a given chromosome, relative to the size of the unlinked portion of the genome (Yeaman 2013). However, the analysis by Yeaman (2013) discounted the effect of linkage in the regions of parameter space where it is potentially most important, so the inferences made about the likelihood of genomic island evolution via increased EP may be incorrect. Moreover, Yeaman (2013) employed a semi-heuristic approximation to the EP (Yeaman & Otto 2011) that has not been formally tested in the case of multiple linked loci under selection. Here, we first verify this approximation by comparison to an alternative, more formally derived one (Aeschbacher & Bürger 2014). Second, we extend the work of Yeaman (2013) to quantify more comprehensively the potential importance of increased EP at linked mutations as an explanation for the evolution of genomic islands. Specifically, we explore predictions about the expected width of a genomic island, the absolute and relative increase in EP due to linkage between selected sites, the waiting time for island evolution, and the effect of different distributions of mutational effect size on these predictions. Third, we determine how fast neutral divergence that has built up between two loci under spatially divergent selection without gene flow erodes upon secondary contact, and compare this to the time it takes for an equivalent two-peak island to arise by divergent selection with on-going gene flow. Overall, our aim is to refine the theoretical arguments about island evolution, provide some testable predictions about island size, and identify how further study could aim to test the other hypotheses for island evolution, particularly through secondary contact and erosion of genetic divergence around selected loci.

Model

To explore the likelihood of genomic islands evolving by increased probability of establishment of new linked mutations (explanation 3.A, Table 1), it is necessary to consider both the effect of linkage on the establishment probability (EP) and the rate of occurrence of linked vs. unlinked mutations. To study the effect of linkage on the EP of a *de novo* mutation that occurs near an already established locally adapted allele, we describe a model that incorporates the relevant evolutionary forces (*i.e*. selection, gene flow, recombination, and genetic drift), but is still simple enough to allow for efficient mathematical analysis. We then use simplifying assumptions about the expected distribution of fitness effects of new mutations and the relative size of the mutational target in linked vs. unlinked regions to parameterize our model and make inferences about the likelihood of genomic island evolution via increased EP.

We consider a discrete-time, monoecious diploid model with continent–island type migration (Haldane 1930; Wright 1931) and divergent selection at two linked biallelic loci A and B. Let the alleles at these two loci be A_1 and A_2, and B_1 and B_2, respectively. We define the migration rate m as the proportion of the island population replaced by immigrants from the continent each generation, and denote by r the probability of recombination between the two loci per generation. Assuming additive interactions across alleles and loci, and ignoring epistasis as well as parental and position effects in heterozygotes, we write the fitnesses of the nine distinguishable genotypes in the island population as

$$\begin{array}{c c} & \begin{array}{ccc} B_1B_1 & B_1B_2 & B_2B_2 \end{array} \\ \begin{array}{c} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{array} & \begin{pmatrix} 1+a+b & 1+a & 1+a-b \\ 1+b & 1 & 1-b \\ 1-a+b & 1-a & 1-a-b \end{pmatrix} \end{array}, \qquad (1)$$

where a and b are the selective advantages on the island of alleles A_1 and B_1 relative to A_2 and B_2, respectively. To enforce positive fitnesses, we require that $0 < a, b < 1$, and $a + b < 1$. Unless otherwise stated, we assume that the continental population is fixed for alleles A_2 and B_2, whereas alleles A_1 and B_1 are local *de novo* mutations endemic to the island population. The conditions for establishment and maintenance of one- and two-locus polymorphisms in the island population have been described before (Bürger & Akerman 2011; Aeschbacher & Bürger 2014). For mathematical convenience, we assume that selection in favour of A_1 is weaker than selection in favour of B_1 ($a < b$). This assumption implies that allele A_1 cannot be maintained in the island population if allele B_1 is swamped by gene flow (Haldane 1930; Bürger & Akerman 2011). It is further in agreement with our principal focus on the effect of linkage on establishment of a weak, locally beneficial *de novo* mutation arising in the presence of an already existing migration–selection polymorphism at a background locus. Throughout, we assume that populations are large enough such that genetic drift can be ignored after an initial stochastic phase during which *de novo* mutations arise in low absolute numbers. In this setting, locally beneficial *de novo* mutations have a strictly positive establishment probability if they can invade in an equivalent fully deterministic model, and *vice versa* (Bürger & Akerman 2011; Aeschbacher & Bürger 2014). Therefore, if in the following we say that "A_1 can be established", this means both that A_1 has a strictly positive establishment probability and that it will invade in a corresponding deterministic model.

Aeschbacher & Bürger (2014) derived several results for this model that will help interpret our findings discussed below. We therefore briefly recapitulate them in the following. First, in a single-locus model (no background locus B), A_1 can be established if and only if $m < a$ (cf. Haldane 1930). Second, if $m \geq a$, the additional effect of linkage to a background locus B means that A_1 can be established if $m < m^*$, where $m^* = \frac{a(b-a+r)}{(a-r)(a-b)+r(1-a)}$ (their Eq. 10). Third, this critical threshold can alternatively be expressed in terms of a critical recombination rate, where linkage must be at least as tight as $r^* = \frac{a(a-b)(1+m)}{a(1+2m)-(1+b)m}$ when $m > a / (1 - 2a + b)$ (their Eq. 11). Fourth, if $m > b/(1 - a)$, then no polymorphism will be maintained at locus B, which automatically implies that A_1 cannot be established because gene flow is too strong.

Based on this general model, we derive three approximations for the expected EP of any given new mutation, $\pi = \pi(a, r)$, expressed as a function of a and r, which are discussed in more detail below and in the supplementary materials. To make predictions about the average establishment probability over all available linked mutations, $\overline{\pi}_L$, we integrate $\pi(a, r)$ across the range of possible recombinational distances r at which new mutations could be linked to the background locus, and across the expected Distribution of Fitness Effects (DFE) for values of a. We assume that the DFE is a gamma distribution,

$$f_a(k, \overline{a}) = \frac{\left(\frac{k}{\overline{a}}\right)^k}{\Gamma(k)} a^{k-1} e^{-\left(\frac{k}{\overline{a}}\right)a}, \tag{2}$$

with shape parameter k and mean $\overline{a}$, where $k, \overline{a} > 0$. If $k = 1$, this simplifies to an exponential distribution, $f_a(1, \overline{a}) = \frac{1}{\overline{a}} e^{-a/\overline{a}}$. With a drawn according to equation (2), our assumption of $a < b$ does not necessarily hold; in practice, we therefore set $\overline{a} \ll b$. When integrating over r, we assume a uniform distribution of r between 0 and ½, usually using an upper limit of $r_f = ½$, because this represents the threshold to free recombination in a discrete-time model. This approach to integration is biologically akin to assuming that mutations occur with equal probability at all positions within the window of the chromosome where $0 < r < r_f$, and the rate of recombination does not vary along the chromosome. For mutations with $r = 1/2$ on the same chromosome, they are essentially unlinked and are not assumed to contribute to $\overline{\pi}_L$. To represent the average EP of an unlinked mutation, $\overline{\pi}_U$, we multiply $\pi(a, r)$ by the DFE, and integrate over a, setting $r = r_f$

= ½. To obtain the approximate size of a genomic island evolving via the benefit of selection at a linked background locus (mechanism 3.A, Table 1), we determine the region of the chromosome over which most new mutations will experience an increased EP via this mechanism. Specifically, we calculate the value of r required to contain 95% of the probability density of $\pi(a, r)$, and refer to this as the 95% window size, or C_{95}.

