Target review

Interpreting the genomic landscape of speciation: finding barriers to gene flow

Mark Ravinet[1,2*], Rui Faria[3,4,5], Roger K. Butlin[5,6], Juan Galindo[7], Nicolas Bierne[8], Marina Rafajlović[9], Mohamed A. F. Noor[10], Bernhard Mehlig[9] & Anja M Westram[5]

[1]Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway.
[2]National Institute of Genetics, Mishima, Shizuoka, Japan
[3]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório Associado, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal
[4]IBE, Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Pompeu Fabra University, Doctor Aiguader 88, 08003 Barcelona, Spain
[5]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK,
[6]Centre for Marine Evolutionary Biology, Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden.
[7]Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain
[8]CNRS, Université Montpellier, ISEM, Station Marine Sète, France
[9]Department of Physics, University of Gothenburg, Sweden
[10]Biology Department, Duke University, Durham, North Carolina, USA

*Corresponding author: mark.ravinet@ibv.uio.no

Running head: The genomic landscape of speciation

Keywords: speciation genomics, population genomics, genome scans, genomic divergence, gene flow, reproductive isolation, selection

Abstract

Speciation, the evolution of reproductive isolation amongst populations, is continuous, complex and involves multiple, interacting barriers. Until it is complete, this process need not affect the genome as a whole and so can lead to a heterogeneous genomic landscape with peaks and troughs of differentiation and divergence. When gene flow occurs during speciation, barriers restricting migration locally in the genome lead to patterns of heterogeneity. However, genomic heterogeneity can also be produced or modified by variation in factors such as background selection and selective sweeps recombination- and mutation-rate variation, and heterogeneous gene density. Extracting the effect of gene flow, divergent selection and reproductive isolation from such modifying factors presents a major challenge to speciation genomics. We argue one of the principal aims of the field is to identify the barrier loci involved in limiting gene flow. We first summarise the expected signature of selection at barrier loci, at the genomic regions linked to them and across the entire genome. We then discuss the modifying factors that complicate the interpretation of the observed genomic landscape. Finally, we end with a roadmap for future speciation research; a proposal for how to account for these modifying factors and to progress towards understanding the nature of barrier loci. Despite the difficulties of interpreting empirical data, we argue that the availability of promising technical and analytical methods will shed further light on the important role gene flow and divergent selection have in shaping the genomic landscape of speciation.

Introduction

Speciation is the evolution of reproductive isolation between populations. This process is often continuous and complex, involving the evolution of multiple, interacting reproductive barriers among populations that do not necessarily affect patterns of variation across the whole genome at once. Since Darwin first discussed the concept of speciation, huge progress has been made in identifying the main reproductive barriers at the phenotypic level for a large number of taxa (Coyne & Orr, 2004). However, our understanding of the genetic basis of these barriers and genomic patterns associated with their evolution has remained limited until recently. Over the last decade, advances in sequencing technology have offered an unprecedented opportunity to overcome this hurdle and to investigate the genetic architecture of reproductive isolation across the entire genome and across the speciation continuum (Seehausen *et al.*, 2014). However, our understanding of the links between patterns of genomic differentiation/divergence (defined in Box 1), phenotypes and reproductive isolation is incomplete. In this review, we highlight the potential and the challenges of using genomic data, alongside other sources of evidence, to understand the evolutionary processes that shape the "genomic landscape" of differentiation and speciation, and to identify barriers to gene flow.

Recent attempts to identify loci involved in reproductive isolation, i.e. barrier loci (see Section 2 and Box 1), from high-density genetic data have largely centred on bottom-up genome scan approaches (*sensu* Barrett & Hoekstra, 2011). Regions of high genomic differentiation ("outlier loci", typically measured using F_{ST}) are often assumed to have arisen due to reproductive barriers, while homogenizing gene flow decrease differentiation elsewhere in the genome. In agreement with classic hybrid zone research (Barton & Bengtsson, 1986; Harrison, 1990; Vines *et al.*, 2003), initial genome scans revealed compelling evidence of genome-wide heterogeneity in differentiation between populations, ecotypes and species (Nosil *et al.*, 2009). While early genome scans had limited resolution, and the genomic distribution of the loci under divergent selection was mostly unknown, cheaper genome sequencing and more streamlined genome assembly pipelines are now overcoming these initial limitations. As a result, accumulating genomic data has started to reveal patterns of heterogeneity in a wide variety of non-model organisms at different stages of divergence (Table 1).

Despite progress in documenting patterns, interpreting the peaks and troughs of differentiation in genome scan data has not been as straightforward as initially assumed (Fig. 1). This has caused problems for researchers hoping to use genome scans to identify signatures of local adaptation (Hoban *et al.*, 2016) and barriers to gene flow during speciation (Noor & Bennett, 2009; Cruickshank & Hahn, 2014). There are several reasons for these difficulties. Firstly, peaks of high differentiation are produced in diverging populations without gene flow as a result of background selection and selective sweeps after isolation (Charlesworth *et al.*, 1993; Noor & Bennett, 2009; Cruickshank & Hahn, 2014; Burri *et al.*, 2015). Although some of these peaks may indicate loci that become barrier loci after contact, many other peaks do not. Instead they may reflect sweeps of universally adaptive alleles, genomic conflict, sexual selection, or drift. Therefore, the effects of barrier loci can be clearly identified only when they have actually recently acted to prevent gene flow in Nature (Harrison & Larson, 2016; Marques *et al.*, 2016; McGee *et al.*, 2016). Allopatric divergence remains important for understanding genome-wide heterogeneity in the absence of gene flow (Noor & Bennett, 2009); however barrier loci can be identified by hybridization either in the field or in the lab. Tests for on-going or recent gene flow are therefore a crucial prerequisite for the identification of barrier loci from genome scans. Secondly, patterns of F_{ST} (or other differentiation and divergence measures) are influenced by multiple factors that vary across the genome, including mutation, demographic history, genetic drift, selection, gene flow, recombination, gene density, and genome architecture; and some of these factors are expected to change during different stages of speciation (Fig. 1). From the speciation perspective, the principal objective is to infer the number, distribution and strength of barriers to gene flow, as well as their influence on other genomic regions. However, extracting this signal from genome scan data in the presence of so many other processes remains challenging.

Starting with the premise that identifying barrier loci is a major objective of speciation research, our aim with this target review is to clarify what we can expect to learn from population genomic data, specifically in examples of speciation involving periods with gene flow. We start by describing the expected patterns of local and genome-wide differentiation generated by barrier loci in idealised scenarios. We then consider how these patterns might be modified by a series of complicating factors, primarily demographic history and non-uniformity of the genome with respect to mutation, recombination and background selection. These may obscure real signatures of divergent selection and gene flow or create spurious patterns that are false positives (Box 2). We argue that it is essential to account for these factors in order to identify features of the genomic landscape related to barrier effects and so critical for the speciation process. We end with a roadmap suggesting ways in which to put inferences from the genomic landscape into context by combining them with other sources of data (e.g. experiments) to gain further insight into the speciation process.

Section 1: Barriers to gene flow in the genomic landscape

Barrier loci and barrier effects

We define barrier loci as positions in the genome that contribute to a reduction in effective migration rate (m_e) relative to the expected rate given the proportion of individuals moving between diverging populations; i.e. loci that contribute to a barrier to gene flow (see also Box 1). These loci may act independently or interact with one another, and the extent of interaction may vary as speciation proceeds. Barrier loci may involve single nucleotide substitutions or other types of mutation such as indels (Chan *et al.*, 2010; Phadnis *et al.*, 2015), or chromosomal rearrangements. These variants may be neutral within populations; *e.g.* genomic incompatibilities evolving via drift, or they may be under selection unrelated to the environment: *e.g.* meiotic drive (Presgraves, 2007). Barrier loci may be under divergent selection, either 'ecological' (Nosil, 2012) or due to reinforcement (Butlin, 1987; Servedio & Noor, 2003). Alleles at barrier loci may also be pleiotropic, affecting multiple barrier traits simultaneously, or they may influence multiple-effect traits (Servedio *et al.*, 2011; Smadja & Butlin, 2011), in either case potentially generating a strong reduction in gene flow, i.e. a strong barrier effect (see Box 1). We note that in some cases a barrier to gene flow may not necessarily require allele frequency differences at the barrier locus at all, as in one-allele models (Felsenstein 1981, Servedio 2000). Such barriers likely show different genomic patterns and may not be detectable in standard genome scans; as such, they are beyond the scope of this review.

In order for an allele at a barrier locus to under divergent selection spread and contribute to a barrier effect in the long term, selection locally favouring this allele must be strong enough to overcome the opposing effect of gene flow (Haldane, 1930; Slatkin *et al.*, 1985; Slatkin, 1987). In small populations the efficacy of selection is reduced by greater drift, and stronger selection is sometimes needed to reach a given degree of differentiation (Yeaman & Otto, 2011). The distribution of barrier locus effect sizes in a given case study is therefore likely to depend on both effective population size (N_e) and migration (m). For large populations with strong extrinsic barriers to the exchange of individuals, barrier effect sizes should vary over a wide range, whereas in small populations exchanging many migrants, only large-effect barrier loci are expected (Yeaman & Whitlock, 2011). The distribution is also expected to vary with progression towards speciation and demographic history; small effect alleles may be more common during periods of geographical isolation than during contact, and late rather than early speciation, although these scenarios need to be investigated more thoroughly. The effect-size distribution of barrier loci remains elusive because although theoretical work shows that even alleles under very weak selection may temporarily contribute to phenotypic divergence (Yeaman, 2015), loci with small fitness effect sizes are difficult to identify from empirical data. The same is true for phenotypic effect sizes; loci of large effect are easier to detect (Rockman, 2012). Empirical work often focuses on loci with large phenotypic and fitness effects, *e.g.* stickleback plate armour (Colosimo *et al.*, 2005) and pelvic spine reduction (Shapiro *et al.*, 2004; Chan *et al.*, 2010), but the general pattern remains unclear (*e.g.* Seehausen *et al.*, 2014).

At equilibrium, differentiation at a single two-allele barrier locus in a pair of hybridising populations of constant size and with constant migration rate depends on the magnitude of the barrier effect, as well as drift. This barrier effect, in turn, is determined by the strength of divergent selection, selection against hybrids or assortment directly influencing the barrier locus. How much this level of differentiation stands out from the genomic background depends on migration m and upon the effective population size N_e (i.e. via drift). These parameters determine the distribution of baseline differentiation. In addition to elevating values of differentiation (F_{ST}) and divergence (d_{XY}) at the barrier locus, the barrier effect also affects surrounding genomic regions (Charlesworth *et al.*, 1997; and see section on loci linked to barrier nucleotides below, as well as Fig. 2), generating peaks of differentiation and divergence that can be detected as outliers in genome scans (Lewontin & Krakauer, 1973; Storz, 2005; Stephan, 2016). In many cases, independent evidence (e.g. experimental data or evidence for parallel evolution) shows that outlier loci are associated with barriers to gene flow (Table 2). However, differentiation is a continuous measure and selection coefficients are continuous as well; therefore separating loci into two distinct classes, outliers and non-outliers, is an oversimplification.

Even if an outlier scan correctly identifies a genomic region containing a barrier locus, narrowing the region down to the barrier locus itself may be difficult. This is partly because measures of differentiation are noisy, due to stochasticity in coalescence as well as sampling (Fig. 3), but also due to the resolution of the scan and the chromosomal scale influenced by the barrier effect: large blocks of linkage disequilibrium can occur in some species. Given the complexity and cost of dealing with whole genome data, particularly in non-model organisms, the vast majority of genome scan studies still make use of reduced representation sequencing approaches (Davey *et al.*, 2011; Andrews *et al.*, 2016). In these cases, outlier markers may frequently show high differentiation because they are linked to a barrier locus, rather than being the direct target of selection. For genomic regions under selection, multiple SNPs may often show elevated differentiation (hence the island concept – see Box 1), although there may be variance among sites because of drift-related stochasticity. For this reason, differentiation in whole-genome data is usually calculated across a window spanning multiple variants rather than using single nucleotides. However the resolution of this approach might mean differentiated regions are missed, especially at the start of the speciation continuum when genetic differentiation decays rapidly with genomic distance (Hoban et al. 2016).

While remaining a formidable challenge in many study systems, identifying the actual loci and substitutions responsible for barrier effects (e.g. underlying divergently selected phenotypic traits or causing hybrid incompatibility) will undoubtedly improve our understanding of the speciation process. In some cases, introgression across hybrid zones may provide the necessary precision for identifying speciation genes or at least understanding how they interact. Otherwise, the strongest evidence for the role of individual substitutions is most likely to come from experimental approaches, such as mapping studies followed by the generation of transgenic individuals (Colosimo *et al.*, 2005; Cong *et al.*, 2013). Importantly, the promising future for approaches such as CRISPR (Bono *et al.*, 2015; see also Section 3) may provide information about pleiotropy, dominance and other effects that are important to understand the role of barrier loci in divergence and speciation (Storz & Wheat, 2010; Seehausen *et al.*, 2014).

Loci linked to barrier loci

Linkage causes the genomic effects of barriers to extend beyond barrier loci, as divergent selection locus reduces the local effective migration rate at linked loci. At equilibrium, the effective migration rate m_e can be approximated as $m_e = m/(1+s/r)$ in the limit of small m, s, r (Barton & Bengtsson 1986). For idealized populations in equilibrium, the relationship between F_{ST} and m_e is simple (Slatkin, 1991); therefore, the expectation is that differentiation peaks at the barrier locus and decreases with physical distance. This is one rationale for the use of reduced-representation genome scans (e.g. those based on RADseq): Rather than necessarily needing to be under selection themselves, markers may indicate the presence of barrier loci by showing elevated F_{ST} due to linkage.

However, the simple relationship between m_e and F_{ST} only holds for the situation of equilibrium between migration, selection, mutation, and drift (Whitlock & McCauley, 1999). In a transient state, where equilibrium is not yet reached (e.g. because the adaptive mutation and increase in frequency occurred only recently, or because of recent secondary contact), the distribution of F_{ST} along the chromosome is strongly contingent on the local genomic history and is not necessarily indicative of m_e. Moreover, at equilibrium and in a transient state, observed patterns of F_{ST} may rarely correspond to theoretical expectations as they are always affected by stochasticity (Fig. 3). Both effects can lead to false positives, i.e. high F_{ST} loci that are not actually indicative of a barrier locus and false negatives, low-F_{ST} regions despite close linkage to a selected locus (see Fig. 3 for examples of both). Many outlier detection methods assume simple demographic models and so may derive a null expected distribution of differentiation that does not correspond to the true distribution (Lotterhos & Whitlock, 2014; Hoban *et al.*, 2016). Clearly, if history and stochasticity are not taken into account, genome scan data may easily be misinterpreted.

One important departure from equilibrium happens during and after a selective sweep (Box 1), where an adaptive allele increases in frequency. In Fig. 2 we demonstrate the development of F_{ST} from a transient state towards equilibrium for a soft sweep under continuous gene flow (Fig. 2A), a hard sweep under continuous gene flow (Fig. 2B), and a hard sweep in allopatry followed by secondary contact for comparison (Fig. 2C). However, this figure shows differentiation averaged over a 5000 independent evolutionary histories, meaning the results obtained are in approximate agreement with the theoretical expectations (not shown). In Fig. 3 we show the outcome of a single evolutionary history to illustrate effects of stochasticity, demonstrating that deviations from the expectation are possible (see Supplementary Material for more details on the simulations run and parameters used to generate these illustrations).

For a sweep under continuous gene flow, average differentiation is increased close to the selected locus during and immediately after the sweep due to a temporary reduction of within-population diversity. The extent of the local sweep effect depends on the strength of selection, upon the starting allele frequencies at the selected locus (i.e. whether the sweep was ‘hard’ or ‘soft’ – see Box 1 and compare Figs. 2A and B), and on the time since the sweep occurred

(Przeworski, 2002; Hermisson & Pennings, 2005; Pennings & Hermisson, 2006; Messer & Petrov, 2013). However, the genomic region where average F_{ST} is increased is relatively small immediately after the sweep, and grows towards equilibrium (i.e. from left to right in Figs. 2A & B). This is because the haplotype (or haplotypes) sweeping to high frequencies contain common alleles at most loci, initially generating little differentiation. Therefore, F_{ST} may initially remain low even in genomic regions where m_e is reduced due to linkage. However, over time, this reduced m_e allows for an accumulation of allele frequency differences due to both drift and new mutations. These patterns indicate that barrier loci that have undergone sweeps in the face of gene flow may be more easily detectable when they are closer to equilibrium, because the proportion of surrounding loci showing elevated differentiation increases with time after the sweep (Fig 2 & B). However, it is unclear how quickly equilibrium is approached (Wood & Miller, 2006; Bierne, 2010; Yeaman *et al.*, 2016). This approach may be slow because it requires both mutation and rare recombination events, suggesting many loci in empirical studies are not at equilibrium.