To explore the time scales over which the last step in the evolution of two-peak genomic islands occur under alternative scenarios (explanations 3.A vs. 3.D), we use a combination of stochastic and deterministic theory. We extend our model to bidirectional gene flow and predict the dynamics of neutral divergence using the structured coalescent. Thereby, we replace the neutral migration rate by appropriate rates of effective gene flow, which are a function of the location of the neutral site and the strength of selection at the two loci causing the adaptive peaks (see SI for details).

Results

Analytical predictions for establishment probability

Two-type branching process approximation in discrete time: The EP under a two-type branching process, $\pi(a, r)$, was derived in Aeschbacher & Bürger (2014), and represents a weighted average over A_1 occurring either on the locally beneficial (B_1) or deleterious (B_2) genetic background (see SI for details). The EP, $\pi(a, r)$, decreases both with the migration rate m and the recombination r, and increases with the selection coefficient a (solid lines in Figure 1). As expected, if migration is sufficiently weak, then A_1 can establish even if unlinked to the background locus B (solid blue line in Figure 1B).

To represent the average probability of a new, linked mutation across all possible recombination rates and sizes of selection coefficient, we integrate across both the rate of recombination and the DFE,

$$\bar{\pi}_L = \int_0^\infty \int_0^{r_f} \pi(a,r) f_a(k,\bar{a}) f_r(0,r_f)\, dr\, da, \qquad (3)$$

where $f_r(r_{\min}, r_{\max})$ is a uniform density between $r_{\min}$ and $r_{\max}$. Other distributions could be used to represent non-homogenous recombination rates or mutation targets, but for simplicity we focus on the uniform case here. We could not find a simple closed form for $\bar{\pi}_L$, so in the following analyses we use numerical integration to investigate how it varies with $\bar{a}$, *b*, *r*, *m*, and *k*. The difference between the EP of an average linked vs. an average unlinked mutation of a certain effect size *a*, i.e. $\bar{\pi}_L^{(r)}(a)$ vs. $\bar{\pi}_U^{(r)}(a)$, increases with *m*, and above the critical threshold *m**, the EP for an individual unlinked mutation is 0 (Figure 2A). Integrating across the DFE, and hence accounting for all possible values of *a*, shows that the mean EP of unlinked mutations, $\bar{\pi}_U$, decreases more gradually and does not rapidly intersect 0, as it includes the effects of rare large mutations that have non-zero EP (Figure 2B).

Slightly-supercritical branching process: To derive an approximation to the branching-process solution described above, we assume that the selective advantage of the focal mutation and the migration rate are both small relative to the selection coefficient at the background locus and the recombination rate, *i.e*. that $a, m \ll b, r$. Under these assumptions, it can be shown that the EP of a linked mutation is

$$\pi(a,r) \approx 2\frac{a(b+r)-(1+b)mr}{(1+b)(b+r)} \qquad (4)$$

if $m < a(b + r) / [r\,(1 + b)]$, and 0 otherwise (see SI for derivation). By this approximation, mutations can establish if the recombination rate r is below a critical value given by

$$\tilde{r}^{-} = \frac{ab}{(1+b)m-a}. \qquad (5)$$

If we further assume that both selection coefficients are much smaller than unity, *i.e.* $0 < a \ll b \ll 1$, then this reduces to $\tilde{r}^{-} = ab/m$, which is identical to the approximate critical recombination rate for invasion of A_1 in a deterministic continuous-time model (see Eq. 4.15 in Bürger & Akerman, 2011). This can be further reduced to $\tilde{r}^{-} \approx b$ when $m \sim a$, which is Barton's rule of thumb that selection against gene flow has an appreciable effect on linked sites only if linkage is sufficiently strong, i.e. $r \ll b$ (Barton 2000).

Under the assumption of an exponential DFE ($k = 1$), a closed-form solution for the average EP with respect to r and a can be derived by using a Taylor series approximation around $b = 0$ (see SI for details). This yields

$$\bar{\pi}_L \approx e^{-\frac{m}{a}}\left\{\bar{a} + \bar{a}b + bm - 2bm\gamma + 2bm\left[E_i\left(\frac{m}{a}\right) + \ln\left(\frac{\bar{a}}{2bm}\right)\right]\right\} - 2\bar{a}b, \qquad (6)$$

where $\gamma = 0.577$ is the Euler–Mascheroni constant and $E_i(z) = -\int_{-z}^{\infty}\frac{e^{-t}}{t}\,dt$ is the exponential integral function. This approximation works well as long as $b \lesssim 0.2$ (see SI and Figure S1 for details), and is very close to the splicing approximation (see below).

Both of these approximations overestimate the exact two-type branching process when either m is large or r is small (Figure 1).

Splicing approximation in discrete time: The rate of increase in frequency of allele A_1 when rare is given approximately by the leading eigenvalue of the Jacobian matrix describing the stability around the equilibrium where B is polymorphic and A is fixed for A_2 (see SI for details; as per Yeaman and Otto 2011). Because this deterministic rate of increase in frequency is equivalent to the effect of natural selection in a single-locus one-population model, this rate can be spliced into the standard probability of fixation for a new mutation (Pr[*fix*] = 2*s*; Haldane 1927), which yields an approximation to the EP of A_1,

$$\pi(a,r) \approx 2\max\left[0, \frac{2+b-r+m(2a-b-r)+\sqrt{R}}{2(1-a+b)} - 1\right], \qquad (7)$$

where

$$R = (1+m)\{b^2(1+m) + 2b(1-m)r + r[r - m(4-4a-r)]\}. \quad (8)$$

As in the case of the two-type branching process, we must resort to numerical integration to obtain approximations to the average probability of establishment of linked alleles for a given selection coefficient a, $\bar{\pi}_L^{(r)}(a)$, and across the entire DFE, $\bar{\pi}_L$. As shown in Figure 1, the splicing approach (dotted lines) yields an approximation to the EP very close to the one obtained from the slightly-supercritical branching process assuming $a, m \ll b, r$ (dashed lines; Eq. 4). Accordingly, it deviates most from the exact two-type branching process (solid lines; Eq. S3) if a is relatively large and, at the same time, either *m* is large or *r* is small. In these cases, the contribution of mutations that occur on the deleterious genetic background B_2 becomes important; both the splicing approach and the slightly-supercritical branching process do not capture this.

Island size and establishment probability of linked vs. unlinked mutations

We now use the above approximations to explore the likelihood of genomic islands evolving by mechanism 3.A (Table 1). As outlined above, we integrate both across all possible recombination rates and mutation effect sizes (assuming a gamma-distributed DFE with a mean effect size of $\overline{a}$, as above) to represent how an average linked vs. unlinked mutation would be affected by the interplay between migration, selection, and recombination. With increasing migration rate in the region of $m \sim \overline{a}$, an increasing fraction of the available mutations become more strongly limited by migration than favoured by selection. This causes a decrease in the mean EP of both linked ($\overline{\pi}_L$) and unlinked ($\overline{\pi}_U$) mutations, but the effect of linkage to the B locus mitigates this effect, so that the mean EP decreases faster for unlinked mutations. These broad patterns are illustrated in Figures 3A & B, which show the effect of *m* on $\overline{\pi}_L$, and Figure 3C, which shows the ratio $\overline{\pi}_L/\overline{\pi}_U$. In each case, there is a pronounced transition in the region of $m \sim \overline{a}$. The decrease in $\overline{\pi}_U$ occurs more rapidly than $\overline{\pi}_L$ because as the migration rate increases, π_U approaches zero for an increasing fraction of unlinked mutations, whereas linked mutations still have non-zero probability of establishment (cf. Figure 2B). Therefore, $\overline{\pi}_L/\overline{\pi}_U$ increases towards infinity if *m* is increased far beyond $\overline{a}$, and while linked mutations can still establish, their absolute EP is much reduced.

Another parallel effect of increasing migration is to reduce the size of the linked region that can contribute to adaptation. As migration increases, an increasing fraction of the chromosome falls above the critical recombination threshold *r**, such that an increasing fraction of the available linked mutations cannot be established. As a consequence, there is a pronounced decrease in the expected size of the region around the background locus that can contribute to local adaptation if $m > \bar{a}$, as shown by C_{95}, the 95% window size (Figure 3D).

The shape of the gamma distribution (*k*) affects the position of the transition zone where i) the EP starts decreasing, ii) the ratio $\bar{\pi}_L/\bar{\pi}_U$ starts increasing, and iii) the size of the 95% window starts decreasing. These transitions occur at lower values of *m* for *k* = 2 (*i.e.* a DFE with relatively more intermediate values of *a*, compared to *k* = 1) and higher values of *m* for *k* = 0.5 (i.e. a DFE with relatively more extreme values of *a*, compared to *k* = 1) (Figure S2). Thus, while *k* does not affect the broad qualitative patterns representing the role of linkage, it does affect the relative importance of linkage at a given value of *m* (*i.e.* the broad patterns are maintained, but shifted with respect to *m* with changes in *k*). Specifically, the average EP of linked mutations, $\bar{\pi}_L$, is higher for smaller *k* because a larger fraction of the probability density occurs with mutations of larger size, which have proportionally larger EP $\bar{\pi}_L^{(r)}$ (Figure S2B). This is also reflected in the width of the genomic window within which most linked mutations establish: smaller *k* and hence relatively more mutations of large effect increase the size of the window (C_{95} in Figure S2D). However, put in relation to the average EP of unlinked mutations, $\bar{\pi}_U$, the effect of the shape is compensated for by the fact that unlinked mutations profit proportionally more from a fat right-hand tail (low *k*) of the input DFE (Figure S2C).

It is also worth noting that the EP is lower for higher values of b when m is very low (see SI for details and a derivation of the critical migration rate). This occurs because the contribution to the total fitness made by locus A becomes relatively smaller as b increases (*i.e.*, the marginal fitness ratio of genotype $A_1A_2B{\bullet}B{\bullet}$ vs. $A_2A_2B{\bullet}B{\bullet}$ decreases with increasing b), which reduces the net advantage of allele A_1, and hence its EP. However, this effect of the relative magnitude of selection coefficients is only important when the facilitating effect of linkage is relatively small (a large relative to b); when linkage is important, then the EP increases with b.

Waiting time for evolution of genomic islands

Taken together, the above results show that linkage becomes critically important for any local adaptation to occur when the migration rate increases beyond the mean selection coefficient ($m > \bar{a}$). However, they also illustrate that in this same region of parameter space, there is a decrease in both the size of the region that can contribute to this adaptation (Figure 3D) as well as the absolute EP of mutations within this region (Figure 3A &B). These latter two factors would be expected to greatly increase the waiting time for the first linked mutation to establish, as the mutational target is smaller with small window sizes, and waiting time scales with the inverse of the EP. To obtain a rough approximation to this expected waiting time, we combine the results from Figure 3, calculating $t_L \approx 1/(2N\mu 2C_{95}\bar{\pi}_L)$ and $t_{\mathrm{U}} \approx 1/(2N\mu C_U\bar{\pi}_U)$, for linked and unlinked mutations, where N is the size of the diploid population, μ is the mutation rate per cM per gamete per generation, C_{95} is the 95% window size in cM as defined above, and C_{U}

is the size of the genome that is unlinked to B, in cM. We multiply C_{95} by 2 to account for the fact that a linked mutation can occur on both sides of the background locus. Unfortunately, it is difficult to parameterize these equations to make quantitative predictions about the expected waiting time, because we know little about the rate of beneficial mutations. In organisms such as humans, where genome size is on the order of 10^9 base pairs (bp), and where there are approximately 10^6 bp/cM and about 10^{-8} mutations/bp/generation, the mutation rate per cM is ~ 0.01 per generation, but it is unclear what fraction of these would be locally beneficial.

Given uncertainty about N and μ, and accounting for appropriate mutational target, we scale the waiting times by the inverse of the substitution rate of an average mutation from the same DFE that arises in a completely isolated panmictic population of size N. Because establishment is independent of recombinational distance from the background locus if $m = 0$, this reference mutational target simply amounts to $2C_{95,m=0} = 2 \times 0.95 \times r_f = 0.95$ for $r_f = 0.5$, independently of all the other parameters. Specifically, we approximate the scaled expected waiting times for linked and unlinked mutations as

$$T_L \approx t_L \left(\frac{1}{4N\mu\bar{a}\,2C_{95,m=0}} \right)^{-1} \approx \frac{4N\mu\bar{a}\,2C_{95,m=0}}{4\ N\mu\, C_{95}\bar{\pi}_L} = \frac{0.95\ \bar{a}}{C_{95}\bar{\pi}_L}, \tag{9a}$$

$$T_U \approx t_U \left(\frac{1}{4N\mu\bar{a}\,2C_{95,m=0}} \right)^{-1} \approx \frac{4N\mu\bar{a}\,2C_{95,m=0}}{2\ N\mu\, C_U\bar{\pi}_L} = \frac{1.9\ \bar{a}}{C_U\bar{\pi}_U}. \tag{9b}$$

This scaling has the advantage of focussing exclusively on the *relative* impact on waiting times of the tension between migration and divergent selection, and therefore seems appropriate for a comparison of linked vs. unlinked locally beneficial *de novo* mutations.

Some uncertainty remains about C_U, the size of the unlinked mutational target. However, we can make reasonable choices. For $b = 0.1$ and a C_U of 100 cM, i.e. equal to twice the map distance with respect to which we defined C_{95}, the difference in waiting times for linked and unlinked mutations is negligible if $m < \bar{a}$ (dotted vs. solid curves in Figure 4A). This is expected, because in the limit of no migration, the background locus is fixed for B_1, and establishment of the focal mutation A_1 does not depend on its recombinational distance from locus B. As *m* increases above $\bar{a}$, T_U increases more rapidly than T_L, because linkage now increases the relative establishment probability and the mutational target for linked mutations (C_{95}) is still large (compare Figures 3C and 3D). At $m \approx 10\bar{a}$, the increase in T_L becomes log-log linear, while T_U keeps increasing exponentially with m. This is where linkage starts to convey a considerable relative advantage. However, absolute waiting times for linked mutations are at least about 100 times higher compared to the case of $m = 0$ (Figure 4A). This log-log linear phase ends if $m \approx b$, at which point gene flow swamps even the beneficial background allele B_1, and T_L quickly increases to infinity.