Importantly, Fig. 2 shows averages across multiple simulations, therefore approximating expected F_{ST} values. These may differ markedly from individual outcomes of the evolutionary process, which are affected by stochasticity (Fig. 3). In Fig. 3, it becomes clear that during the transient state, a hard sweep may cause multiple loci to show high differentiation, which are interspersed by low-F_{ST} regions. This can be explained by the fact that the haplotype the selected allele occurs on harbours common and rare neutral alleles. These hitchhiking rare alleles will increase in frequency with the sweep, resulting in transient high-F_{ST} peaks that may be quite distant from the selected locus, especially if selection is strong and the sweep is rapid. In genome scan, such peaks could easily be mistaken for further selected loci, and distinguishing between them and the actual locus under selection may be difficult; nonetheless this effect is less likely for soft sweeps, where rare alleles are very unlikely to rise to high frequency. Over time, differentiation at distant loci will be lost due to gene flow, recombination and drift, reducing the probability of such false positives as equilibrium is approached. However, it should also be noted that F_{ST} is always affected by stochasticity, even at equilibrium.

In some cases the contrast between F_{ST} and d_{XY} is likely to be helpful for distinguishing between transient states and equilibrium, facilitating the correct interpretation of outlier loci (Cruickshank & Hahn, 2014; Delmore *et al.*, 2015; Irwin *et al.*, 2016). Relative measures such as F_{ST} may miss the distinct effects on diversity and divergence (Charlesworth et al. 1997), and peaks of differentiation can be present for both recent local sweeps (transient) and in equilibrium (see above and Fig. 2). Measures of absolute divergence such as d_{XY} in regions surrounding barrier loci take longer to increase via the establishment of new mutations. Recent local sweeps should be characterised by F_{ST} peaks lacking elevated d_{XY}, while in equilibrium both F_{ST} and d_{XY} are expected to be higher in the vicinity of barrier loci because of the reduction of effective migration rate (Fig. 4). Unfortunately, such distinct behaviour of F_{ST} and d_{XY} might not apply to more complex scenarios involving secondary or intermittent contact. These scenarios need further investigation.

The spread of barrier effects to linked neutral loci is uncontroversial. More contentious is the effect of a barrier locus on divergence of linked loci that are also under divergent selection. Some F_{ST} outlier analyses have identified loci that occur in proximity to QTL, e.g. at a distance of ~10 cM in pea aphids (Via & West, 2008), and divergently selected loci may cluster in the genome (Yeaman, 2013). Moreover, in some species, highly differentiated genomic regions appear to increase in size along the speciation continuum (Feder *et al.*, 2012a; Renaut *et al.*, 2012). These findings suggest that further divergence might be more likely in the vicinity of existing barrier loci, and that this might lead to a growth of highly differentiated genomic regions. Conceptual thinking has focused on one potential explanatory mechanism, divergence hitchhiking (Via & West, 2008; Feder *et al.*, 2012a; b). Under this framework, reduced m_e around divergently selected loci may facilitate the establishment of new mutations under weak divergent selection in their vicinity (Feder *et al.*, 2012a; Nosil & Feder, 2012a; Via, 2012), causing an increase in size of differentiated regions(Feder *et al.*, 2012a; Via, 2012, see also next section). However, using multi-locus simulations Feder and colleagues (2010; 2012b) demonstrated that divergent selection facilitated the establishment of weakly adaptive mutations only under limited conditions when selection is strong, N_e is small and migration is low. Furthermore, if divergence hitchhiking does occur, Hill-Robertson interference may prevent weakly adaptive alleles from establishing when they arise in habitats or genomes where they are maladaptive and they are unable to escape via recombination (Feder *et al.*, 2012b; Yeaman, 2015). Clustering under high migration load can be facilitated by chromosomal rearrangements or when linkage allows multiple weakly selected alleles to surpass the m_e threshold preventing homogenisation (Yeaman, 2013, 2015). Alternatively, clustering may occur when weak differentiation is better protected from loss via drift when linkage to a strongly diverged locus is tight (Rafajlovic *et al*. 2016). However, if drift is strong enough to allow new adaptive loci to regularly replace those lost via stochasticity, selection against recombinants can favour clustering without the need of recombination modifiers (Yeaman, 2013; Yeaman *et al.*, 2016). Theoretical and empirical evidence therefore suggests that selection against recombinants under high migration load may facilitate the clustering but not necessarily the establishment of barrier loci.

Barriers and genome-wide effects

When there are only few barrier loci, their genome-wide effect is small because most of the genome can easily recombine from one background to another. However, as speciation progresses (Coyne & Orr, 2004) and the number of barrier loci becomes large, separating the effects of different barrier loci becomes more difficult. Barrier loci may experience a reduction in local m_e both due to direct selection and due to indirect effects of linked and unlinked loci. Neutral loci throughout the genome are subject to indirect effects too, potentially resulting in a strong genome-wide barrier. Barton (1983) showed that a sharp transition from independent barrier effects to such genome-wide effects depends on the ratio of total selection to total recombination among loci. He called this ratio the ‘coupling coefficient’. The effect of coupling applies to all types of barriers (Kruuk *et al.*, 1999; Bierne *et al.*, 2011), including primary divergence with gene flow (Barton & de Cara, 2009; Abbott *et al.*, 2013). Beyond the transition to genome-wide barriers, the genomic landscape of differentiation should tend to become less structured, making barrier loci progressively more difficult to detect against increasing background differentiation. Estimating the strength of selection on individual barrier loci becomes difficult following the transition, as indirect effects increasingly contribute to their differentiation.

Selection on multiple traits, i.e. multifarious selection, is thought to be more likely to facilitate speciation than strong selection on a single trait (Rice & Hostert, 1993; Nosil *et al.*, 2008; Nosil, 2013). Similarly, selection against migrants at multiple loci results in a stronger barrier to gene flow, reducing effective migration rate across the genome when overall selection is the same (Barton & Bengtsson, 1986; Feder *et al.*, 2012b). This allows new locally-adaptive mutations to establish, independent of their genomic position, even if their effect size is relatively small; it also facilitates an increase in genome-wide divergence at neutral regions due to drift (Feder *et al.*, 2012b). This process has been termed genome hitchhiking (Feder *et al.*, 2012a) and it essentially describes the impact of multifarious divergent selection when coupling is strong. Flaxman and colleagues (2014) used simulations to demonstrate that statistical associations amongst a large number of genes combined with divergent selection can interact to drive a rapid transition from local to genome-wide barrier effects. This genome-wide congealing (GWC) is resembles the coupling transition predicted by Barton (Flaxman *et al.*, 2014; Tittes & Kane, 2014). During progression towards speciation in their model, numerous, weakly selected mutations occur but are unable to generate differentiation due to the effects of gene flow. Following a transition from local to genome-wide barriers, however, the contribution of these mutations to reproductive isolation increases as the genome-wide m_e is reduced below a threshold and LD increases (Tittes & Kane, 2014). Importantly, GWC does not require physical linkage or periods of allopatry that might elevate LD amongst loci (Tittes & Kane, 2014). Nonetheless, Flaxman *et al*. (2014) demonstrate that genomic features such as chromosome length or clustering of adaptive loci on specific chromosomes, as well as periods of geographical isolation, can drastically reduce the waiting time to GWC. Simulations show that both genome hitchhiking and genome-wide congealing are able to occur under a wide range of parameters provided there is selection on many loci (Feder & Nosil, 2010; Nosil & Feder, 2012b). However, as with divergence hitchhiking, empirical evidence showing that genome hitchhiking allows weakly adaptive alleles to establish remains elusive.

Section 2: Other factors modifying the genomic landscape

As we have seen, even in relatively simple situations with fixed population sizes and constant migration, the genomic landscape is complicated by linkage, history, and the accumulation of barrier effects. We have yet to consider additional modifying factors such as demographic history, genome-wide heterogeneity in mutation and recombination rates, background selection, and gene density.

Demographic and evolutionary history

Understanding the demographic and evolutionary history of population and species pairs is necessary to generate expected patterns of genomic differentiation. Fluctuations in effective population size (N_e) can have a profound effect in this regard; *e.g.* when N_e is small, the effect of drift is greater whereas selection is more efficient when N_e is large (Charlesworth *et al.*, 2003; Charlesworth, 2009; Charlesworth & Charlesworth, 2010). Pronounced changes in N_e such as bottlenecks can shift the mean and variance of baseline genomic differentiation, making it difficult to identify highly differentiated regions (Ferchaud & Hansen, 2016). N_e is an important parameter to estimate because, as well as determining the effectiveness of selection, it influences scaled mutation and recombination rates; for example, scaled mutation rate, $N_e\mu$ determines the rate at which adaptive mutations enter a population (Hartl & Clark, 2007; Charlesworth & Charlesworth, 2010).

We have emphasised the need to test for gene flow (see Box 3) to better appreciate the relative role of alternative processes explaining a landscape of heterogeneous genomic differentiation (see Box 2). When populations or species meet, the landscape may point to barrier loci resistant to gene flow (Harrison & Larson, 2016), but without accounting for divergence history, it is not clear whether these populations have diverged *in situ* or have resisted genome-wide homogenization upon secondary contact between divergent lineages (Bierne *et al.*, 2013; Feder *et al.*, 2013). First, with recent secondary contact, peaks of differentiation may just reflect loci that differentiated due to drift during allopatry, and have yet to be homogenised by gene flow. Such spurious outliers may obscure or hinder the detection of true barrier loci. Second, the genomic signatures of selected loci may also differ between primary divergence and secondary contact (Fig. 2). With primary divergence, during and immediately after a local selective sweep, transient high differentiation peaks will occur at large distances from the selected locus but these are eroded by recombination and migration (Fig. 3). In contrast, during allopatry, this erosion does not happen, generating large regions of high differentiation, which will be maintained for some time after secondary contact. Therefore, for local sweeps of comparable age, differentiated regions will often be much larger (and therefore potentially easier to detect) in secondary compared to primary divergence as migration has had less time to act. Recent studies have explicitly tested for primary vs. secondary contact, allowing for a more accurate interpretation of genome scan data; a wide array of tools is available for this sort of approach (Sousa & Hey, 2013; see also Box 3). Some have provided support for primary divergence (Nosil *et al.*, 2012; Butlin *et al.*, 2014) whereas others indicate that secondary contact after a period of isolation best explains heterogeneous differentiation (Tine *et al.*, 2014; Martin *et al.*, 2015b; Roesti *et al.*, 2015; Rougemont *et al.*, 2016).

Even sophisticated statistical frameworks for testing divergence hypotheses only consider a small proportion of the ‘universe of potential historical scenarios’ (Knowles, 2009). Divergence history varies across the genome due to non-uniformity in effective migration rate, effective population size and recombination (Maddison, 1997; Roux *et al.*, 2014, 2016; Mallet *et al.*, 2016). Gene-tree vs. species-tree discordance can occur because of introgression (Maddison, 1997; Knowles & Maddison, 2002; Geneva *et al.*, 2015; Rosenzweig *et al.*, 2016), but also because of incomplete lineage sorting (ILS) (Hobolth *et al.*, 2011b; Dutheil & Hobolth, 2012). Described as 'deep coalescence' by Maddison (1997), ILS occurs when the most-recent common ancestor for a genealogy exists before speciation begins, resulting in counter-intuitive three taxa phylogenies (Scally *et al.*, 2012) or distortions of divergence time estimates between two species (Leaché *et al.*, 2013). ILS therefore increases the variance of genomic divergence estimates, making it difficult to identify true outliers and also potentially introducing false positives. ILS affects a greater proportion of the genome when speciation events occur close in time and the ancestral effective population size is large (Barton, 2006; Hobolth *et al.*, 2011b). This presents an obvious challenge to studies of multiple species pairs or adaptive radiations (Mallet *et al.*, 2016). Furthermore, stochasticity in divergence times and ILS at neutral loci can generate false signals of both genomic divergence and gene flow between species pairs (Barton, 2006; Pease & Hahn, 2013; Cruickshank & Hahn, 2014). Incorporating demographic history in tests for selection is difficult as incorrect specification of the history, potentially generated by ILS patterns, increases error rates (Lotterhos & Whitlock, 2014; Aeschbacher *et al.*, 2016; Fraïsse *et al.*, 2016a; Hoban *et al.*, 2016; Le Moan *et al.*, 2016). Approaches that do not use demographic models may be preferable in some cases although these too are prone to bias (Hoban *et al.*, 2016).

Speciation is undoubtedly complex, unfolding in space and time with populations overlapping, contracting and re-expanding (Butlin *et al.*, 2008; Abbott *et al.*, 2013; Seehausen *et al.*, 2014). This complexity suggests that most species have probably evolved with gene flow occurring at some point in their evolutionary history (Smadja & Butlin, 2011) and that the process cannot easily be delineated into primary vs. secondary contact or with vs. without gene flow (Bierne *et al.*, 2013; Cruickshank & Hahn, 2014). A genic perspective on speciation predicts that divergence history will vary across the genome (Wu, 2001; Wu & Ting, 2004) therefore the history of barrier loci might not necessarily reflect the history of populations, as *Heliconius* butterflies, *Anopheles* mosquitoes and marine-freshwater sticklebacks appear to show (Bierne *et al.*, 2013; Mallet *et al.*, 2016). Adaptive alleles may evolve during a period of geographical isolation but introgress between divergent lineages via hybridisation and only act as barrier loci in a later phase of *in-situ* divergence between populations (Bierne *et al.*, 2011, 2013). Ancient divergence times for adaptive variants in several systems also suggest that these alleles are maintained as standing variation and spread between populations as a result of gene flow, repeatedly becoming involved in divergence (Colosimo *et al.*, 2005; Lamichhaney *et al.*, 2015;

Fraïsse *et al.*, 2016b). Coupling between independently evolved ancient adaptive alleles and incompatibilities due to selection across environmental gradients may drive progress towards speciation over shorter timescales (Barton & de Cara, 2009; Bierne *et al.*, 2011, 2013; Abbott *et al.*, 2013). As well as ancient adaptive variants, intrinsic genomic incompatibilities arising from epistatic interactions appear to segregate within species (Shuker *et al.*, 2005; Corbett-Detig *et al.*, 2013). Although such incompatibilities are difficult to detect, their presence suggests the possibility of widespread potential for coupling with adaptive alleles.

Mutation rate variation

In the absence of gene flow and selection, neutral diversity within and divergence between populations scales with mutation rate. In the human genome, for example, nucleotide diversity is positively correlated with *de novo* mutation rate, which in turn accounts for a third of sequence divergence variation between humans and chimpanzees (Francioli *et al.*, 2015). Mutation rate variation amongst species, populations and individuals and the implications of this for evolutionary inference are relatively well understood (Drummond *et al.*, 2006; Ho & Larson, 2006; Hodgkinson & Eyre-Walker, 2011). However, absolute mutation rates (i.e. the number of mutations per site and generation) are also non-uniform across the genome (Hodgkinson & Eyre-Walker, 2011; Ness *et al.*, 2015). Mutation probability is influenced by G:C bases and neighbouring base identity (Hodgkinson & Eyre-Walker, 2011; Ness *et al.*, 2015). Replication timing also has an effect, with longer exposure to mutagens during transcription in late replicating regions (Hodgkinson & Eyre-Walker, 2011; Francioli *et al.*, 2015). Mutation rate is often higher on Y-chromosomes than the X or autosomes because 100% of Y chromosomes occur in males, experiencing higher mutation rates due to spermatogenesis (Hodgkinson & Eyre-Walker, 2011). Despite knowledge of mechanisms causing mutation rate variation, it remains contentious whether systematic genome-wide variation occurs at a scale that might bias genome scans. For example, while Ness *et al.* (2015) detected fine-scale heterogeneity in mutation rate, they found no clear variation amongst 200 kbp genome windows (Hodgkinson & Eyre-Walker, 2011), suggesting that the extent of any bias in genome scans will also differ with the scale of the analyses.