Realistically, the mutational target size for unlinked mutations is much larger than 100 cM in most organisms. As expected, for $C_U > 100$ cM, the waiting time to establishment for unlinked mutations becomes much reduced compared to linked mutations if m is weak enough, suggesting that in this region of parameter space, a strong signature of clustering would be unlikely to evolve. The larger the unlinked target size, the higher the migration rate at which linked mutations are expected to arise earlier than unlinked ones (i.e. where dashed curves intersect with solid ones in Figure 4A). When this occurs, t_L tends to be very long (relative to the waiting time in the absence of migration) except at the highest values of $\overline{a}$ (blue curves in Figure 1A). This illustrates that in the region where a strong signal of clustering is expected via mechanism 3.A, the waiting time for the evolution of this pattern will be very long relative to adaptation in allopatry. The shape of the gamma distribution (*k*) has a strong effect on absolute waiting times, but does not substantially affect relative difference between linked and unlinked locally beneficial *de novo* mutations (Figure S3). However, there is some effect of *b*, as the transition to long waiting times with increasing *m* for linked mutations is sharper and occurs at lower *m* with small *b* (Figure 4B).

Divergence with gene flow vs. secondary contact

Genomic islands consisting of multiple causal loci can evolve under various mechanisms summarised in category 3 (Table 1). The same amount of divergence is expected at equilibrium in models of secondary contact and *de novo* adaptation (Bürger & Akerman 2011; Aeschbacher and Bürger 2014), which makes it challenging to distinguish between these alternative explanations based on empirical observations. However, by studying the dynamics leading to these equilibrium patterns, it may be possible to exclude some mechanisms based on a comparison of associated time scales with the demographic history of populations and species of interest. To illustrate this argument, we concentrate on the competing mechanisms of increased establishment probability of linked mutations in the face of gene flow (explanation 3.A in Table 1) vs. the erosion of neutral divergence upon secondary contact (explanation 3.D). Moreover, we extend our model to allow for symmetric gene flow between the two demes of equal size, as compared to the asymmetric continent–island regime assumed above. We restrict our treatment to genomic islands made up of two causal loci, and we focus on the last evolutionary step needed to complete equivalent two-locus islands at the equilibrium of migration, selection, and genetic drift. By equivalent we mean that the map distance between the two loci A and B under selection, the migration rate *m* at parapatric stages, and the selection coefficients *a* and *b* are identical in the two scenarios.

Under explanation 3.A, we define the last evolutionary step as the rise of a peak of divergence at locus A by establishment of a weak, locally beneficial mutation in one of the two demes. Here, we condition on there being an established migration–selection polymorphism at the linked background locus B, and on the new mutation being successful. We start counting time when the locally beneficial mutation occurs. As above, we focus on the case where linkage to B is essential for establishment of the new mutation. In contrast, under explanation 3.D, we view the erosion of neutral divergence as the last evolutionary step during secondary contact following adaptive divergence at loci A and B in allopatry (assuming the period of allopatry has been long enough to allow local adaptation, which is not included in the following estimates). Here, we condition on locally adaptive mutations having been fixed before secondary contact, and we start counting time at the onset of secondary contact.

In both cases, we ignore mutation at loci A and B, and we assume that neutral mutation rates are low, so that the infinite-alleles model of mutation holds and F_{ST} can be approximated as

$$F_{\mathrm{ST}} \approx \frac{T_T - T_W}{T_T} \qquad (10)$$

(Slatkin 1991, Wilkinson-Herbots 1998) Here, T_T and T_W are the expected coalescence times between two samples taken at random from the entire population and from within a deme, respectively. These times depend on the local effective population size N_e and the migration rate m in a neutral model. We derive these coalescence times in the SI using classical results for the structured coalescent (Griffiths 1981; Takahata 1988; Notahara 1990; Wilkinson-Herbots 1998, 2008, 2012). To account for the effect of selection, we replace m by the appropriate effective rate of gene flow m_e (Petry 1983; Aeschbacher & Bürger 2014; Akerman & Bürger 2014).

Under these assumptions, erosion of neutral divergence during secondary contact after adaptive divergence occurs on a much faster time scale than the rise of a new peak in F_{ST} under *de novo* selection with gene flow (Figure 5). With $a = 0.01$, $b = 0.1$, and $m = 0.02$, a potentially identifiable two-peak island is formed already after about $t = 100$ generations in the secondary-contact scenario (pale-blue curves in Figure 5A), whereas the peak arising at locus A is barely noticeable at this time; only after about $t = 1000$ generations is there a clear second peak (Figure 5B). Under purely deterministic dynamics, i.e. assuming infinitely large local population sizes, the waiting time for the decay of neutral divergence in terms of F_{ST} is well approximated by the inverse of the effective rate of gene flow, $1/m_e$ (SI; Figure S4). Even with a finite local population size N_e, $1/m_e$ seems to provide a good approximation to the time until F_{ST} reaches its equilibrium in the secondary-contact case (Figure 5C). In contrast, it takes much more time for F_{ST} at locus A to rise upon occurrence of a locally beneficial mutation in the presence of gene flow (Figure 5D and Figure S5B and S5D).

The large differences in time scales over which equivalent genomic islands arise and become potentially detectable suggest that explanation 3.D may be much more likely than 3.A if the two populations or species of interest have diverged relatively recently. Given that the above analysis assumes that a mutation destined to establish has already occurred under 3.A, the total waiting time for mechanism 3.A would be even longer than shown in Figure 5 (as per Figure 4). These differences remain even if adaptive divergence before secondary contact has not been complete, as long as the allopatric phase has not been too short (Figure S5A and S5C). We note that the mutations contributing to local adaptation in allopatry (mechanism 3.D) are not expected to be more clustered than random, unless the underlying mutational target is itself clustered (e.g., due to previous evolution via mechanism 3C). Thus, mechanism 3.D also depends on multiple mutations occurring in close enough proximity to each other that a contiguous island is observed, at least for some time after secondary contact.

Discussion

Identification and interpretation of genomic islands is challenging for two interrelated reasons: combinations of evolutionary scenarios can lead to similar patterns of genomic diversity and divergence, and widely used measures of divergence differ in sensitivity and specificity to detect and differentiate between alternative scenarios. The original explanation for island evolution, by which they arise in genomic regions where on-going gene flow is effectively reduced due to divergent selection (Turner et al. 2005), has been challenged and refined to include the various scenarios presented in Table 1. Whereas empirical study will serve to resolve some of these issues, here we have used theoretical approaches to explore the importance of linkage for establishment of new mutations during an initial phase of local adaptation.

Evolution of genomic islands by increased establishment probability

There are three main factors that affect the likelihood of genomic islands evolving by increased EP of linked mutations (mechanism 3.A): the relative increase in the EP of linked vs. unlinked mutations, their absolute EP, and the size of the mutational target within the linked region that experiences increased EP. The relative increase in EP for linked vs. unlinked mutations will be greatest when there is very strong selection on a single locus (the B locus) accompanied by high migration and weaker selection on other loci (represented by the A locus here). If the distribution of fitness effects of new mutations is gamma distributed, linked mutations will have much higher EP relative to unlinked ones when the migration rate exceeds the average selection coefficient substantially ($m >> \bar{a}$; Figure 3C). However, as the migration rate increases above $\bar{a}$, the other factors lose importance: Both the absolute EP and the size of the region in which linkage increases EP decrease rapidly (Figure 3A & D). Because the size of the mutational target and the absolute EP both decrease, the waiting time for the establishment of the first linked mutation increases in this region of parameter space. This results in adaptation proceeding at a rate much slower than would occur in the absence of migration (Figure 4). Thus, there is a ‘goldilocks zone’ where the balance between migration and selection is just right for mechanism 3.A: migration is high enough that unlinked mutations have low EP, but not so high that the waiting time for a linked mutation to establish is very long. In other words, this zone is roughly between the migration rate that maximizes the ratio of establishment probabilities in Figure 3C, and the one that minimizes the time in Figure 4). In species with a continuous or stepping-stone pattern of population structure spanning a broad environmental gradient, migration rates may often be too geographically restricted to substantially constrain adaptation, further limiting the potential importance of this mechanism.

In light of the sensitivity of this mechanism to the balance between migration and selection, both demography and environment would have to be quite stable over relatively long periods of time for evolution of genomic islands exclusively by increased EP of linked mutations. In such cases, genomic islands would evolve most rapidly in organisms with large population size, as this increases the total number of mutations available for selection and therefore reduces the waiting time. Also, as the average number of crossovers per chromosomes is found to be relatively insensitive to physical chromosome length (Hillers and Villeneuve 2003), we would expect that organisms with larger genomes have higher mutation rates per centimorgan, and that mechanism 3.A might therefore be more important in such organisms. .

Two assumptions are implicit in the above discussion: the underlying mutational target for a given polygenic trait is uniformly distributed throughout the genome and the rate of recombination is homogeneous. If there are recombination coldspots around the B locus due to segregating inversions or the previous fixation of recombination modifiers, then increased EP due to linkage would extend over a greater physical distance and therefore mutational target. Similarly, if functionally related loci tend to cluster together in the genome (Nützmann & Osbourn 2014), then the rate of decay in the EP per base pair would be unchanged, but the mutational target within this region would be increased. Both of these non-homogeneous features of the genome can co-occur with the loci involved in local adaptation by chance or as a result of previous bouts of local adaptation that favoured the establishment or fixation of such modifiers or rearrangements (mechanism 3.C). In either case, this could considerably increase the potential for the evolution of genomic islands via increased probability of establishment of linked mutations, as suggested by Yeaman (2013).

It is worth noting that Figure 4 provides a very rough comparison of the situation for linked vs. unlinked mutations, as it uses the somewhat arbitrary C_{95} to determine the mutational target for linked mutations. This also assumes that the physical distance scales linearly with the rate of recombination, which does not account for the reduction in effective recombination rate when there are even numbers of crossovers between a pair of loci in a given meiosis. Also, by simply contrasting the expected establishment times and ignoring the variances, we do not provide a rigorous assessment of what would constitute a statistically significant signature of an island (even a few clustered loci with many non-clustered ones might be detectable as an island). Finally, our approach ignores the fact that there might be multiple background loci that could independently initialise genomic islands. The latter effect will depend on the DFE, and appropriate treatment of this would require modelling *b* as being drawn from that DFE, too. Comprehensively addressing these issues is beyond the scope of this paper, but provides an obvious problem to address in the future, and our conclusions throughout should be considered with these caveats in mind.

Evolution from new mutations vs. erosion and secondary contact

Whereas mechanism 3.A depends explicitly on the mutation rate, the evolution of islands via erosion following secondary contact (mechanism 3.D) is independent of mutation rate and therefore can proceed much more rapidly (assuming that populations in allopatry have had sufficient time to accumulate the differences that are eroded in secondary contact). Here, we show that island formation via secondary contact and erosion of divergence (mechanism 3.D) will be much more rapid than via establishment of new mutations, even once a mutation that is destined to establish has already occurred (Figure 5). We also show that the expected time to island formation via secondary contact and erosion scales approximately with the inverse of the effective migration rate ($1/m_e$; Figure S4) for a wide range of population sizes. While we have not considered adaptation from standing variation here, we suspect that the likelihood of island formation from standing variation is similar to, if not lower than, that for new mutations for the same reasons as we discussed above: in the parameter space where linkage is critical to establishment, the absolute EP is low, so there would need to be many mutations present as standing variation to give a signature. In addition, the fact that these mutations are segregating at some intermediate, albeit potentially low, frequency implies that they have already overcome the curse of stochastic loss. But this initial phase is exactly when linkage is expected to convey the largest relative advantage. Mutations segregating as standing variation may depend much less on this initial benefit. However, further work is necessary to rigorously explore the contrast between standing and *de novo* variation in this context, and special attention should be paid to the cause of standing variation, as migration among populations inhabiting similar environments can introduce pre-adapted variants.

The absence of genomic islands

While finding a strong signature of genomic islands can reveal much about the genomic basis of adaptation, the lack of any observable peaks in F_{ST} does not necessarily mean that no adaptation is occurring. While the two-locus models covered here predict that no local adaptation will occur when $m > m^*$, divergent adaptation at the phenotypic level can be maintained at equilibrium through a quantitative genetic response mediated by very small changes in allele frequency coupled with positive covariance among alleles, provided sufficient genetic variance is maintained by mutation (Le Corre & Kremer 2003; Yeaman 2015). In such cases, divergence at the underlying loci may be transient, and no signature of islands is expected, nor any significant peaks in F_{ST}. Alternatively, if there are multiple alleles or haplotypes present at a single locus that yield the same locally adapted phenotype, the signature of F_{ST} would likely be greatly reduced, especially at loci linked to the causal selected site(s). Studies of experimental evolution in bacteria commonly find that many different amino acid residues within a given gene evolve in response to the same selective pressure (e.g., Vogwill *et al.* 2014; Bailey *et al.* 2015). Similarly, a catalogue of the loci identified in adaptation has found that the same locus often is involved repeatedly with different alleles (Martin & Orgogozo 2013). If several such alleles at a single locus are maintained in a population, the expected signatures of divergence would be weaker. Such patterns could likely evolve much more readily with adaptation from standing variation. While we have restricted our study here to two-deme, two-allele models, further study of more realistic population structures and the possibility of haplotypes of equal fitness is necessary to more completely understand their impact on statistical signatures of adaptation. Finally, it is worth noting that even in cases where we would predict stable genomic islands to form based on our model, in empirical studies these may not be statistically distinguishable from the genomic background due to the highly stochastic nature of the genealogical process, mutation, and recombination. This would be especially problematic in cases where migration between populations is low and F_{ST} at neutral sites is high (Feder & Nosil 2010). In such cases, however, there may still be a genome-wide aggregate signal in the form of a negative correlation between divergence and local recombination rates (Nachman & Payseur 2012; Brandvain *et al.* 2014).