Irrespective of the scale at which it varies, mutation rate is an important population genetic parameter used to scale estimates of parameters such as effective population size (N_e) and divergence time (t) derived from genomic data. Since N_e is typically estimated from θ ($4N_e\mu$ - scaled mutation rate on autosomes, where μ = absolute mutation rate), assuming a uniform mutation rate will inflate estimates of N_e for mutational hotspots, obscuring the extent to which drift or selection contributes to divergence in these genomic regions (Charlesworth, 2009). Furthermore, given the importance of estimating demographic parameters for determining how and when speciation has occurred (see *Demographic and evolutionary history*), uniform mutation rates incorrectly applied across the genome may obscure the history of barrier loci and the speciation process (Scally & Durbin, 2012). Mutation rate variation also has implications for genomic differentiation; high mutation rate at some genomic regions may downwardly bias local measures of relative differentiation, e.g. F_{ST}, obscuring loci putatively under selection (Foll & Gaggiotti, 2008). Absolute divergence measures such as d_{XY} are also subject to bias due to mutation rate variation; a low mutation rate will result in low levels of divergence, potentially giving a false impression of constraint or introgression (Geneva *et al.*, 2015; Rosenzweig *et al.*, 2016).

Genome-wide mutation rate variation should be taken into consideration in order to interpret the genomic landscape accurately. To-date, our understanding of intra-genomic mutation rate variation remains limited and is drawn from a relatively small number of model organisms. Quantifying this heterogeneity is a major endeavour even with high throughput sequencing technologies (Ness *et al.*, 2015). Nonetheless, there is considerable promise for incorporating mutation rate estimates into predictive models (Francioli *et al.*, 2015; Ness *et al.*, 2015; see *Section 3: Roadmap*).

Background selection and selective sweeps at non-barrier loci

Advantageous mutations involved in adaptive evolution are of greatest interest in speciation research as in many cases, these generate the barrier alleles we wish to detect (see Section 2; Seehausen *et al.*, 2014). However, they are rare; most non-neutral *de novo* mutations are likely to be deleterious (Ohta, 1992; Eyre-Walker & Keightley, 2007), and their removal from populations by selection, i.e. background selection, can shape the genomic landscape of variation in a similar way to positive selection on adaptive alleles (Charlesworth *et al.*, 1993; Stephan, 2010). Purging of deleterious mutations by purifying selection removes neutral variation at linked sites, reducing genetic diversity and local effective population size (Charlesworth *et al.*, 1993; Charlesworth, 2012; Cutter & Payseur, 2013).

Like the other processes described in this section, the extent of background selection varies across the genome. Evidence from *Drosophila melanogaster* suggests it is highest on autosomes, accounting for 58% of the observed variation in nucleotide diversity across 100 kbp windows (Comeron, 2014). Simulations based on theoretical approximations show the effects of background selection on patterns of diversity are greatest when deleterious mutation rate is high and recombination rate is low, i.e. when linked neutral sites are unable to escape via recombination from new mutations entering a population (Charlesworth *et al.*, 1993; Charlesworth, 2012). Background selection should also be higher in genomic regions with a high density of coding sequence, where mutations are more likely to have deleterious effects; this is supported by lower diversity in these regions (Lohmueller *et al.*, 2011; Cutter & Payseur, 2013; Enard *et al.*, 2014). Whether or not mutations are deleterious within a coding region may vary with proximity to optimum fitness on a adaptive landscape; when a population is close to maximum fitness, a greater proportion of mutations will be deleterious, causing a shift away from the optimum (Orr, 1998; Cutter & Payseur, 2013). On a genome-wide level, drastic reductions in effective population size can limit background selection as the frequencies of new deleterious mutations are more strongly influenced by drift (Charlesworth, 2012).

Despite being different processes, background and positive selection may produce similar patterns of reduced intraspecific diversity and increased interspecific genomic differentiation in genome scans using relative measures like F_{ST} (Noor & Bennett, 2009; Cruickshank & Hahn, 2014). Distinguishing between them is important in order to identify barrier loci under divergent selection and rule out false positives; ideally, positive selection should be tested against a null-evolutionary model that incorporates background selection (Cutter & Payseur, 2013; Comeron, 2014; Zeng & Corcoran, 2015; Elyashiv *et al.*, 2016). Predictive models incorporating background selection are able to estimate the contribution of the process to differentiation (Lohmueller *et al.*, 2011; Comeron, 2014; Zeng & Corcoran, 2015; Elyashiv *et al.*, 2016). Similarly, outlier analyses and demographic inferences that account for signatures of background selection are more robust, with fewer false positives (Ewing & Jensen, 2016; Huber *et al.*, 2016). To-date however, only a few studies have attempted to account for background selection in the context of speciation and barrier loci (e.g. Roesti *et al.*, 2013; Burri *et al.*, 2015; Delmore *et al.*, 2015; Feulner *et al.*, 2015).

Global selective sweeps of universally adaptive alleles (see Box 1), i.e. those adaptive in both diverging populations, may also generate signatures similar to barrier loci. Divergence history may involve phases of allopatric isolation, during which universally adaptive mutations can become fixed in only one subpopulation because gene flow is absent. This generates a peak of differentiation that will decay with the introgression of the adaptive allele to the other subpopulation when contact and gene flow are restored. However, homogenisation of allele frequencies after secondary contact does not occur instantaneously, and peaks of differentiation will be maintained during early phases of gene flow, potentially being misinterpreted as indicating barrier loci (see Fig 1). Similar effects may occur at loci that do not contribute to adaptation to the environment or speciation at all, but that are subject to sexual selection,genomic conflict or drift occurring independently in geographically isolated subpopulations.

Even with continuous gene flow, recent sweeps of universally favourable alleles may temporarily generate high differentiation peaks. The spread of favourable mutations amongst subpopulations will take time and can cause temporary allele frequency differences, especially if subpopulations are large or the magnitude of gene flow between them is relatively low. Furthermore, the original hard sweep will strongly reduce diversity in regions flanking the selected locus, leading to a single haplotype at high frequency in the source population. The lag time between mutations arising and spreading means recombination events between the flanking haplotype and others are more likely to occur in the second subpopulation (i.e. a soft sweep). Consequently, different haplotypes will increase in frequency in the other subpopulation, leading to elevated differentiation at regions flanking the selected locus, but not the selected locus itself, generating two adjacent peaks (Bierne, 2010; Roesti *et al.*, 2014). This signature may be distinguishable from a single peak of divergent selection, but only if sufficiently large chromosomal regions are studied.

Recombination rate variation

With a uniform recombination rate across the genome, the width of a genomic region of differentiation surrounding a barrier locus is directly proportional to the strength of the barrier effect (Barton & Bengtsson, 1986). In reality, however, recombination rate varies widely across the genome of most species studied (Jensen-Seaman & Furey, 2004). This may be associated with chromosome type (i.e. sex chromosomes vs. autosomes), distance to the centromere, GC content, CpG motifs, transposable elements, polyA and polyT sequences, gene density and recombination modifiers (Butlin, 2005; Smukowski & Noor, 2011 and references therein), or, on a fine scale, with recombination hotspots (Myers *et al.*, 2010; Massy, 2013). Since many of these factors are associated, determining the true cause of recombination rate variation is difficult but its effects on genomic variation are more predictable. A barrier locus will influence a larger genomic region when it occurs in a low-recombination region compared to a high-recombination region (Stephan, 2010; Cutter & Payseur, 2013). Therefore, it might be easier to detect in a genome scan, but harder to narrow down to small functional regions or individual nucleotides. This alone is justification enough to account for recombination rate variation when interpreting patterns of differentiation across the genome (Nachman & Payseur, 2012; Roesti *et al.*, 2012). However, a strong correlation between recombination rate and nucleotide diversity (Begun & Aquadro, 1992; Cutter & Payseur, 2013) suggests that recombination rate variation can confound interpretation of the genomic landscape in other ways too.

While recombination rate has a mutagenic effect, this does not appear to be correlated with genomic divergence (Noor, 2008; Charlesworth & Campos, 2014). Indeed, controlling for mutation rate variation shows recombination determinines the extent human-chimpanzee divergence in other ways (Francioli *et al.*, 2015). Background selection reducing genetic diversity in regions of low recombination is a compelling explanation for these patterns (Charlesworth *et al.*, 1993; Cutter & Payseur, 2013). Neutral alleles in low recombination regions are more frequently in LD with deleterious mutations and so experience a stronger purging effect (Charlesworth *et al.*, 1993; Charlesworth, 2012). This leads to a reduction in within-population diversity, while measures of absolute divergence (d_{XY}) remain largely unaffected, provided gene flow is sufficiently low (Charlesworth *et al.*, 1997; Noor & Bennett, 2009; Cruickshank & Hahn, 2014; Zeng & Corcoran, 2015; but see also Phung *et al.*, 2016). However, measures of relative differentiation (F_{ST}) will be inflated and some regions may appear as outliers. High differentiation between species has indeed been observed in low recombination regions, e.g. close to centromeres (*e.g.* Nachman & Payseur, 2012; Roesti *et al.*, 2012). Nonetheless, it remains unclear how low gene flow between populations must be for background selection in recombination cold spots to cause false positive signals of differentiation.

Importantly, recombination can influence selection beyond its signature in genome scans. High recombination allows the independent evolution of individual selected positions, counteracting Hill-Robertson interference (Stephan, 2010; Gossmann *et al.*, 2014). The more efficient fixation of adaptive mutations can potentially lead to greater divergence in high recombination regions (Bullaughey *et al.*, 2008). Conversely, when recombination is absent, lower efficacy of selection at eliminating weakly deleterious mutations can lead to increased non-synonymous divergence (Haddrill *et al.*, 2007; Bullaughey *et al.*, 2008). However, the effects of gene flow on the relationship between recombination and efficacy of selection have not been studied. Additionally, regions of reduced recombination may allow existing barrier loci to shield closely linked, newly established barrier loci under weaker selection from stochastic loss (Rafajlovic *et al.*, 2016). Clusters of barrier loci may be more likely to evolve in low recombination regions and it is possible that recombination suppressors evolve because they enhance clustering effects (Yeaman, 2013).

The speciation process can also be expected to alter how recombination varies across the genome; divergent selection between populations connected by gene flow should favour the spread of recombination modifiers such as chromosomal rearrangements that decrease recombination between barrier loci (Kirkpatrick & Barton, 2006; Ortiz-Barrientos *et al.*, 2016). Because recombination is suppressed in heterokarotypes, linkage between barrier loci can be maintained within chromosomal rearrangements and these are expected to show higher differentiation and divergence than collinear regions that will be homogenised by gene flow (Noor et al 2001; Jackson et al 2016). As with other low recombination regions, alternative explanations must be ruled out. For example, ancient rearrangements, pre-dating speciation, may show inflated divergence and differentiation compared to the genome-wide average (Noor & Bennett, 2009).

Gene density

With the large number of assembled and annotated genomes now available, mapping gene positions and estimating gene density is possible for more and more taxa. This has clearly shown that genes are not randomly distributed across the genome (Hurst *et al.*, 2004; Sémon & Duret, 2006; Al-Shahrour *et al.*, 2010). First, genes may cluster and form gene-rich regions, while other parts of the genome may contain hardly any functional loci (Nobrega, 2003; Hellsten *et al.*, 2010). Genes may also be grouped by function, and the expression of these groups may be regulated simultaneously (Hurst *et al.*, 2004; Al-Shahrour *et al.*, 2010). The causes for this are not clear but likely involve tandem duplications, chromatin structure and shared regulatory elements (see Hurst *et al.*, 2004 for a review). Irrespective of their cause, clusters of functionally similar and co-expressed genes are likely to be favoured by selection (Hurst *et al.*, 2002; Al-Shahrour *et al.*, 2010), although clustering may also evolve neutrally (Sémon & Duret, 2006). The non-random distribution of genes in the genome, as well as their functional grouping, can influence processes acting throughout the genome, playing an important role in shaping the landscape of genomic differentiation.

Functional genomic regions, which includes genes as well as transcription factor binding sites, rDNA and regions coding for microRNAs, are more likely to experience positive and background selection than non-functional regions, where mutations have no consequence. Because background selection can reduce N_e locally in the genome, a negative correlation between gene density and polymorphism is expected (Nordborg *et al.*, 2005; Hobolth *et al.*, 2011b; Flowers *et al.*, 2012). Similarly, a higher probability of local selective sweeps in these parts of the genome will reduce within-population diversity (Stephan, 2010). High recombination can limit the impact of such reductions in diversity; polymorphism is positively correlated with recombination rate (Hey & Kliman, 2002; Nordborg *et al.*, 2005). Indeed, it has been demonstrated that gene density can show a positive relationship with recombination rate (Duret & Arndt, 2008; Flowers *et al.*, 2012). This may simply be an emergent property of the transcription process, which increases recombination rate (Kim & Jinks-Robertson, 2012). Alternatively, a higher recombination rate in gene-dense regions might be directly favoured by selection, because both positive and negative selection are more efficient when the extent of Hill-Robertson interference between multiple selected sites is reduced (Hey & Kliman, 2002, see also *Recombination rate variation*).

Importantly, gene density influences the efficacy of selection independently of recombination rate; e.g. selection efficiency is negatively correlated with gene density in regions of both high and low recombination (Hey & Kliman, 2002). However, this only holds true above a threshold level of high gene density, suggesting a trade off between selective interference and the advantages of co-expression of clustered genes (Hey & Kliman, 2002). This potentially has implications for the spatial proximity of barrier loci in the genome. Increased Hill-Robertson interference due to high gene density relative to recombination rate may be advantageous for the maintenance of clusters of adaptive genes under divergent selection. Beneficial combinations are less likely to be broken up, but will take longer to come together. Barrier loci in gene-dense regions may also need higher selection coefficients to overcome the reduction in local effective population size caused by background selection.

The grouping of genes with related functions can also be expected to influence large-scale mechanisms in the speciation process when gene flow is occurring, e.g. the evolution of inversion polymorphisms or divergence hitchhiking (see *Loci linked to barrier nucleotides*). Functional grouping means multiple loci affecting the same divergently selected trait or suite of traits may be physically linked (Hurst *et al.*, 2004; Al-Shahrour *et al.*, 2010). Inversions are mainly adaptive if they capture multiple barrier loci (Kirkpatrick & Barton, 2006; Faria & Navarro, 2010), and the potential for capturing multiple barrier loci in an inversion when gene flow is occurring is higher if they are grouped. Divergence hitchhiking occurs when adaptive mutations arise close to an established barrier locus and are shielded from gene flow. When functionally related genes are closely linked, new mutations occurring in the same genomic region are more likely to be adaptive than if genes are randomly distributed, this increases the potential for divergence hitchhiking. Similarly, new adaptive mutations would also be better protected against stochastic loss (Rafajlovic et al. 2016).

Section 3: A roadmap for the genomic landscape

The genomic landscape of differentiation has now been described in many species. Both the number of examples and the genomic resolution are increasing, with many studies now providing nucleotide-level descriptions for a large proportion of the genome with multiple replicates (examples in Table 1). The problem however is not to generate these descriptions but to interpret them; a difficult challenge because we know that the landscape depends on multiple factors. To identify barrier loci properly, the parameter of primary interest is the local effective rate of gene flow, m_e. This is determined by the actual migration rate and the local barrier effect, which comprises the direct barrier effect (if any) and the influence of other barrier loci, mediated by recombination. The influence of any indirect barrier effect will depend on local recombination rate and gene density. Both direct and indirect effects, in turn, may be confounded by the impact of population history on the genome, itself dependent on local mutation rate, recombination rate, background selection or global and local selective sweeps not related to species specialisation and speciation.

With so many modifying factors, interacting in complex ways, the prospects for disentangling the genomic landscape might seem bleak. We believe this conclusion premature; in this section, we outline a roadmap for future research in speciation genomics to overcome the issues faced by the field. Our roadmap will not be feasible in all study systems, but it should represent a guideline for researchers to work with. Over the last 15 years, since the publication of Wu's (2001) 'genic view', a huge number of empirical studies have provided previously unimagined insight into how speciation has progressed, and this number is still increasing. We argue that with a carefully considered approach, ongoing speciation research will provide us with an even greater understanding of the "mystery of mysteries".

Step 1: Know the study system

Although perhaps obvious, a strong biological background for a study system cannot be over-emphasised. Many of the most insightful recent speciation genomics studies have been on taxa with a rich literature on many aspects of their biology such as three-spined sticklebacks (McKinnon & Rundle, 2002; Jones *et al.*, 2012), *Heliconius* butterflies (The Heliconius Genome Consortium, 2012) and African cichlids (Keller *et al.*, 2012; Brawand *et al.*, 2014). This background includes a solid understanding of the ecology, reproductive biology, life history strategies and geographical distribution with a special focus on phylogeography and evolutionary history. Crucially, genetic data should be supplemented with other evidence, from a variety of sources such as fossil and historical records or experimental data on movement between populations, in order to constrain the range of testable scenarios and to provide limits on parameter estimates. Information on the mechanisms of pre- and postzgotic isolation and the contributions of different components to overall isolation will also aid in the interpretation of barrier loci. Knowledge of the biological background of a system should be used to inform sampling strategies. We additionally recommend broadening the geographic and taxonomic range of sampling where possible to account for unsuspected sources of introgression (e.g. Martin *et al.*, 2015a).