Comparison to results of Yeaman (2013)

The analysis presented here uses an approach similar to that of Yeaman (2013), combining approximations for the EP with representations of the size of linked vs. unlinked mutation targets, but extends this analysis and corrects a deficiency in presentation. Because Yeaman (2013) numerically integrated the EP only over the recombination rate but not the DFE, the analysis was restricted to considering the relative advantage of linkage for a single mutation. As such, when migration was high enough that the EP of an unlinked mutation became 0, the ratio of π_L/π_U (referred to there as the DHA, or 'Divergence Hitchhiking Advantage') became undefined. This was plotted as a value of 0 in Yeaman's Figure 1, which is unfortunate as it suggests that the relative advantage of linkage is low in this region of parameter space, when in fact it is infinitely high and mutations of that size could not establish without linkage. Thus, inferences based on Figure 1 from Yeaman (2013) discounted the contribution of potentially the most important region of parameter space (i.e. the shaded region shown here in Figure 2A). Here, we correct this oversight by integrating across both recombination and the DFE of possible mutations at the A locus, so that $\bar{\pi}_L$ always includes the contribution of some mutations with non-zero EP (i.e., some mutations have $m^* > m$), and the mean EP, $\bar{\pi}_L$ therefore never becomes 0 (Figure 2B). With this approach, we show how in the region where $\bar{\pi}_L/\bar{\pi}_U$ tends to infinity, the advantage of linkage is counterbalanced by reduced size of mutation target and increased time to establishment. Thus, our analysis leads to the same qualitative conclusions as Yeaman (2013): that mechanism 3.A is unlikely to be a broad explanation for the evolution of genomic islands independently of some other factors affecting the distribution of the mutational target or the local rate of recombination. These conclusions are insensitive to whether we use the splicing approach (as in Yeaman 2013) or the two-type branching process (Aeschbacher & Bürger 2014) to approximate the establishment probability. Although there are quantitative differences between these two approximations if migration is relatively strong or recombination relatively weak (Figure 1), there is good agreement between them in terms of the ratio of establishment probabilities, $\bar{\pi}_L/\bar{\pi}_U$, and the 95% window size, C_{95} (Figure S6).

Conclusions and future directions

Understanding the importance of the various mechanisms for the evolution of genomic islands will inevitably require additional theory and more empirical work. Future theoretical work could further explore the role of the DFE, study the effects of variable selection strength at the background locus, and incorporate dominance, epistasis, and other modes of selection (e.g. background selection). This would inform the development of methods for robust inference about the mechanisms underlying genomic patterns of diversity. In parallel, more comprehensive study of the extent of gene flow and hybridization in natural populations is needed, as well as complementary lines of evidence on gene function and phenotypic effect. This will inform the current debate about the importance of gene flow in generating genomic islands (Cruickshank & Hahn 2014; Sætre 2014; Feulner *et al.* 2015; Monnahan *et al.* 2015) and help answer a number of open questions: Is on-going gene flow commonly strong enough to make linkage essential for divergence, or does gene flow only modulate evolving or previously existing patterns? Are empirically observed islands produced by one or multiple sites under divergent selection? How commonly do loci contribute to reproductive isolation independent of adaptation? Do these islands reflect regions of reduced recombination as a preadaptation to clustered architectures of traits under divergent selection, or did clustered architectures and reduced local recombination rates evolve in response to such selection? The distinction between preadaptation vs. response to selection may also be seen as a question of timescale. For instance, in sticklebacks that repeatedly colonised post-glacial lakes (Rogers *et al.* 2013), species may have encountered the same type of environmental heterogeneity many times over deep evolutionary time. In such case, what is now a pre-adaptation could be a consequence of previous response to selection for reduced recombination between locally adapting loci. From the theoretical arguments outlined here, it seems unlikely that increased establishment probability due to linkage will provide a broad explanation for the evolution of genomic islands without some sort of preadaptation, either in local recombination rate or the genomic distribution of mutational targets. Comparative genomic studies may provide the best means to identify how commonly such genome-scale changes have shaped the chromosomal landscape upon which local adaptation and reproductive isolation are built.

Acknowledgements

We would like to thank Michael Whitlock, Nick Barton, Alexandre Blanckaert, Joachim Hermisson, and Ivan Juric for helpful discussion during the preparation of this manuscript. This work was funded by the Genome Canada LSARP program, with co-funding from Genome BC and the Forest Genetics Council of BC (co-Project Leaders S. N. Aitken and A. Hamann), Alberta Innovates Health Solutions (Translational Health Chair to SY), the Swiss National Science Foundation (Advanced Postdoc.Mobility fellowship P300P3_154613 to SA), and the Austrian Science Fund (FWF; grant P25188 to RB).

References

Aeschbacher S, Bürger R (2014) The effect of linkage on establishment and survival of locally beneficial mutations. *Genetics*, **197**, 317–336.

Akerman A, Bürger R (2014) *The consequences of gene flow for local adaptation and differentiation: a two-locus two-deme model*.

Bailey SF, Rodrigue N, Kassen R (2015) The Effect of Selection Environment on the Probability of Parallel Evolution. *Molecular Biology and Evolution*, **32**, 1436–1448.

Bank C, Bürger R, Hermisson J (2012) The limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-islands model. *Genetics*, **191**, 845–863.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57 (Pt 3)**, 357–376.

Barton N, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. *PLoS Genetics*, **10**, e1004410.

Bürger R, Akerman A (2011) The effects of linkage and gene flow on local adaptation: A two-locus continent-island model. *Theoretical Population Biology*, **80**, 272–288.

Charlesworth B (1979) Selection on recombination in. , 581–589.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular biology and evolution*, **15**, 538–543.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

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.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 461–474.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution, **64**, 1729–1748.

Feulner PGD, Chain FJJ, Panchal M *et al.* (2015) Genomics of divergence along a continuum of parapatric population differentiation. *PLoS genetics*, **11**, e1004966.

Fisher R (1930) *The Genetical Theory of Natural Selection*. Clarendon, Oxford.

Gillespie JH (2000) The neutral theory in an infinite population. *Gene*, **261**, 11–18.

Griffiths RC (1981) The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. *Journal of Mathematical Biology*, **12**, 251–261.

Haldane JBS (1930) A mathematical theory of natural and artificial selection. VI. Isolation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **26**, 220–230.

Harr B (2006) Genomic islands of differentiation between house mouse subspecies. , 730–737.

Hermisson J (2009) Who believes in whole-genome scans for selection? *Heredity*, **103**, 283–284.

Hillers KJ, Villeneuve AM (2003) Chromosome-wide control of meiotic crossing over in C. elegans. *Current Biology*, **13**, 1641–1647.

Kirkpatrick M (2006) Chromosome Inversions, Local Adaptation and Speciation. *Genetics*, **173**, 419–434.

Lenormand T, Otto SP (2000) The evolution of recombination in a heterogeneous environment. *Genetics*, **156**, 423–438.

Lindtke D, Buerkle CA (2015) The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution*, **69**, 1987-2004.

López E, Pradillo M, Oliver C *et al.* (2012) Looking for natural variation in chiasma frequency in Arabidopsis thaliana. *Journal of Experimental Botany*, **63**, 887–894.

Martin A, Orgogozo V (2013) The loci of repeated evolution: a catalog of genetic hotspots of phenotypic adaptation. *Evolution*, **67**, 1235-1250.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research Cambridge*, **23**, 23–35.

Monnahan PJ, Colicchio J, Kelly JK (2015) A genomic selection component analysis characterizes migration-selection balance. *Evolution*, n/a–n/a.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3321–3323.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Nachman, MW. and Payseur, BA. (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Phil. Trans. R. Soc. B* 367:409–421.

Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. , **2001**.

Nordborg M (1997) Structured coalescent processes on different time scales. *Genetics*, **146**, 1501–1514.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Notahara M (1990) The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29**, 59–75.

Nützmann H-W, Osbourn A (2014) Gene clustering in plant specialized metabolism. *Current Opinion in Biotechnology*, **26**, 91–99.

Petry D (1983) The Effect on Neutral Gene Flow of Selection at a Linked Locus. , **313**, 300–313.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution*, **16**, 351–358.

Rogers SM, Bowles E, Mee JA The consequences of genomic architecture on ecological speciation in postglacial fishes. , 1–25.

Sætre G-P (2014) Genome scans and elusive candidate genes: detecting the variation that matters for speciation. *Molecular ecology*, **23**, 4677–8.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research Cambridge*, **58**, 167–175.

Slatkin M, Excoffier L (2012) Serial founder effects during range expansion: A spatial analog of genetic drift. *Genetics*, **191**, 171–181.

Strasburg JL, Sherman NA, Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation ? What can patterns of differentiation across plant genomes tell us about adaptation and speciation ?

Takahata N (1988) The coalescent in two partially isolated diffusion populations. *Genetical Research*, **52**, 213–222.

Trans P, Lond RS, Barton NH (2000) Genetic hitchhiking Genetic hitchhiking. , 1553–1562.

Turner TL, Hahn MW, Nuzhdin S V (2005) Genomic Islands of Speciation in Anopheles gambiae. , **3**.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Vogwill T, Kojadinovic M, Furió V, MacLean RC (2014) Testing the Role of Genetic Background in Parallel Evolution Using the Comparative Experimental Evolution of Antibiotic Resistance. *Molecular biology and evolution*, **31**, 3314–3323.

Wilkinson-Herbots H (1998) Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, **37**, 535–585.

Wilkinson-Herbots HM (2008) The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theoretical Population Biology*, **73**, 277–288.

Wilkinson-Herbots HM (2012) The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theoretical Population Biology*, **82**, 92–108.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright S (1943) Isolation by distance. *Genetics*, **28**.

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1743–51.

Yeaman S (2015) (in press) Local adaptation by small-effect alleles. *Am Nat*, **186**.

Yeaman S, Otto SP (2011) Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, **65**, 2123–2129.

Yeaman S, Whitlock MC (2011) The Genetic Architecture of Adaptation under Migration-Selection Balance. *Evolution*, **65**, 1897–1911.

Figures

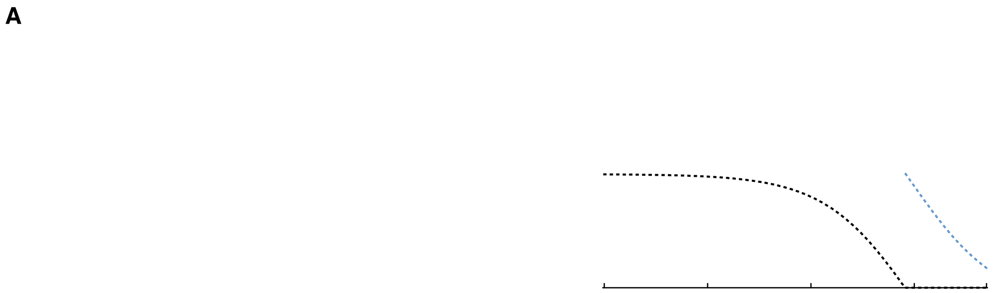

Figure 1. Comparison of three approximations to the establishment probability of a new linked mutation. The establishment probability based on the two-type branching process (solid lines), a slightly-supercritical two-type branching process assuming $a, m \ll b, r$ (dashed lines), and the splicing approach (dotted lines) as a function of migration rate (A) and recombination rate (B). A) r = 0.01; B) m = 0.01. In both cases, b = 0.1

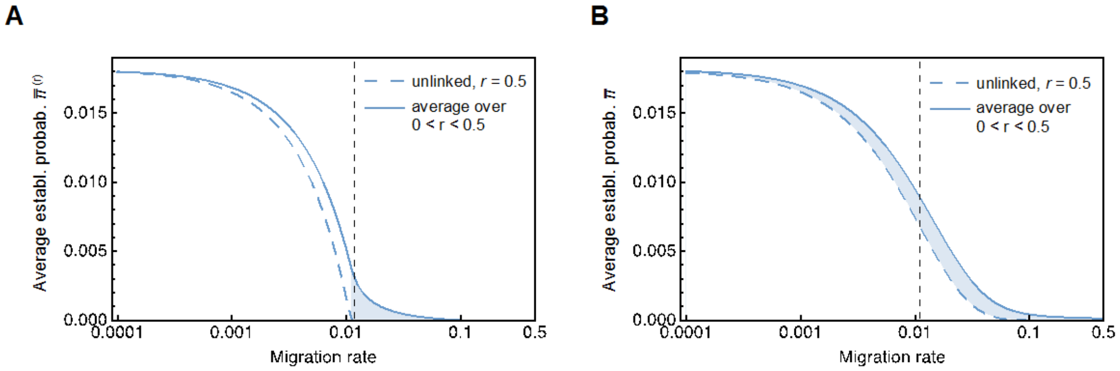

Figure 2. The effect of linkage on the average establishment probability of locally beneficial mutations. A) Contrast between an unlinked ($\bar{\pi}_U^{(r)}$, dashed curve) and an average linked ($\bar{\pi}_L^{(r)}$, solid curve) mutation of effect $a = 0.01$. B) As in A), but after averaging over an exponential distribution of fitness effects (DFE) with mean $\bar{a} = 0.01$ (dashed for $\bar{\pi}_U$ vs. solid for $\bar{\pi}_L$). In both panels, the selection coefficient at the background locus is $b = 0.1$, and the vertical dashed line indicates the critical migration rate below which a single unlinked mutation of effect $a = 0.01$ can be established. Results are shown for the two-type branching process.