Step 2: Establish the extent of gene flow and understand the demographic history

Gene flow is clearly fundamental for studying the genomic basis of reproductive isolation. A study system should therefore be sampled where divergent populations or species meet (Marques *et al.*, 2016; McGee *et al.*, 2016). Testing for and quantifying the extent of gene flow is a crucial prerequisite for interpreting genomic analyses correctly; ideally both genomic and additional evidence of gene flow (e.g. individuals in natural populations showing evidence of introgression) should be identified. Quantifying gene flow is explicitly linked to an understanding of the demographic history of a pair of populations or species. Reconstructing the evolutionary history is desirable as it can have important effects on the genomic landscape (see *Section 2: Demographic and evolutionary history*). Care should be taken to distinguish between population level processes such as fluctuations in effective population size (Li & Durbin, 2011) and genome-wide variation in demographic parameters (Roux *et al.*, 2014, 2016). Fortunately, both can be incorporated into flexible hypothesis testing frameworks such as coalescent modelling and Approximate Bayesian Computation (Ewing & Jensen, 2016; Roux *et al.*, 2016). Given the importance of this step, Box 3 discusses methods that are useful to test for the presence of gene flow and to infer demographic history in more detail.

Step 3: Capture the best possible picture of the genomic landscape

A wealth of next-generation sequencing approaches exists for representing the genomic landscape accurately, nearly all of which have been used in a genome scan context (Table 1). Relatively inexpensive and easy to apply to non-model organisms, reduced representation techniques such as RAD-seq, RNA-seq and target capture sequencing have quickly gained ground as popular tools for population genomics (Davey *et al.*, 2011; Andrews *et al.*, 2016). These methods can clearly identify patterns of heterogeneity and outlier loci (examples in Table 1). They have also successfully been used to successfully reconstruct population history (Shafer *et al.*, 2015), estimate genome-wide recombination rate variation (Roesti *et al.*, 2013) and identify signatures of selection (Roesti *et al.*, 2015). Although *de novo* assembly of reduced representation markers can prove useful for identifying outlier loci (Le Moan *et al.*, 2016; Ravinet *et al.*, 2016; Rougemont *et al.*, 2016), ideally a reference genome and genetic map are required to place markers in a genomic context. With such resources, it is possible to test whether divergent loci cluster in the genome (Renaut *et al.*, 2013; Marques *et al.*, 2016), to estimate the size of differentiated regions (Nadeau *et al.*, 2012, 2013) and to ask whether higher differentiation is found predominantly in regions of low recombination (Roesti *et al.*, 2013; Tine *et al.*, 2014; Delmore *et al.*, 2015; Marques *et al.*, 2016).

However, reduced representation sequencing may not always be the ideal choice for identifying barrier loci because of their relatively low genome coverage (*e.g.* 0.45% of 0.4 Gb three-spine stickleback genome; Hohenlohe *et al.*, 2010). Markers will rarely be the direct targets of selection. In low recombination regions, physical distance between barrier loci and markers that are outliers is likely to be large; in high-recombination regions, barrier loci are less likely to be detected in the first place as the scale of LD is small. Furthermore, these methods may bias studies in favour of identifying barrier loci with single nucleotide substitutions, overlooking structural variants, rearrangements and changes in genome organization that can only be detected reliably by using long insert mate pair libraries (Jones *et al.*, 2012). Most importantly, users should be aware of the pitfalls and biases unique to each different reduced representation method that may ultimately distort the picture of the genomic landscape; e.g. null alleles and sequence length bias in RAD-seq (Davey *et al.*, 2013; Gautier *et al.*, 2013; Ravinet *et al.*, 2016) and bias towards conserved genic regions or overexpressed alleles in RNA-seq (Hoban *et al.*, 2016).

Whole-genome re-sequencing is becoming increasingly affordable as an alternative to reduced representation approaches and has been used successfully in multiple taxa (see Table 1 for examples). Although it still requires a well-assembled reference, resequencing provides good genome-wide coverage, mitigating the problem of not targeting barrier loci. Furthermore, resequencing can help to identify structural variation, duplications, copy number variation, translocations and inversions that prove elusive with a reduced marker set. Hybrid assemblies combining both long and short read technologies have proven successful in producing high quality assemblies incorporating structural variation (English *et al.*, 2012; Wang *et al.*, 2015). Nonetheless, difficult to assemble features such as highly repetitive regions are likely to be missed even with new approaches (Hoban *et al.*, 2016). For those with fewer resources, resequencing might seem daunting. However a feasible option is to sequence a small number of individuals (i.e. one or two) to high depth and many other individuals to much lower depth (Glazer *et al.*, 2015). This hybrid approach also allows high-depth data to be used for other purposes such as demographic inference, genome annotation and assessing structural variation. Pool-seq, i.e. sequencing with barcoding of population samples rather than individuals, can also be used to estimate population allele frequencies and reduce sequencing costs (Schlötterer *et al.*, 2014; Christe *et al.*, 2016).

Step 4: Measure genomic factors that contribute to the differentiation landscape

Measuring factors influencing the genomic landscape is difficult, but not insurmountable. Genome-wide recombination rate variation can be documented by mapping in experimental crosses (Roesti *et al.*, 2013) or pedigrees (Kong *et al.*, 2002; Kawakami *et al.*, 2014). LD-based methods using population genetic data are also able to estimate average realised recombination across the population and over time (Tine *et al.*, 2014), which may be more relevant in the landscape context (Smukowski & Noor, 2011). Whichever approach is used, high-density genomic markers and large numbers of individuals are essential since it is clear that recombination rate can vary on a small genomic scale (Roesti *et al.*, 2013; Kawakami *et al.*, 2014). Furthermore, if possible, a comparative recombination mapping approach, i.e. using all taxa studied, should be taken to account for differences between closely-related species (Renaut *et al.*, 2013).

Directly measuring genome-wide variation in mutation rate is likely to be more difficult, especially in non-model organisms with long generation times. Estimates at putatively neutral sites using phylogenetic methods remain valuable (Kondrashov & Kondrashov, 2010; Scally & Durbin, 2012). However these estimates are prone to bias depending on the timescale over which they are estimated (Ho *et al.*, 2005; Ho, 2014), and they do not incorporate deleterious or weakly deleterious mutations: i.e. they are substitution, not mutation rates. If possible, whole-genome sequencing within families using parent-offspring trios provides a direct measurement of genome-wide mutation rate heterogeneity and also allows classification of mutations, as adaptive, deleterious or neutral (Francioli *et al.*, 2015). Mutation accumulation lines offer an experimental approach in lab-based populations; natural selection is reduced and mutations are allowed to accumulate even if they would otherwise have negative fitness consequences (Ness *et al.*, 2015).

Estimating gene density relies on a high quality reference genome and precise annotation; with accurate annotation, gene density is relatively easy to quantify (Hurst *et al.*, 2002; Al-Shahrour *et al.*, 2010). Precise genome annotation, aided with transcriptomic data, should also mean that measures of gene density are feasible for most organisms following genome assembly. However, greater effort needs to be made to better annotate regions that are not protein-coding but still play a functional role, e.g. regulatory regions. Importantly, measuring gene density via annotation may also provide insight into other confounding factors influencing the genomic landscape, potentially overcoming limitations for non-model organisms. For example, recombination hotspots may be predicted by identifying transposons and sequence motifs recognised by recombination modifier genes (Myers *et al.*, 2010). Similarly, models using the spatial distribution of CpG dinucleotides, flanking sequence and other mutation rate modifiers could potentially be used to estimate mutation rate variation (Francioli *et al.*, 2015; Ness *et al.*, 2015).

Step 5: Identify selection at barriers, taking modifying factors into account

To properly identify the signature of selection properly, controlling for factors that modify or mimic the signature of barriers to gene flow is essential. Previous work has attempted to do this, at least in part, e.g. removing the effects of recombination rate variation by either correcting local estimates of differentiation for regional differentiation (Roesti *et al.*, 2012), correlating differentiation with recombination rate (Renaut *et al.*, 2013) or focusing on barrier loci in high-recombination regions (Marques *et al.*, 2016). Clearly much of the focus to-date has been on recombination rate variation although mutation rate has been tentatively linked to genomic differentiation using indirect measures such as synonymous divergence (d_S - Renaut *et al.*, 2014). Human-chimpanzee sequence divergence models incorporating both mutation and recombination rate variation also show promise in partitioning these effects (Francioli *et al.*, 2015).

Ultimately, the aim should be to infer selection using models that account for variation in multiple confounding factors. It is now possible to detect hard selective sweeps in a single population by including fixed differences with an outgroup to account for mutation rate variation and by scaling the site frequency spectrum by estimates of background selection derived from mutation and recombination rate variation and genome annotation data (Huber *et al.*, 2016). However, this has yet to be extended to cases of divergence with gene flow. Methods using genome-wide measures of recombination rate variation and nucleotide diversity in order to estimate the intensity and timing of selection and gene flow are also now available and can be extended to include background selection (Aeschbacher *et al.*, 2016). Such methods can only be used if independent measurements of these factors (see Step 4) are combined with genome scan data. Modelling the genomic landscape with local estimates of recombination rate, mutation rate and gene density, can then be used to ask whether we need to invoke divergent selection and gene flow to explain peaks of high differentiation (Cruickshank & Hahn, 2014).

Systems of parallel divergence or speciation may also be helpful in separating the effects of various factors (Irwin *et al.*, 2016). For example, when recombination rate variation is correlated among closely related taxa, high differentiation in low-recombination regions that appear in multiple species pairs is more likely to have arisen due to background selection (Burri *et al.*, 2015). This is especially true if contrasts involve different types of barriers to gene flow, and if the same highly differentiated regions occur in comparisons with and without gene flow. However, as a caveat, differentiated regions shared amongst contrasts may sometimes still be due to the same loci under divergent selection. Disentangling these explanations is only possible with information on gene density, mutation rate, the types of barriers involved, and the history of gene flow. Nonetheless, even with these data we can still only identify candidate barrier regions: experimental and functional approaches are necessary to identify barrier loci unequivocally.

Step 6: Independent evidence for barrier loci

Crucially, genomic data alone cannot provide conclusive evidence of barrier loci. Disentangling effects is difficult precisely because some modifying factors (e.g. demographic history) are estimated from data used to measure the landscape of differentiation. Even with good genomic evidence of selection on a candidate region, other processes, such as local adaptation following or unrelated to speciation, can be invoked (Cruickshank & Hahn, 2014). For this reason, the search for evidence of selection should extend beyond the genome scan. In principle, there are two ways of obtaining independent evidence for selection; we can either directly test for signatures of selection on a given locus; or we can test for a link between the genotype and the phenotype, and separately test for selection on the phenotype (Table 2). The advantage of the former is that it provides a more direct test of selection; the advantage of the latter is that knowing the associated phenotypic change allows for a complete "story" and a better understanding of the system.

Selection experiments in the field or laboratory, followed by genome-wide or candidate locus sequencing, are an excellent example of the former approach (Soria-Carrasco *et al.*, 2014; Egan *et al.*, 2015). Although not possible in all organisms, such studies have already identified loci involved in reproductive isolation and adaptive divergence (Colosimo *et al.*, 2005; Barrett *et al.*, 2008; Arnegard *et al.*, 2014). Genomic data beyond the binary sampling often used for outlier scans can also be very helpful to collect independent evidence of selection. For example, barrier loci are expected to show steep allele frequency clines in regions where gene flow is occurring (Box 2). Data from instances of parallel divergence may also be used to test whether the same genomic regions show differentiation repeatedly (although see caveats described in Step 5; Table 2).

Various approaches have been used in order to test associations of candidate loci with divergent phenotypes (or, ideally, phenotypes for which tests of divergent selection have been performed), including QTL crossing experiments, association and admixture mapping. Combining mapping with genome scan data can help identify when QTL coincide with outlier loci and also provides further evidence that these loci are under selection in the wild (Via & West, 2008; Renaut *et al.*, 2010; Berner *et al.*, 2014). Differences in gene expression between populations at candidate genes under divergent selection might also be informative (Poelstra *et al.*, 2014). In systems where decent genome annotation exists, this may identify associations between candidate loci and known divergent traits (Lamichhaney *et al.*, 2015, 2016).

Nonetheless, the majority of these approaches stop short of directly demonstrating how a barrier allele alters the function to produce phenotypic consequences and ultimately results in reproductive isolation (Seehausen *et al.*, 2014). In some cases, molecular assays of protein function are possible; but often conclusive evidence is only really possible using transgenic or gene interference methods which to-date have largely been limited to model organisms such as *Drosophila* (Thomae *et al.*, 2013; Satyaki *et al.*, 2014; Phadnis *et al.*, 2015). With the rapid adoption of CRISPR, a method applicable to a much wider range of organisms, transgenic experiments are likely to become an important part of speciation research (Bono *et al.*, 2015). Gene insertion, knockouts and reciprocal transplant experiments, for example, will be able to provide direct evidence of barrier nucleotide function in non-model organisms (Bono *et al.*, 2015).

Concluding remarks

The genomic landscape of speciation is, like the process itself, complex. A wide variety of processes and mechanisms can shape differentiation and divergence between species pairs, beyond divergent selection and gene flow. Like a true physical landscape, determining which processes have played an important role in its formation is difficult but not insurmountable. Accounting for modifying factors in genome scan data will undoubtedly require sophisticated approaches but will also need additional evidence such as independent measures of recombination and mutation rate variation, and, maybe most importantly, independent evidence for selection (e.g. from experiments). The field of speciation genomics is already progressing towards disentangling modifying factors and directly measuring selection on candidate loci in the field and the lab, with a greater emphasis on experimental design and new analysis methods. Furthermore, with new molecular tools and more advanced sequencing technologies on the horizon, conclusive evidence for barrier loci will likely become easier to achieve for those working outside the realm of model species. We look forward to further developing our understanding of how genomic heterogeneity evolves and how this understanding can used to identify loci involved in reproductive isolation with greater precision and reliability.

Box 1 – Clearer definitions

A wealth of technical terms, often without clear definition, makes an attempt to understand the literature on speciation genomics a daunting task (Harrison, 2012). In this review, we argue for the importance of identifying **barrier loci**, positions in the genome that contribute to barriers to gene flow between populations. These include loci under divergent ecological selection, but also loci involved in other barriers, e.g. mate choice, or intrinsic postzygotic isolation. When a **locally beneficial allele**, adaptive in a single population, arises, divergent positive selection will cause it to increase in frequency, resulting in a **local selective sweep. The barrier effect** is a reduction of effective migration rate relative to actual migration between populations that occurs at the barrier locus (i.e. the direct effect) but can also extend beyond it (i.e. the indirect effect). In surrounding genomic regions, the barrier effect will initially allow a build-up of **genomic differentiation**, i.e. a difference in allele frequency, between populations, typically documented using a relative measure such as F_{ST}. Over time, the barrier effect will allow neutral mutations to establish, resulting in **genomic divergence** between populations, typically measured using d_{XY}. In contrast to barrier loci, when **globally beneficial alleles** arise they will increase in frequency due to positive (but crucially, not divergent) selection and spread amongst populations in contact. Both globally and locally adaptive alleles may undergo **hard sweeps**, i.e. from de novo mutation or introgression, or **soft sweeps**, i.e. from standing genetic variation. **Genome scans**, comparisons between pairs of populations or species at multiple loci across the genome (typically thousands of loci nowadays), can quantify the genomic landscape of differentiation and divergence when placed on a physical or genetic map. These are used to identify **outliers**, *i.e.* loci or regions that fall outside the expected equilibrium neutral distribution of differentiation or divergence, which may be influenced by barrier effects.

Box 2 – Searching for islands in a sea of metaphors

Genomic differentiation may be heterogeneous during much of the speciation process (Nosil 2012; Table 1). Under the genic view of speciation, the genome is porous to gene flow while reproductive isolation is incomplete (Wu, 2001; Wu & Ting, 2004). A large number of genome scans have identified distinct genomic regions ("islands") of greater differentiation than the putatively neutral genomic background ("sea level) that tends toward homogenization by gene flow (Nosil *et al.*, 2009). First described as "genomic islands of speciation" in *Anopheles* mosquitoes, these regions were assumed to harbour loci underlying reproductive isolation (Turner *et al.*, 2005). The genomic island metaphor has proved popular and has been valuable for driving empirical progress; a wide array of studies searching for "speciation islands" in multiple taxa has been published in the last decade.

Other terms have also been coined to describe genomic heterogeneity. These may not explicitly invoke speciation, e.g. "genomic islands of differentiation" (Harr, 2006) or "genomic islands of divergence" (Nosil *et al.*, 2009). Large differentiation regions, potentially containing multiple speciation genes have been referred to as "continents of divergence" (Michel *et al.*, 2010; Egan *et al.*, 2015). These metaphors have led to conceptual frameworks, such as Feder et al.'s (2012) four-phase model, which incorporate processes such as divergence and genome hitchhiking (see main text) to explain how differentiation across the genome evolves. Although the metaphors have proved useful for describing observed patterns and communicating a complex concept to a wider audience, introducing attractive terminology runs the risk of encouraging ambiguity (Harrison, 2012). For example, differentiation is more likely to vary continuously during speciation rather than showing clearly defined "islands" or "continents". Metaphors also lead to arbitrary and unproductive discussions on how to define them: what level of differentiation defines an island and when or at what length does an "island" become a "continent"?

Although genomic regions of high differentiation undoubtedly exist (Table 1), they are not necessarily caused by the interplay between gene flow and divergent selection; they may in fact be "incidental islands" that emerge when gene flow is absent (Noor & Bennett, 2009; Turner & Hahn, 2010; Cruickshank & Hahn, 2014). Divergent and indirect selection (i.e. hitchhiking and background selection) can reduce within-population diversity in geographically isolated and potentially locally adapted populations, leading to high- F_{ST} regions that may not be related to speciation, while much of the genome remains undifferentiated due to incomplete lineage sorting. This process results in a specific genomic signature with high levels of differentiation (F_{ST}, a relative measure) and low levels of absolute divergence (d_{XY}) at loci affected by local adaptation or background selection. In this case, divergence due to direct and indirect selection occurs in the absence of gene flow, potentially after speciation is completed or even just while local adaptation is occurring. It is necessary to rule out this alternative explanation before interpreting regions of elevated differentiation as barrier loci. For that, it is crucial to test for ongoing or recent gene flow (Box 3).

Importantly, even if gene flow does occur, elevated divergence/differentiation alone is not sufficient to identify barrier loci; additional evidence is necessary (see Roadmap). Given that "islands" may not be involved in speciation at all, we suggest avoiding any terminology linking highly differentiated genomic regions to speciation unless further evidence suggests this is, in fact, the case.

Box 3 – Inferring and quantifying gene flow

Barrier effects can only be detected in the presence of recent or ongoing gene flow. Inferring gene flow, outside of a genome scan and preferably in the context of evolutionary history, is an important first step for interpreting the genomic landscape of speciation. However, given the complexity of speciation history and the high probability that, in many cases, gene flow is not constant over time, this presents a major difficulty for speciation research.

Identifying recent gene flow using population clustering methods that reliably detect F1, F2 and backcross hybrids is relatively straightforward (Pritchard *et al.*, 2000; Anderson & Thompson, 2002; Falush *et al.*, 2003). Emphasis should be placed on identifying introgression over several generations: i.e. on the presence of backcrossed individuals. Clinal analysis of allele frequencies across hybrid zones or across the genome overcomes a significant current disadvantage of clustering techniques as it allows for reliable migration estimates (Barton, 1983; Barton & Hewitt, 1985; Gompert & Buerkle, 2011). Other evidence for recent or ongoing gene flow makes use of the biogeographical distributions of species, e.g. asking if genetic differentiation is lower in sympatry compared to allopatry (Noor & Bennett, 2009; Marques *et al.*, 2016). For example, *Heliconius* butterfly studies show greater divergence between allopatric races than between those in sympatry or parapatry, suggesting ongoing gene flow (Nadeau *et al.*, 2012, 2013; Martin *et al.*, 2013). Similarly, very recently diverged populations (i.e. hundreds of generations) with documented hybridization events suggest low genomic differentiation is maintained, at least in part, by gene flow (Lescak *et al.*, 2015; Marques *et al.*, 2016). Finally, non-genetic evidence of migration or potential migration between populations using mark-recapture experiments (Bolnick *et al.*, 2009), mate-choice experiments (Nosil *et al.*, 2002; McKinnon *et al.*, 2004) and phenotypic variation (Lescak *et al.*, 2015) can bolster the argument that low background differentiation in a genome scan is due to ongoing gene flow.

Several key approaches incorporate demographic history, making it possible to infer both gene flow and mechanisms of divergence (Sousa & Hey, 2013). Site frequency spectrum (SFS) methods can rapidly approximate the joint allele frequency distribution between populations, allowing comparisons of divergence with and without gene flow and the estimation of migration rate (Gutenkunst *et al.*, 2009; Excoffier *et al.*, 2013). Isolation-with-Migration (IM) models have also recently been extended to incorporate whole-genome data and overcome some simplifying assumptions such as absence of recombination (Hobolth *et al.*, 2011a; Mailund *et al.*, 2012). Approximate Bayesian Computation (ABC) is more computationally expensive but can incorporate thousands of loci resulting in high precision parameter estimation (Robinson *et al.*, 2014; Shafer *et al.*, 2015). ABC is flexible, allowing variation in migration rates amongst loci to be incorporated (Roux *et al.*, 2013, 2014) or the inclusion of haplotype-based statistics for estimating gene flow (Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010). However, we note that model-based inference is limited to distinguishing amongst the models tested; parameter estimates are therefore meaningful only in the context of a specific model. Since these are generally simplifications, the results must be interpreted with caution.

Modelling approaches typically perform poorly when estimating gene flow timing (Roux *et al.*, 2013), but this may be possible to overcome when there is sufficient biogeographical and phylogenetic information to resolve periods of contact between populations (Garrigan *et al.*, 2012; Nadachowska-Brzyska *et al.*, 2013). This is the rationale behind comparative statistics such as ABBA-BABA that test for an excess of derived alleles at positions across the genome (Green *et al.*, 2010; Durand *et al.*, 2011; Martin *et al.*, 2014). By incorporating different taxa with known divergence times, it is possible to infer the time interval when introgression may have occurred (Martin *et al.*, 2013; Eaton *et al.*, 2015). Methods comparing the size of introgressed haplotypes ('migrant tracts') to an expected distribution under migration within *T* generations may provide relatively accurate estimates of the timing of gene flow (Pool & Nielsen, 2009). However, this requires accurate haplotype phasing and has very little power to date admixture more than 1000 generations in the past (Pool & Nielsen, 2009; Liang & Nielsen, 2014). Identity-by-state tracts, i.e. the distance between polymorphisms on a haplotype, also require phased data to calculate but provide a promising means for estimating the timing and extent of gene flow, as well as other demographic parameters (Harris & Nielsen, 2013). An extension of the Markov coalescent approach for estimating effective population size as a function of time can now use haplotype data from multiple individuals to determine cross-coalescence rate (i.e. coalescent events within and between populations) providing accurate estimates of the timing and rate of last migration without a specified demographic model (Li & Durbin, 2011; Schiffels & Durbin, 2014).

Acknowledgments

Many of the ideas for this review were first formulated from discussions between co-authors during our symposium on "The genomic landscape of speciation" at ESEB 2015, Lausanne, Switzerland. We were kindly sponsored by Floragenex, Oregon, USA, and also by Stab Vida, Portugal. Mark Ravinet was funded by a JSPS Postdoctral Fellowship for Foreign Researchers. RF was funded by FCT under the Programa Operacional Potencial Humano – Quadro de Referência Estratégico Nacional from the European Social Fund and the Portuguese Ministério da Educação e Ciência through the postdoctoral fellowship SFRH/BPD/89313/2012 and research project PTDC/BIA-EVF/113805/2009 and FCOMP-01-0124-FEDER-014272. JG was funded by a postdoctoral fellowship from Xunta de Galicia (Modalidade B). AMW and RKB are funded by NERC.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., *et al.* 2013. Hybridization and speciation. *J. Evol. Biol.* **26**: 229–46.

Aeschbacher, S., Selby, J.P., Willis, J.H. & Coop, G. 2016. Population-genomic inference of the strength and timing of selection against gene flow. *bioRxiv* 1–21.

Al-Shahrour, F., Minguez, P., Marques-Bonet, T., Gazave, E., Navarro, A. & Dopazo, J. 2010. Selection upon genome architecture: Conservation of functional neighborhoods with changing genes. *PLoS Comput. Biol.* **6**.

Anderson, E.C. & Thompson, E.A. 2002. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**: 1217–1229.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. & Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **in press**. Nature Publishing Group.

Arnegard, M.E., McGee, M.D., Matthews, B., Marchinko, K.B., Conte, G.L., Kabir, S., *et al.* 2014. Genetics of ecological divergence during speciation. *Nature*, doi: 10.1038/nature13301. Nature Publishing Group.

Barrett, R.D.H. & Hoekstra, H.E. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**: 767–80.

Barrett, R.D.H., Rogers, S.M. & Schluter, D. 2008. Natural selection on a major armor gene in threespine stickleback. *Science (80-.).* **322**: 255–257.

Barton, N. & Bengtsson, B.O. 1986. The barrier to genetic exchange between hybridising populations. *Heredity (Edinb).* **57**: 357–376.

Barton, N. & Hewitt, G. 1985. Analysis of Hybrid Zones. *Annu. Rev. Ecol. Syst.* **16**: 113–148.

Barton, N.H. 2006. Evolutionary Biology: How did the human species form? *Curr. Biol.* **16**: 648–650.

Barton, N.H. 1983. Multilocus Clines. *Evolution (N. Y).* **37**: 454–471.

Barton, N.H. & de Cara, M.A.R. 2009. The evolution of strong reproductive isolation. *Evolution* **63**: 1171–90.

Begun, D.J. & Aquadro, C.F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.

Berner, D., Moser, D., Roesti, M., Buescher, H. & Salzburger, W. 2014. Genetic architecture of skeletal evolution in European lake and stream stickleback. *Evolution* **68**: 1792–805.

Bertorelle, G., Benazzo, A. & Mona, S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* **19**: 2609–2625.

Bierne, N. 2010. The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution (N. Y).* **64**: 3254–3272.

Bierne, N., Gagnaire, P. & David, P. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. **59**: 72–86.

Bierne, N., Welch, J., Loire, E., Bonhomme, F. & David, P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* **20**: 2044–72.

Bilandžija, H., Ma, L., Parkhurst, A. & Jeffery, W.R. 2013. A potential benefit of albinism in Astyanax cavefish: Downregulation of the oca2 gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. *PLoS One* **8**: 1–14.

Bolnick, D.I., Snowberg, L.K., Patenia, C., Stutz, W.E., Ingram, T. & Lau, O.L. 2009. Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution (N. Y).* **63–8**: 2004–2016.

Bono, J.M., Olesnicky, E.C. & Matzkin, L.M. 2015. Connecting genotypes, phenotypes and fitness: Harnessing the power of CRISPR/Cas9 genome editing. *Mol. Ecol.* **24**: 3810–3822.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., *et al.* 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**: 375–381.

Bullaughey, K., Przeworski, M. & Coop, G. 2008. No effect of recombination on the efficacy of natural selection in primates. *Genome Res.* **18**: 544–554.

Burri, R., Nater, A., Kawakami, T., Mugal, C.F., Olason, P.I., Smeds, L., *et al.* 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res.* **25**: 1656–1665.

Butlin, R. 1987. Speciation by reinforcement. *Trends Ecol. Evol.* **2**: 8–13.

Butlin, R.K. 2005. Recombination and speciation. *Mol. Ecol.* **14**: 2621–2635.

Butlin, R.K., Galindo, J. & Grahame, J.W. 2008. Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philos. Trans. R. Soc. London Ser. B* **363**: 2997–3007.

Butlin, R.K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., *et al.* 2014. Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution* **68**: 935–49.

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F.W., Alfoldi, J., Barrio, a. M., *et al.* 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science (80-.).* **345**: 1074–1079.

Chan, Y.F., Marks, M.E., Jones, F.C., Villareal, G., Shapiro, M.D., Brady, S.D., *et al.* 2010.

Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (80-.).* **327**: 302–305.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**: 195–205.

Charlesworth, B. 2012. The role of background selection in shaping patterns of molecular evolution and variation: Evidence from variability on the Drosophila X chromosome. *Genetics* **191**: 233–246.

Charlesworth, B. & Campos, J.L. 2014. The relations between recombination rate and patterns of molecular variation and evolution in Drosophila. *Annu. Rev. Genet.* **48**: 383–403.

Charlesworth, B. & Charlesworth, D. 2010. *Elements of Evolutionary Genetics*. Roberts & Company Publishers, Greenwood Village, Colorado.

Charlesworth, B., Charlesworth, D. & Barton, N.H. 2003. The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Syst.* **34**: 99–125.

Charlesworth, B., Morgan, M.T. & Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–303.

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. *Genet. Res.* **70**: 155–74.

Christe, C., Stölting, K.N., Paris, M., Fraïsse, C., Bierne, N. & Lexer, C. 2016. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol. Ecol.* **in press**.

Colosimo, P.F., Hoseman, K.E., Balabhadra, S., Villareal, G., Dickson, M., Grimwood, J., *et al.* 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science (80-.).* **307**: 1928–1933.

Comeron, J.M. 2014. Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. *PLoS Genet.* **10**.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., *et al.* 2013. Multiplex Genome Engineering Using CRISPR/Cas System. *Science (80-.).* **339**: 819–824.

Corbett-Detig, R., Zhou, J. & Clark, A. 2013. Genetic incompatabilities are widespread within species. *Nature* **504**: 135–137. Nature Publishing Group.

Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sinaeur, New York.

Cruickshank, T.E. & Hahn, M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**: 3133–3157.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O., Csillery, K. & Francois, O. 2010.

Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25**: 410–418.

Cutter, A.D. & Payseur, B. a. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**: 262–274. Nature Publishing Group.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L. 2013. Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* **22**: 3151–64.

Davey, J.W., Hohenlohe, P. a, Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**: 499–510. Nature Publishing Group.

Delmore, K.E., Hübner, S., Kane, N.C., Schuster, R., Andrew, R.L., Câmara, F., *et al.* 2015. Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Mol. Ecol.* **24**: 1873–1888.

Drummond, A.J., Ho, S.Y.W., Phillips, M. & Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *Plos Biol.* **4**: e88.

Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. 2011. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**: 2239–52.

Duret, L. & Arndt, P.F. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4**.

Dutheil, J.Y. & Hobolth, A. 2012. Ancestral Population Genomics. In: *Evolutionary Genomics: Statistical and Computational Methods. Volume 2* (M. Anismova, ed). Springer, New York, USA.

Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A. & Cavender-Bares, J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution (N. Y).* **69**: 2587–2601.

Egan, S.P., Ragland, G.J., Assour, L., Powell, T.H.Q., Hood, G.R., Emrich, S., *et al.* 2015. Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecol. Lett.* **18**: 817–825.

Elyashiv, E., Sattath, S., Hu, T.T., Strustovsky, A., McVicker, G., Andolfatto, P., *et al.* 2016. A genomic map of the effects of linked selection in Drosophila. *PLoS Genet.* 1–38.

Enard, D., Messer, P.W. & Petrov, D. a. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24**: 885–95.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., *et al.* 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7**: e47768.

Ewing, G.B. & Jensen, J.D. 2016. The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* **25**: 135–141.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**: e1003905.

Eyre-Walker, A. & Keightley, P.D. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet.* **8**: 610–8.

Falush, D., Stephens, M. & Pritchard, J.K. 2003. Inference of population structure: Extensions to linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Faria, R. & Navarro, A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends Ecol. Evol.* **25**: 660–9. Elsevier Ltd.

Feder, J.L., Egan, S.P. & Nosil, P. 2012a. The genomics of speciation-with-gene-flow. *Trends Genet.* 1–9. Elsevier Ltd.

Feder, J.L., Flaxman, S.M., Egan, S.P., Comeault, A. a. & Nosil, P. 2013. Geographic Mode of Speciation and Genomic Divergence. *Annu. Rev. Ecol. Evol. Syst.* **44**: 73–97.

Feder, J.L., Gejji, R., Yeaman, S. & Nosil, P. 2012b. Establishment of new mutations under divergence and genome hitchhiking. *Philos. Trans. R. Soc. B Biol. Sci.* **367**: 461–474.

Feder, J.L. & Nosil, P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution (N. Y).* **64–5**: 1729–1747.

Ferchaud, A.L. & Hansen, M.M. 2016. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: Three-spine sticklebacks in divergent environments. *Mol. Ecol.* **25**: 238–259.