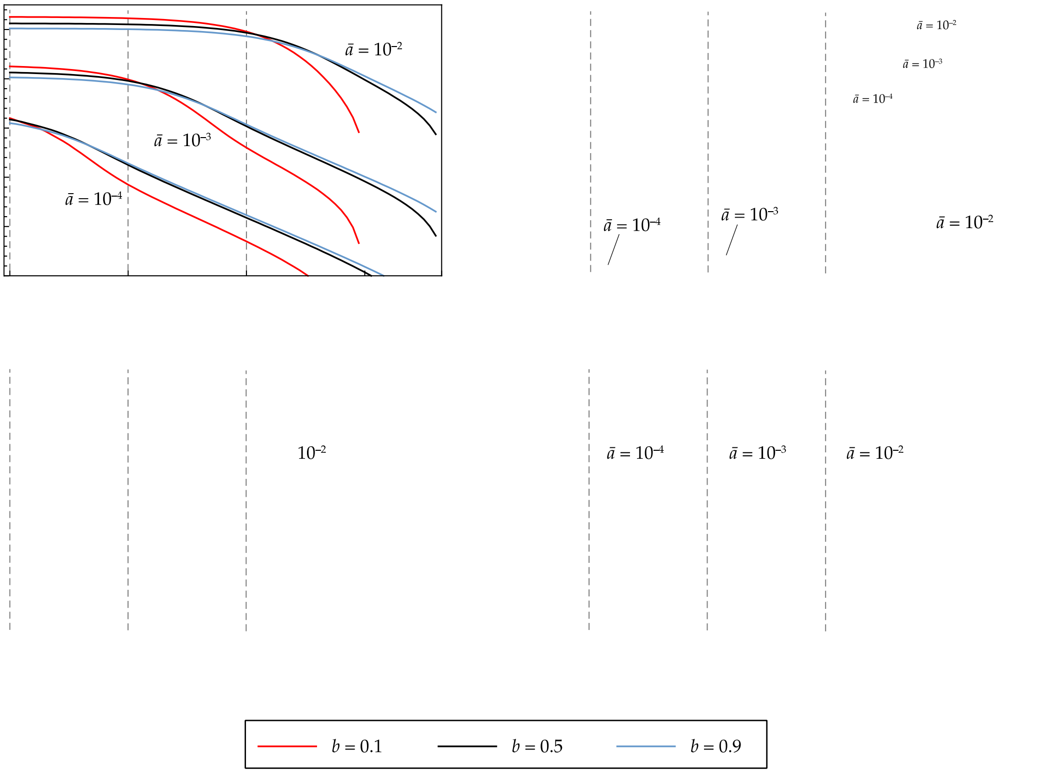

Figure 3. The effect of the migration rate on the establishment probability and window size. The average establishment probability $\bar{\pi}_L$ as a function of the migration rate for various strengths of selection on the $\log_{10}$ (A) and natural scale (B), and relative to the average establishment probability of an unlinked mutation, $\bar{\pi}_L/\bar{\pi}_U$ (C). D) The size of the window within which 95% of all successfully establishing linked *de novo* mutations occur (C_{95}). All panels show results for the two-type branching process with different values of $\bar{a}$ and b, and $k = 1$ (exponential DFE).

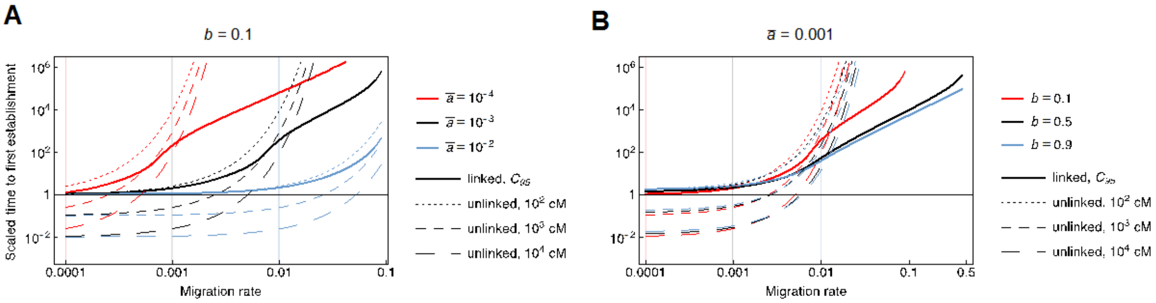

Figure 4. Waiting times to the first establishing linked and unlinked locally beneficial mutation. Expected waiting times are shown as a function of the migration rate for linked (solid curves) and unlinked (dashed curves) mutations for various mean selection coefficients $\overline{a}$ at the focal allele (A) and different selection coefficients b at the background locus (B). Times for unlinked mutations are shown assuming three different values of the unlinked mutational target C_U (different degrees of dashing for 100, 1000, and 10000 cM). Waiting times are scaled by the corresponding substitution rate in a completely isolated panmictic population of the same size (see text for details). In both panels, *k* = 1 (exponential DFE; see Figure S3 for other shapes), and predictions are based on the two-type branching process. Light grey lines included to facilitate comparison between parameter combinations.

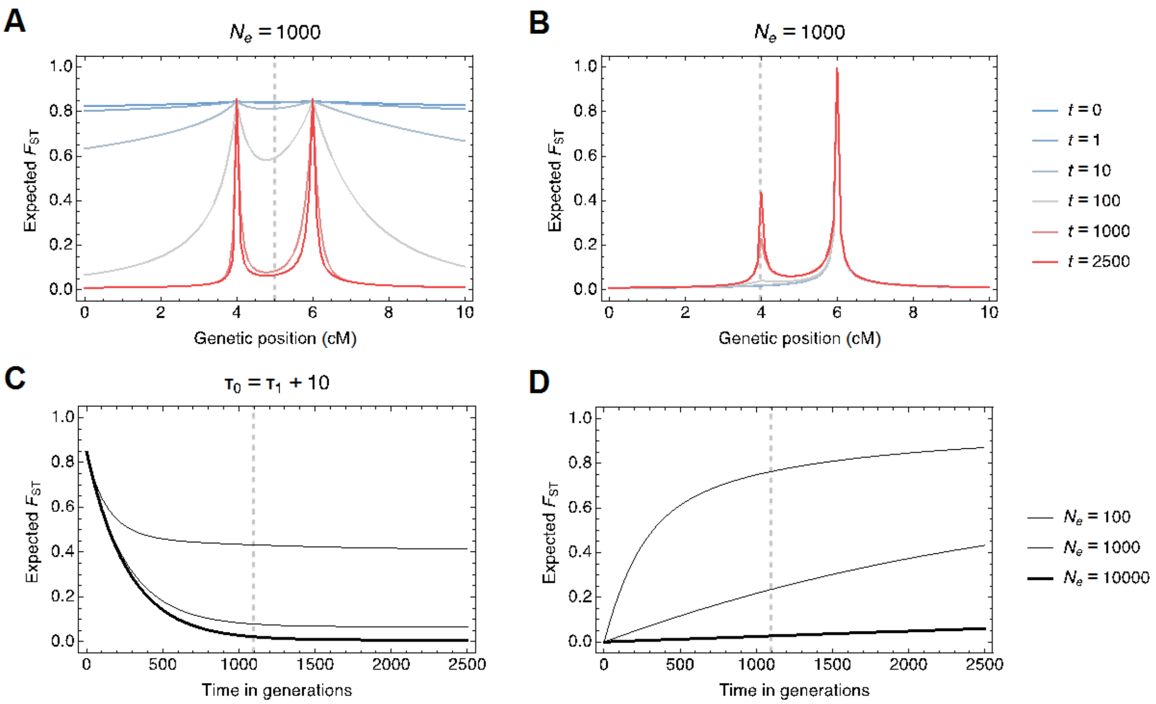

Figure 5. Dynamics of F_{ST} upon secondary contact and during the rise of a new selected allele in finite populations. A, B) Divergence between two demes of size N_e = 1000 in terms of F_{ST} at a neutral site as a function of its genetic position at various points in time. A) Formation of a two-peak island by erosion of neutral divergence in the neighbourhood of two selected loci A and B positioned at 4 and 6 cM. Before secondary contact, A and B have undergone adaptive divergence over 20 N_e generations in allopatry. B) Formation of a two-peak island through the rise of a locally beneficial *de novo* mutation at locus A in the face of gene flow and in the vicinity of a previously established migration–selection polymorphism at B. C, D) Dynamics of F_{ST} at a neutral site in the centre of the two-peak island (C) and at the position of the rising selected locus (D) for different effective population sizes N_e. Vertical lines indicate the inverse of the effective rate of gene flow, $1/m_e$, in the centre of the island, as a deterministic approximation to the waiting time for the erosion of neutral divergence. In all panels, F_{ST} is approximated as a ratio of coalescence times under the appropriate demographic model, with actual migration rates replaced by the deterministic approximation to the effective rate of gene flow, m_e, at the position of the neutral site (see SI for details). Parameters are a = 0.01, b = 0.1, and m = 0.02.