Feulner, P.G.D., Chain, F.J.J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., *et al.* 2015. Genomics of Divergence along a Continuum of Parapatric Population Differentiation. *PLOS Genet.* **11**: e1004966.

Flaxman, S.M., Wacholder, A.C., Feder, J.L. & Nosil, P. 2014. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* 4074–4088.

Flowers, J.M., Molina, J., Rubinstein, S., Huang, P., Schaal, B.A. & Purugganan, M.D. 2012. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol. Biol. Evol.* **29**: 675–687.

Foll, M. & Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–93.

Fraïsse, C., Belkhir, K., Welch, J.J. & Bierne, N. 2016a. Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* **25**: 269–286.

Fraïsse, C., Belkhir, K., Welch, J.J. & Bierne, N. 2016b. Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* **25**: 269–286.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., *et al.* 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**: 822–826. Nature Publishing Group.

Gagnaire, P.A., Pavey, S.A., Normandeau, E. & Bernatchez, L. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing. *Evolution (N. Y).* **67**: 2483–2497.

Garrigan, D., Kingan, S.B., Geneva, A.J., Andolfatto, P., Clark, A.G., Thornton, K.R., *et al.* 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res.* **22**: 1499–511.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., *et al.* 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 3165–3178.

Geneva, A.J., Kingan, S.B. & Garrigan, D. 2015. A new method to scan genomes for introgression in a secondary contact model. *PLoS One* 1–15.

Glazer, A.M., Killingbeck, E.E., Mitros, T., Rokhsar, D.S. & Miller, C.T. 2015. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3* **5**: 1463–72.

Gompert, Z. & Buerkle, C.A. 2011. Bayesian estimation of genomic clines. *Mol. Ecol.* **20**: 2111–27.

Gompert, Z., Comeault, A.A., Farkas, T.E., Feder, J.L., Parchman, T.L., Buerkle, C.A., *et al.* 2014. Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17**: 369–379.

Gossmann, T.I., Santure, A.W., Sheldon, B.C., Slate, J. & Zeng, K. 2014. Highly variable recombinational landscape modulates efficacy of natural selection in birds. *Genome Biol. Evol.* **6**: 2061–2075.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., *et al.* 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–22.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**: e1000695.

Haddrill, P.R., Halligan, D.L., Tomaras, D. & Charlesworth, B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol.* **8**: R18.

Haldane, J.B.S. 1930. A mathematical theory of natural and artificial selection-I. *Math. Proc. Cambridge Philos. Soc.* **26**: 220–230.

Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**: 730–737.

Harris, K. & Nielsen, R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**: e1003521.

Harrison, R.G. 1990. Hybrid zones: windows on evolutionary processes. In: *Oxford surveys in evolutionary biology*, pp. 69–128.

Harrison, R.G. 2012. *The language of speciation*.

Harrison, R.G. & Larson, E.L. 2016. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* **25**: 2454–2466.

Hartl, D.L. & Clark, A.G. 2007. *Principles of population genetics*. Sinauer Associates, Sunderland, Massachusetts, USA.

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V., *et al.* 2010. The genome of the Western clawed frog *Xenopus tropicalis*. *Science* **328**: 633–6.

Hermisson, J. & Pennings, P.S. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–52.

Hey, J. & Kliman, R.M. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* **160**: 595–608.

Ho, S.Y.W. 2014. The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.* **29**: 496–503. Elsevier Ltd.

Ho, S.Y.W. & Larson, G. 2006. Molecular clocks: when times are a-changin'. *Trends Ecol. Evol.* **22**: 79–83.

Ho, S.Y.W., Phillips, M.J., Cooper, A. & Drummond, A.J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* **22**: 1561–1568.

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., *et al.* 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *Am. Nat.* **188**: 000–000.

Hobolth, A., Andersen, L.N. & Mailund, T. 2011a. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics* **187**: 1241–3.

Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H. & Mailund, T. 2011b. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**: 349–56.

Hodgkinson, A. & Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**: 756–766. Nature Publishing Group.

Hoekstra, H.E., Hirschmann, R.J., Bundey, R.A., Insel, P. a & Crossland, J.P. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science (80-.).* **313**: 101–104.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *PloS Genet.* **6**: e1000862.

Huber, C.D., DeGiorgio, M., Hellmann, I. & Nielsen, R. 2016. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol. Ecol.* **25**: 142–156.

Hurst, L.D., Pál, C. & Lercher, M.J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**: 299–310.

Hurst, L.D., Williams, E.J.B. & Pal, C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* **18**: 604–606.

Irwin, D.E., Alcaide, M., Delmore, K.E., Irwin, J.H. & Owens, G.L. 2016. Recurrent selection explains genomic regions of high relative but low absolute differentiation in the greenish warbler ring species. *Mol. Ecol.* **25**: 4488–4507.

Jensen-Seaman, M. & Furey, T. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 528–538.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., *et al.* 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

Kawakami, T., Smeds, L., Backstrom, N., Husby, A., Qvarnstrom, A., Mugal, C.F., *et al.* 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Mol. Ecol.* **23**: 4035–4058.

Keller, I., Wagner, C.E., Greuter, L., Mwaiko, S., Selz, O.M., Sivasundar, a, *et al.* 2012. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol. Ecol.* 2848–2863.

Kim, N. & Jinks-Robertson, S. 2012. Transcription as a source of genome instability. *Nat. Rev. Genet.* **13**: 204–214. Nature Publishing Group.

Kirkpatrick, M. & Barton, N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* **173**: 419–34.

Kitano, J., Ross, J.A., Mori, S., Kume, M., Jones, F.C., Chan, Y.F., *et al.* 2009. A role for neo-sex chromosomes in stickleback speciation. *Nature* **461**: 1079–1083.

Knowles, L.L. 2009. Statistical Phylogeography. *Annu. Rev. Ecol. Syst.* **40**: 593–612.

Knowles, L.L. & Maddison, W.P. 2002. Statistical phylogeography. *Mol. Ecol.* **11**: 2623–2635.

Kondrashov, F.A. & Kondrashov, A.S. 2010. Measurements of spontaneous rates of mutations in the recent past and the near future. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365**: 1169–76.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., *et al.* 2002. 2002:A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241–7.

Kruuk, L.E., Baird, S.J., Gale, K.S. & Barton, N.H. 1999. A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics* **153**: 1959–71.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., *et al.* 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**: 371–375.

Lamichhaney, S., Han, F., Berglund, J., Wang, C., Sallman, A.M., Webster, M.T., *et al.* 2016. A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science (80-.).* **6284**: 470–474.

Leaché, A.D., Harris, R.B., Maliska, M.E. & Linkem, C.W. 2013. Comparative species divergence across eight triplets of spiny lizards (Sceloporus) using genomic sequence data. *Genome Biol. Evol.* **5**: 2410–9.

Le Moan, A., Gagnaire, P. ‐A. & Bonhomme, F. 2016. Parallel genetic divergence among coastal‐marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Mol. Ecol.*, doi: 10.1111/mec.13627.

Lescak, E.A., Bassham, S.L., Catchen, J., Gelmond, O., Sherbick, M.L., von Hippel, F.A., *et al.* 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proc. Natl. Acad. Sci. U. S. A.* 201512020.

Lewontin, R. & Krakauer, J. 1973. DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS. *Genetics* 175–195.

Li, H. & Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–6.

Liang, M. & Nielsen, R. 2014. The lengths of admixture tracts. *Genetics* **197**: 953–967.

Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., *et al.* 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* **7**.

Lotterhos, K.E. & Whitlock, M.C. 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* **23**: 2178–2192.

Luttikhuizen, P.C., Drent, J., Peijnenburg, K.T.C. a, van der Veer, H.W. & Johannesson, K. 2012. Genetic architecture in a marine hybrid zone: comparing outlier detection and genomic clines analysis in the bivalve Macoma balthica. *Mol. Ecol.* **21**: 3048–61.

Maddison, W.P. 1997. Gene trees in species trees. *Syst. Biol.* **46**: 523–536.

Mailund, T., Halager, A.E., Westergaard, M., Dutheil, J.Y., Munch, K., Andersen, L.N., *et al.* 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet.* **8**: e1003125.

Mallet, J., Besansky, N. & Hahn, M.W. 2016. How reticulated are species? *BioEssays* **38**: 140–149.

Marques, D.A., Lucek, K., Meier, J.I., Mwaiko, S., Wagner, C.E., Excoffier, L., *et al.* 2016. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLOS Genet.* **12**: e1005887.

Martin, C.H., Cutler, J.S., Friel, J.P., Dening Touokong, C., Coop, G. & Wainwright, P.C. 2015a. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution (N. Y).* **69**: 1406–1422.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., *et al.* 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23**: 1817–28.

Martin, S.H., Davey, J.W. & Jiggins, C.D. 2014. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**: 244–257.

Martin, S.H., Eriksson, A., Kozak, K.M., Manica, A. & Jiggins, C.D. 2015b. Speciation in Heliconius Butterflies : Minimal Contact Followed by Millions of Generations of Hybridisation. *BioRxiv* 1–24.

Massy, B. De. 2013. Initiation of Meiotic Recombination : How and Where ? Conservation and Specificities Among Eukaryotes. *Annu. Rev. Genet.* **47**: 563–599.

McGaugh, S.E. & Noor, M.A.F. 2012. Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 422–9.

McGee, M.D., Neches, R.Y. & Seehausen, O. 2016. Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. *Mol. Ecol.* **25**: 260–268.

McKinnon, J.S., Mori, S., Blackman, B.K., David, L., Kingsley, D.M., Jamieson, L., *et al.* 2004. Evidence for ecology's role in speciation. *Nature* **429**: 294–298.

McKinnon, J.S. & Rundle, H.D. 2002. Speciation in nature: the threespine stickleback model systems. *Trends Ecol. Evol.* **17**: 480–481.

Messer, P.W. & Petrov, D. a. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**: 659–69. Elsevier Ltd.

Michel, A., Sim, S., Powell, T.H.Q., Taylor, M.S., Nosil, P. & Feder, J.L. 2010. Widespread genomic divergence during sympatric speciation. *Proc. …* **107**: 9724–9729.

Myers, S., Bowden, R., Tumian, A., Bontrop, R., Freeman, C., MacFie, T., *et al.* 2010. Drive against hotspot motifs in primates implicates the PRDM9 Gene in Meiotic Recombination. *Science (80-.).* 876–879.

Nachman, M.W. & Payseur, B. a. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 409–21.

Nadachowska-Brzyska, K., Burri, R., Olason, P.I., Kawakami, T., Smeds, L. & Ellegren, H. 2013. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data. *PLoS Genet.* **9**: e1003942.

Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., *et al.* 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol. Ecol.* **22**: 814–826.

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., *et al.* 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 343–53.

Neafsey, D.E., Lawniczak, M.K.N., Park, D.J., Redmond, S.N., Coulibaly, M.B., Traoré, S.F., *et al.* 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science* **330**: 514–7.

Ness, R.W., Morgan, A.D., Vasanthakrishnan, R.B., Colegrave, N. & Keightley, P.D. 2015. Extensive de novo mutation rate variation between individuals and across the genome of Chlamydomonas reinhardtii. *Genome Res.* **25**: 1739–1749.

Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., *et al.* 2015. A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. *Nat. Genet.* **47**: 1–7. Nature Publishing Group.

Nobrega, M.A. 2003. Scanning Human Gene Deserts for Long-Range Enhancers. *Science (80-.).* **302**: 413–413.

Noor, M.A.F. 2008. Mutagenesis from meiotic recombination is not a primary driver of sequence divergence between Saccharomyces species. *Mol. Biol. Evol.* **25**: 2439–2444.

Noor, M.A.F. & Bennett, S.M. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity (Edinb).* **103**: 439–444.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., *et al.* 2005. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* **3**: 1289–1299.

Nosil, P. 2013. Degree of sympatry affects reinforcement in drosophila. *Evolution (N. Y).* **67**: 868–872.

Nosil, P. 2012. *Ecological Speciation*. Oxford University Press, Oxford, UK.

Nosil, P., Crespi, B.J. & Sandoval, C.P. 2002. Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature* **417**: 440–443.

Nosil, P. & Feder, J.L. 2012a. Genomic divergence during speciation: causes and consequences. *Philos. Trans. R. Soc. London Ser. B* **367**: 332–342.

Nosil, P. & Feder, J.L. 2012b. Widespread yet heterogeneous genomic divergence. *Mol. Ecol.* **21**: 2829–32.

Nosil, P., Funk, D.J. & Ortiz-Barrientos, D. 2009. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**: 375–402.

Nosil, P., Gompert, Z., Farkas, T.E., Comeault, a. a., Feder, J.L., Buerkle, C. a., *et al.* 2012. Genomic consequences of multiple speciation processes in a stick insect. *Proc. R. Soc. B Biol. Sci.* 5058–5065.

Nosil, P., Harmon, L.J. & Seehausen, O. 2008. Ecological explanations for (incomplete) speciation. *Trends Ecol. Evol.* **24**: 145–156.

Ohta, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annu. Rev. Ecol. Syst.* **23**: 263–286.

Orr, H.A. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution (N. Y).* **52**: 935–949.

Ortiz-Barrientos, D., Engelstädter, J. & Rieseberg, L.H. 2016. Recombination Rate Evolution and the Origin of Species. *Trends Ecol. Evol.* **31**: 226–236. Elsevier Ltd.

Pease, J.B. & Hahn, M.W. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution* **67**: 2376–84.

Pennings, P.S. & Hermisson, J. 2006. Soft sweeps II - Molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**: 1076–1084.

Phadnis, N., Baker, E.P., Cooper, J.C., Frizzell, K.A., Hsieh, E., de la Cruz, A.F.A., *et al.* 2015. An essential cell cycle regulation gene causes hybrid inviability in Drosophila. *Science* **350**: 1552–5.

Phadnis, N. & Orr, H.A. 2009. Sterility and Segregation Distortion in Drosophila Hybrids. **323**: 376–379.

Phifer-Rixey, M., Bomhoff, M. & Nachman, M.W. 2014. Genome-wide patterns of differentiation among house mouse subspecies. *Genetics* **198**: 283–297.

Phung, T.N., Huber, C.D. & Lohmueller, K.E. 2016. Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. *PLoS Genet* **12**: 1–27.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., *et al.* 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science (80-.).* **344**: 1410–4.

Pool, J.E. & Nielsen, R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**: 711–9.

Presgraves, D.C. 2007. Speciation genetics: epistasis, conflict and the origin of species. *Curr. Biol.* **17**: R125-7.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160**: 1179–89.

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C. & Panova, M. 2016. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. *Mol. Ecol.* **25**: 287–305.

Renaut, S., Grassa, C.J., Yeaman, S., Moyers, B.T., Lai, Z., Kane, N.C., *et al.* 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.* **4**: 1827. Nature Publishing Group.

Renaut, S., Maillet, N., Normandeau, E., Sauvage, C., Derome, N., Rogers, S.M., *et al.* 2012. Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philos. Trans. R. Soc. B* **367**: 354–363.

Renaut, S., Nolte, A.W., Rogers, S.M., Derome, N. & Bernatchez, L. 2010. SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). *Mol. Ecol.* **20**: 545–559.

Renaut, S., Owens, G.L. & Rieseberg, L.H. 2014. Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Mol. Ecol.* **23**: 311–324.

Rice, W. & Hostert, E.E. 1993. LABORATORY EXPERIMENTS ON SPECIATION - WHAT HAVE WE LEARNED IN 40 YEARS. *Evol. Heal. Dis.* **47**: 1637–1653.

Robinson, J.D., Bunnefeld, L., Hearn, J., Stone, G.N. & Hickerson, M.J. 2014. ABC inference of multi-population divergence with admixture from un-phased population genomic data. *Mol. Ecol.* **23**: 4458–4471.

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**: 1–17.

Roda, F., Ambrose, L., Walter, G.M., Liu, H.L., Schaul, A., Lowe, A., *et al.* 2013. Genomic evidence for the parallel evolution of coastal forms in the Senecio lautus complex. *Mol. Ecol.* **22**: 2941–52.

Roesti, M., Gavrilets, S., Hendry, A.P., Salzburger, W. & Berner, D. 2014. The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**: 3944–3956.

Roesti, M., Hendry, A.P., Salzburger, W. & Berner, D. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Mol. Ecol.* **21**: 2852–2862.

Roesti, M., Kueng, B., Moser, D. & Berner, D. 2015. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **41**: 1–14. Nature Publishing Group.

Roesti, M., Moser, D. & Berner, D. 2013. Recombination in the threespine stickleback genome - Patterns and consequences. *Mol. Ecol.* **22**: 3014–3027.

Rogers, S.M. & Bernatchez, L. 2005. Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (Coregonus clupeaformis). *Mol. Ecol.* **14**: 351–361.

Rosenzweig, B.K., Pease, J.B., Besansky, N.J. & Hahn, M.W. 2016. Powerful methods for detecting introgressed regions from population genomic data. *Mol. Ecol.* 2387–2397.

Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S., *et al.* 2016. Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and non-parasitic lamprey ecotypes. *Mol. Ecol.*, doi: 10.1111/mec.13664.

Roux, C., Fraïsse, C., Castric, V., Vekemans, X., Pogson, G.H. & Bierne, N. 2014. Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone. *J. Evol. Biol.* **27**: 1662–1675.

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N. & Bierne, N. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *bioRxiv* 513–

516.

Roux, C., Tsagkogeorga, G., Bierne, N. & Galtier, N. 2013. Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. *Mol. Biol. Evol.* **30**: 1574–87.

Satyaki, P.R. V, Cuykendall, T.N., Wei, K.H.-C., Brideau, N.J., Kwak, H., Aruna, S., *et al.* 2014. The Hmr and Lhr hybrid incompatibility genes suppress a broad range of heterochromatic repeats. *PLoS Genet.* **10**: e1004240.

Scally, A. & Durbin, R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**: 745–53.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., *et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175. Nature Publishing Group.

Schiffels, S. & Durbin, R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**: 919–925. Nature Publishing Group.

Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature Publishing Group.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P. a, *et al.* 2014. Genomics and the origin of species. *Nat. Rev. Genet.* **15**: 176–92. Nature Publishing Group.

Sémon, M. & Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* **23**: 1715–1723.

Servedio, M.R. & Noor, M. 2003. The role of reinforment in speciation: theory and data. *Annu. Rev. Ecol. Evol. Syst.* **34**: 339–364.

Servedio, M.R., Van Doorn, G.S., Kopp, M., Frame, A.M. & Nosil, P. 2011. Magic traits in speciation: “magic” but not rare? *Trends Ecol. Evol.* **26**: 389–97.

Shafer, A.B. a., Gattepaille, L.M., Stewart, R.E. a. & Wolf, J.B.W. 2015. Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. *Mol. Ecol.* 328–345.

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., *et al.* 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**: 717–723.

Shuker, D.M., Underwood, K., King, T.M. & Butlin, R.K. 2005. Patterns of male sterility in a grasshopper hybrid zone imply accumulation of hybrid incompatibilities without selection. *Proc. Biol. Sci.* **272**: 2491–7.

Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science (80-.).* **236**: 787–792.

Slatkin, M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res.* **58**: 167–175. Universitetsbiblioteket i Oslo.

Slatkin, M., Reviews, A. & Review, A. 1985. Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* **16**: 393–430.

Smadja, C.M. & Butlin, R.K. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Mol. Ecol.* **Early view**: 5123–40.

Smukowski, C.S. & Noor, M.A.F. 2011. Recombination rate variation in closely related species. *Heredity (Edinb).* **107**: 496–508. Nature Publishing Group.

Soria-Carrasco, V., Gompert, Z., Comeault, a. a., Farkas, T.E., Parchman, T.L., Johnston, J.S., *et al.* 2014. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science (80-.).* **344**: 738–742.

Sousa, V. & Hey, J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* **14**: 404–14. Nature Publishing Group.

Stephan, W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. *Philos. Trans. R. Soc. London Ser. B-Biological Sci.* **365**: 1245–1253.

Stephan, W. 2016. Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.* **25**: 79–88.

Storz, J.F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* **14**: 671–688.

Storz, J.F. & Wheat, C.W. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution (N. Y).* **64**: 2489–2509.

Terai, Y., Seehausen, O., Sasaki, T., Takahashi, K., Mizoiri, S., Sugawara, T., *et al.* 2006. Divergent Selection on Opsins Drives Incipient Speciation in Lake Victoria Cichlids. *PLOS Biol* **4**: e433.

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–8. Nature Publishing Group.

Thomae, A.W., Schade, G.O.M., Padeken, J., Borath, M., Vetter, I., Kremmer, E., *et al.* 2013. A pair of centromeric proteins mediates reproductive isolation in Drosophila species. *Dev. Cell* **27**: 412–24. Elsevier Inc.

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R.S.T., *et al.* 2014.

European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat. Commun.* **5**: 5770.

Tittes, S. & Kane, N.C. 2014. The genomics of adaptation, divergence and speciation: A congealing theory. *Mol. Ecol.* **23**: 3938–3940.

Trier, C.N., Hermansen, J.S., Sætre, G.-P. & Bailey, R.I. 2014. Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. *PLoS Genet.* **10**: e1004075.

Turner, T.L. & Hahn, M.W. 2010. Genomic islands of speciation or genomic islands and speciation? *Mol. Ecol.* **19**: 848–50.

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

Twyford, A.D. & Friedman, J. 2015. Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion. *Evolution* 1–33.

Via, S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos. Trans. R. Soc. London Ser. B* **366**: In press.

Via, S. & West, J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.* **17**: 4334–4345.

Vines, T.H., Köhler, S.C., Thiel, M., Ghira, I., Sands, T.R., MacCallum, C.J., *et al.* 2003. The maintenance of reproductive isolation in a mosaic hybrid zone between the fire-bellied toads Bombina bombina and B. variegata. *Evolution* **57**: 1876–1888.

Wang, M., Beck, C.R., English, A.C., Meng, Q., Buhay, C., Han, Y., *et al.* 2015. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* **16**: 214.

Whitlock, M.C. & McCauley, D.E. 1999. Indirect measures of gene flow and migration: F-ST not equal 1/(4Nm+1). *Heredity (Edinb).* **82**: 117–125.

Wood, B.P. & Miller, J.R. 2006. Linked selected and neutral loci in heterogeneous environments. *J. Math. Biol.* **53**: 939–975.

Wu, C.-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* **14**: 851–865.

Wu, C.-I. & Ting, C.-T. 2004. Genes and speciation. *Nat. Rev. Genet.* **5**: 114–22.

Yeaman, S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U. S. A.* **110**: E1743-51.

Yeaman, S. 2015. Local Adaptation by Alleles of Small Effect. *Am. Nat.* **186**: S74–S89.

Yeaman, S., Aeschbacher, S. & Bürger, R. 2016. The evolution of genomic islands by increased establishment probability of linked alleles. *Mol. Ecol.* 2542–2558.

Yeaman, S. & Otto, S.P. 2011. Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution* **65**: 2123–9.

Yeaman, S. & Whitlock, M.C. 2011. The genetic architecture of adaptation under migration-selection balance. *Evolution* **65**: 1897–911.

Zeng, K. & Corcoran, P. 2015. The effects of background and interference selection on patterns of genetic variation in subdivided populations. *Genetics* **201**: 1539–1554.

Figures

Figure 1: Factors potentially shaping the genomic landscape. Different demographic histories, features of the genome and processes can produce apparently equivalent landscapes of differentiation. During primary divergence, barrier loci and their barrier effects increase differentiation. However, local selective sweeps not related to speciation may also produce peaks of divergence. Also regions of reduced recombination can give rise to such peaks. Under secondary contact, gene flow must eventually erode differentiation that has built up during isolation due to drift and potentially local adaptation. Yet mutation cold-spots may suggest that gene flow has recently occured - when its effect in reality was negligible.

Figure 2. Relative differentiation F_{ST} averaged over 5000 independent evolutionary histories during different speciation scenarios. The figure shows an F_{ST} heatmap (see the colour bar for reference) as a function of time since the start of selection, and as a function of physical distance from the locus under selection in three different scenarios. Primary divergence with a hard sweep (A) and soft sweep (B) and secondary contact with a hard sweep during a period of isolation (C). Solid lines show the frequency of the allele that sweeps through the population where it is beneficial. Parameters: N = 500 individuals per deme. Migration rate per individual, deme, and generation: m = 0.004. Mutation rate per allele, locus, individual, and generation: μ = 4 x 10^{-5}. Selection coefficient: s = 0.2. In all cases, there was an initial phase of neutral evolution lasting for at least 2 x 10^4 generations (selection coefficient was set to s = 0 during this phase)..In the case of soft sweep (panel B) the allele frequency at the selected locus in either population was conditioned to be between 30% and 70% when selection started. Approach to equilibrium is slow but the patterns obtained at the end of simulations are similar to those expected at equilibrium. Note the logarithmic timescale on the *x*-axis, and different spacing between neighbouring loci on the *y*-axis.

Figure 3: Relative differentiation F_{ST} obtained in a single stochastic realisation of the model in the case of a hard sweep occurring in a primary contact. All parameters are same as in Figure 2. Note that the range of F_{ST} values obtained in the individual realisation is larger than the range used in the colour bar (the highest F_{ST} values are >0.4, very close to unity), but for simpler interpretation and comparison to the results shown in Fig. 2, all values here are truncated to the range used in Fig. 2.

Figure 4: Comparison of average relative differentiation F_{ST} and average absolute differentiation d_{XY} at the neutral locus at distance 500 kb from the locus under selection during a hard sweep in primary contact, as a function of time since the start of selection. Shown is the result obtained in. Blue: F_{ST} , red: d_{XY}. Averages are made over 10000 independent evolutionary histories. All other parameters are same as in Figure 2.

Tables

Table 1: Examples of systems where evidence of heterogeneous genomic differentiation or divergence has been identified using genome-scan approaches. Note that this table is not intended to be an exhaustive summary.

Study system	NGS approach	Genome scan approach	Main findings	Reference
European rabbit subspecies: *Oryctolagus cuniculus cuniculus* & *O. c. algirus*	Target sequencing, RNA-seq	Sliding window estimates of F_{ST}, d_{XY} ,RND, number of fixed differences and ratio of fixed differences to shared polymorphism	Low genome-wide mean F_{ST}; numerous regions (1.8%) highly differentiated. Overrepresentation on sex chromosome and centromeres suggest minor role for selection. Sweeps do not account for majority of differentiation peaks.	Carneiro *et al.*, (2014)
Fruit fly subspecies: *Drosophila pseudoobscura pseudoobscura* & *D. persimilis*	Whole-genome sequencing	Sliding window estimates of d_{XY} and nucleotide diversity.	High nucleotide diversity and divergence in inversions compared to collinear regions due to reduced recombination	Mcgaugh & Noor, (2012)
Sunflowers: *Helianthus annuus* & *H. petiolaris*	RNA-seq	Sliding window and spatial autocorrelation statistics based on F_{ST}	Lower overall differentiation in sympatry, number and size of genomic islands did not differ with geography, strong negative correlation with recombination regardless of spatial context	Renaut *et al.*, (2013)
Marine and freshwater three-spined stickleback ecotypes: *Gasterosteus aculeatus*	Whole-genome sequencing	Sliding window estimates of F_{ST}, nucleotide diversity and Hidden Markov model detection of outlier regions	150-242 genomic regions of high differentiation across genome. Evidence of parallel reuse of standing variation in different populations.	Jones *et al.*, (2012)
Walking stick ecotypes: *Timema cristinae*	Whole-genome sequencing	Point estimates of F_{ST} at SNP positions, HMM detection of outlier regions	Median F_{ST} greater between geographically separated populations compared to adjacent – 8-30% genome highly differentiated; also evidence of non-parallelism.	Soria-Carrasco *et al.*, (2014)
House mouse subspecies:	RNA-seq	Sliding window estimates of	Higher differentiation on sex	Phifer-Rixey *et al.*,

Mus musculus musculus & *M. m. domesticus*		F_{ST}, d_{XY} and allele frequency differences (δ)	chromosomes. Many regions of high differentiation between species but low within subspecies. Evidence of local selective sweeps and/or barrier loci.	(2014)
Hawthorne and apple Apple maggot fly ecotypes: *Rhagotelis pomonella*	Microsatellites and allozymes	Point estimates of F_{ST}	Two genomic outlier regions on separate chromosomes, suggesting support for genomic island and continent hypotheses.	Michel *et al.*, (2010)
Normal benthic and dwarf limnetic whitefish ecotypes: *Coregonus clupeaformis*	RAD-sequencing	Sliding window estimates of F_{ST} and barrier strength (m/m_e)	Positive correlation of mean and variance of F_{ST}, outlier region size and LD with morphological differentiation. Island size influenced by LD, selection strength and demography. Incomplete parallelism of outliers.	Gagnaire *et al.*, (2013)
Annual and perennial Yellow Monkey Flower ecotypes: *Mimulus guttatus*	Genotyping-by-sequencing	Sliding window estimates of FCT, nucleotide diversity and divergence (d_{XY})	Outlier regions distributed across genome, but enriched in an inversion with barrier loci. Co-linear regions probably homogenized by gene flow.	Twyford & Friedman, (2015)
M and S mosquito forms: *Anopheles gambiae*	SNP-genotyping array	Sliding window estimates of nucleotide diversity and F_{ST}	Regions of high differentiation at centromeres, low nucleotide diversit in high recombination regions suggest recent sweeps and ‘incidental islands’	Neafsey *et al.*, (2010)
Neotropical butterfly species: *Heliconius melpomne*, *H. cydno* & *H. timareta*	Whole-genome-sequencing	Sliding window estimates of F_{ST} and ABBA-BABA tests for gene flow	Low F_{ST} between sympatric species, higher differentiation and lower gene flow on sex chromosomes and at loci underlying divergent wing patterns	Martin *et al.*, (2013)
Flycatchers: *Ficedula sp.*	Whole-genome-sequencing	Non-overlapping window estimates of F_{ST}, d_{XY}	Strong correlations between F_{ST} amongst independent species comparisons and with recombination rate suggests heterogeneity caused by	Burri *et al.*, (2015)

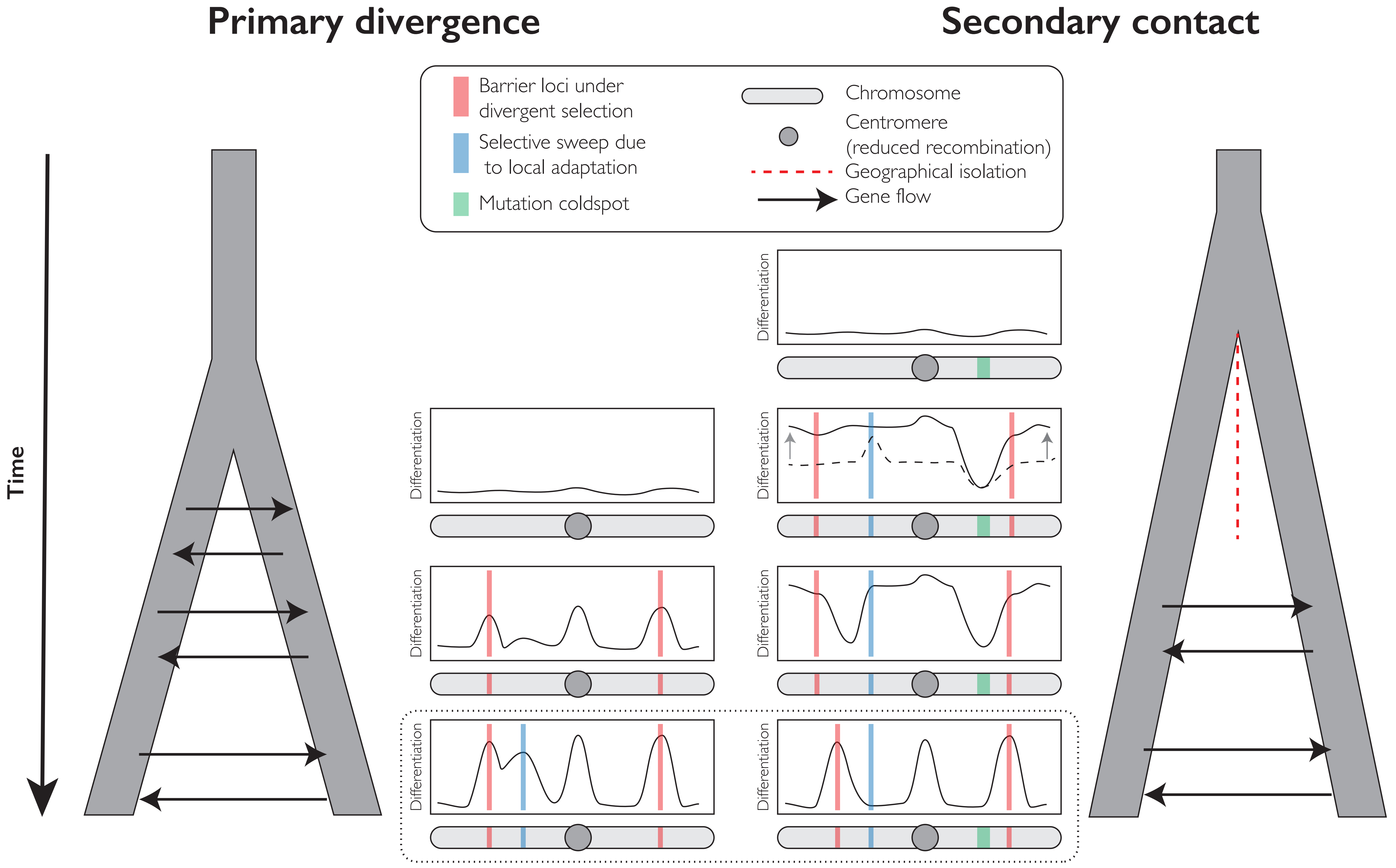
background selection

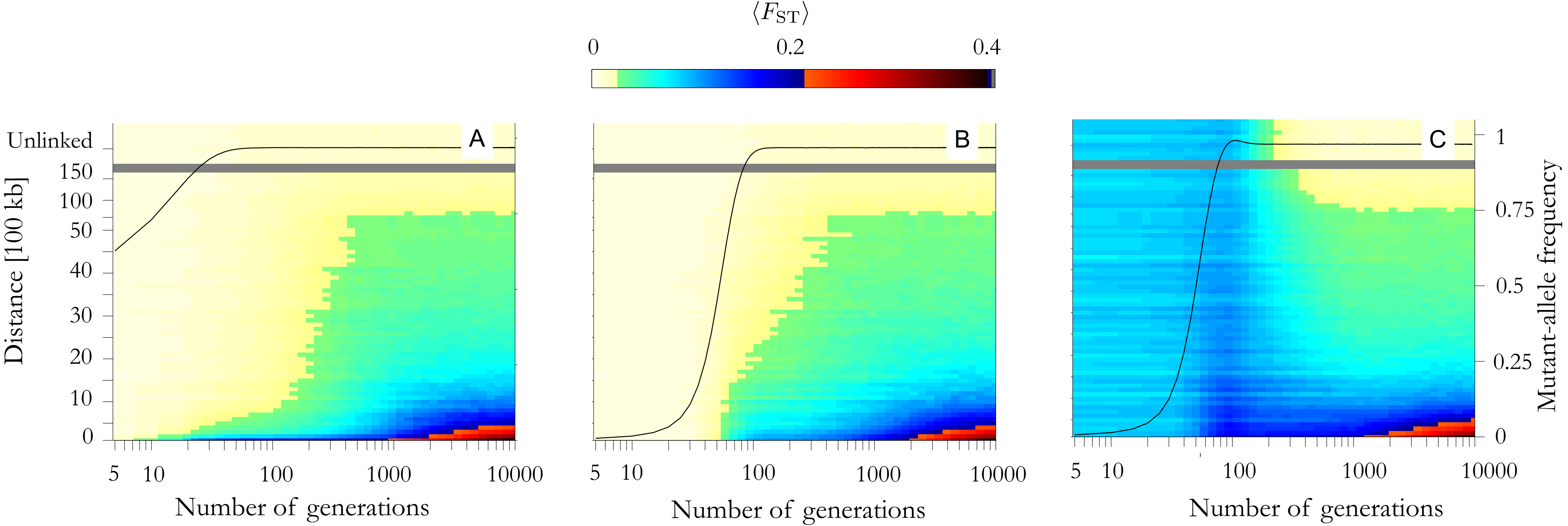
Table 2: Examples of studies where alongside genome scan data, additional evidence besides genome-scan data has been used to demonstrate that selection occurs at outliers. Here we delineate between studies that demonstrate a genotype-phenotype link (table section in grey), which requires separate evidence of selection on the phenotype, and studies that show signatures of selection on the genotype (unfilled table section). We note that in some cases, e.g. lateral plate armour in three-spined sticklebacks, there are overlaps between these categories.

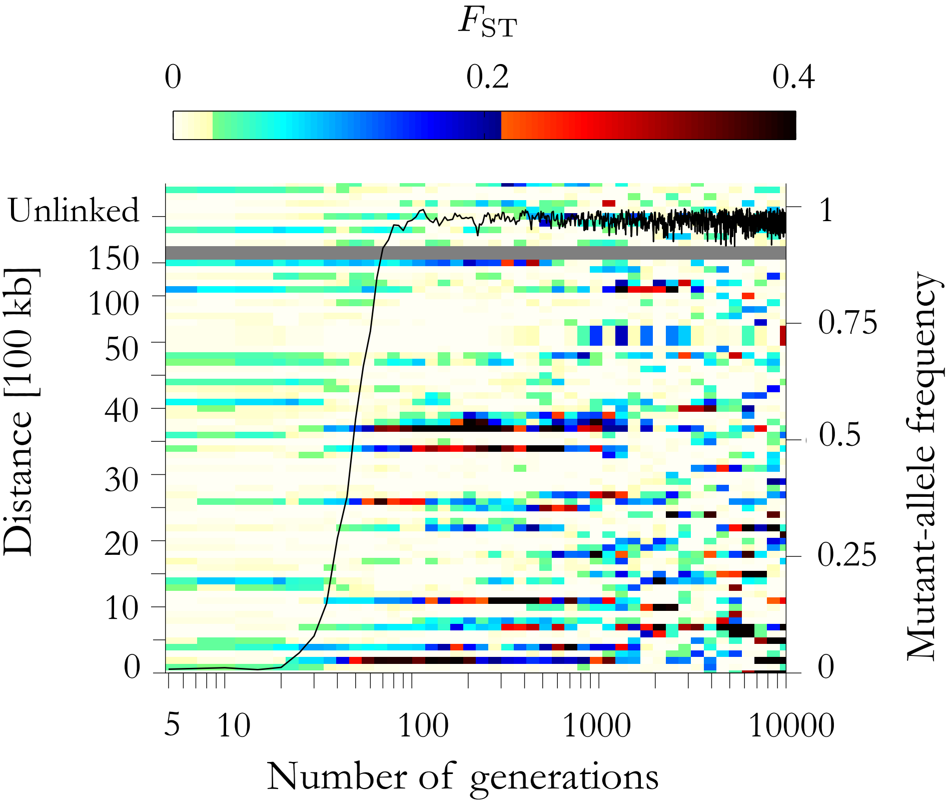
Type of evidence	**Description**	**Caveats**	**Examples**
QTL mapping and other mapping approaches	Identifies genomic basis of known divergent trait or hybrid incompatibility; correspondence of QTL with outliers provides strong evidence for selection	Narrowing genomic region requires large numbers of individuals and high density of markers. Potential bias towards large-effect or clustered loci.	Overlap between QTL and outliers for benthic – limnetic whitefish (Rogers & Bernatchez, 2005). Allele frequency shifts at SNPs linked to QTL for skeletal morphology in lake-stream sticklebacks (Berner *et al.*, 2014). Reduction in sperm number maps to sex chromosomes in Pacific Ocean-Japan Sea stickleback cross (Kitano *et al.*, 2009).
Gene ontology analysis	Test whether outliers have functions that are expected to be divergent (based on observations of phenotypic divergence or known selection pressures	Relatively weak evidence, limited by annotation quality.	Groundsels on different soil types often have different outliers, but similar annotations (Roda *et al.*, 2013). Flowering time genes divergent across latitudinal gradient in sunflowers (Renaut *et al.*, 2013).
Molecular assay	Functional assays of gene products using *in vitro* methods	Usually cannot be formed using study organism.	Cichlid opsin light absorbance (Terai *et al.*, 2006). Expression of Pocket mice

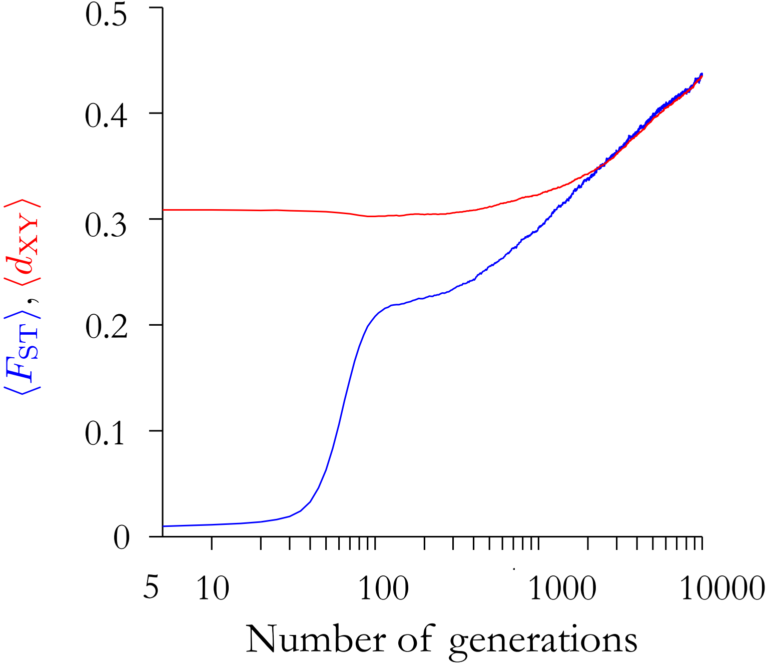
			Mc1r alleles in cultured cells (Hoekstra *et al.*, 2006).
Transgenics	Insertion or deletion of alleles into different genetic background and observation of phenotype	Technically difficult for most organisms	Insertion of high-plated Eda allele into low-plated genomic background (Colosimo *et al.*, 2005) and restoration of pelvic spine phenotype in sticklebacks (Chan *et al.*, 2010).
Knockout/knockdown	Deletion, disruption or suppression of genes underlying divergence traits to demonstrate phenotypic effects	Can only demonstrate loss of function. Target fidelity is difficult to control.	Knockdown of genes related to albinism in cavefish (Bilandžija *et al.*, 2013) and *doublesex* gene controlling mimicry patterns in *Papillo* butterflies (Nishikawa *et al.*, 2015)-
Genomic clines	Steep clines expected for loci under strong ecological selection across ecotones or for genes involved in reproductive isolation across hybrid zones	Recombination, mutation rate and population demography can distort clinal data	Overlap between outlier loci and steep genomic clines in bivalve subspecies (Luttikhuizen *et al.*, 2012). Loci with steep clines at genes known to be involved in RI (Trier *et al.*, 2014).
Parallel evolution	Parallel differentiation at the same locus, genomic region or gene class across multiple geographically and phylogenetically independent species/population pairs	Parallel differentiation caused by shared genomic constraints – i.e. background selection and low recombination – must be ruled out. Parallel differentiation also produced by secondary contact.	Same loci involved in marine-freshwater stickleback divergence across large geographical scales (Jones *et al.*, 2012). Increased differentiation amongst stream populations flanking genomic regions involved in phenotypic differentiation in lake-stream sticklebacks due to propagating selective sweeps (Roesti *et al.*, 2014).
Experimental crosses	Observation of segregation distortion, hybrid sterility or hybrid inviability allows for identification of	Cross designs can be complicated and often only possible in model	Crosses between *Drosophila* subspecies show male sterility and segregation distortion (Phadnis & Orr, 2009).

	intrinsic incompatibilities	organisms – particularly when inviability is present. Cannot always identify extrinsic selection against hybrids.	Evidence of ecological incompatibilities from limnetic-benthic stickleback crosses in artificial environments (Arnegard *et al.*, 2014).
Transplant experiments	Transplanting hybrids or individuals from divergent habitats into a maladaptive environment results in changes in allele frequency or a reduction in fitness and survival.	Not feasible for some species and also difficult to discount selection on other adaptive loci	Switching stick insects between host plants (Gompert *et al.*, 2014); transplant of marine sticklebacks with known lateral plate Eda genotype demonstrates reduced fitness and allele frequency shifts (Barrett *et al.*, 2008).


Primary divergence
Secondary contact
Barrier loci under divergent selection
Selective sweep due to local adaptation
Mutation coldspot
Chromosome
Centromere (reduced recombination)
Geographical isolation
Gene flow
Time
Differentiation
Differentiation
Differentiation
Differentiation
Differentiation
Differentiation
Differentiation


$\langle F_{\mathrm{ST}} \rangle$
0
0.2
0.4
A
B
C
Unlinked
150
100
50
40
30
20
10
0
Distance [100 kb]
5
10
100
1000
10000
Number of generations
Number of generations
Number of generations
1
0.75
0.5
0.25
0
Mutant-allele frequency


F_{ST}
0
0.2
0.4
Unlinked
150
100
50
40
30
20
10
0
Distance [100 kb]
1
0.75
0.5
0.25
0
Mutant-allele frequency
5
10
100
1000
10000
Number of generations


0.5
0.4
0.3
0.2
0.1
0
$\langle F_{\mathrm{ST}} \rangle$, $\langle d_{\mathrm{XY}} \rangle$
5
10
100
1000
10000
Number of generations

Supplementary material

In this Supplementary Material, the model used to generate Figs. 2-4 in the main text is explained.

Appendix S1. MODEL

We model two populations, each with N diploid individuals. It is assumed that the two populations exchange migrants at a rate m per generation, individual, population. All loci are assumed to be bi-allelic. The two allelic types are denoted by A and a. One locus is assumed to be under divergent selection. At this locus, one of the two alleles (A) is favoured in the first population, and the other (a) is favoured in the second population. In the first population, the fitnesses of genotypes AA, Aa, and aa are 1, $1 - s/2$, and $1 - s$. In the second population the corresponding fitnesses are $1 - s$, $1 - s/2$, and 1. When $s = 0$, all loci are evolving neutrally. In all simulations, during the initial 20000 generations (or longer) we set $s = 0$. In what follows this phase is referred to as *the neutral-evolution phase*.

In addition to the locus under selection, we assume that there are L_{linked} neutral loci at increasing recombination distances from the locus under selection. One of the neutral loci is fully linked to the selected locus. For the remaining loci, recombination occurs at a rate r between a pair of adjacent loci on the chromosome. Furthermore, there are L_{unlinked} neutral loci unlinked to the selected locus.

The lifecycle of individuals is modelled in the following order: migration of virgin adults, mating locally within each population, recombination, fecundity selection, mutation. All neutral loci are assumed to be subject to mutation at a rate μ per allele, locus, individual, population, generation. We use a symmetric two-allele mutation model: a mutation changes allele a to A, and vice versa.

Following the neutral-evolution phase (see above), we model a primary-contact divergence, or a secondary-contact divergence.

In the primary-contact divergence case we assume that divergent selection starts immediately after the neutral-evolution phase, so that s is larger than zero (and constant over time) at the locus assumed to be under selection. Here we distinguish two cases that are briefly discussed next. First, we assume that the locus under selection has no genetic variation prior to the initiation of divergent selection (all alleles are of type A), and we introduce only one mutation (allele a) in the second population, where this allele is beneficial. This corresponds to “a hard sweep”. After this mutation has been introduced, we neglect any further mutations at the locus under selection. Note that the mutation can, by chance, be lost during the initial phase of divergence. In the simulations we discard such cases, and the results we present are conditional on that divergence due to the introduction of a single mutation is successful. Second, we assume that, when divergent selection starts, the locus under selection has genetic variation that has been accumulated during the neutral-evolution phase. This corresponds to “a soft sweep”. To be able to make a clear distinction between a hard-sweep case and a soft-sweep case, in these simulations we condition on that, when selection starts, the allele-frequencies at the locus under selection are between 0.3 and 0.7 in both populations.

By contrast to the primary-contact model, in the secondary-contact model we assume that, after the neutral-evolution phase, there is a period of complete isolation between the two populations (the migration rate is set to zero). In this period of complete isolation, it is assumed that $s = 0$ during the first N generations. Thereafter, we model a hard sweep while the populations are still isolated. This is modelled similarly to the case of hard sweep in a primary contact (see above), but here there is no migration while the sweep occurs. When the frequency of the locally beneficial allele becomes one in the population where the sweep occurs, we reintroduce migration between the two populations (secondary contact). The migration rate during the secondary contact is assumed to be equal to the migration rate prior to the period of complete isolation.

Note that in cases where a sweep occurs during a period of isolation between the populations, the beneficial mutant allele sweeps to fixation. This is not true for sweeps occurring during a primary contact because migration introduces locally deleterious alleles, but the frequency of the locally deleterious allele is expected to be smaller than the frequency of the locally beneficial allele. The difference in the two frequencies depends on the migration-selection-drift balance. Note also that in the secondary-contact case, the same migration-selection-drift balance is attained after migration between the two populations has been reestablished.

In the simulations, we measure in generation τ the total-population heterozygosity $\Pi_{\mathrm{T},\tau}$, as well as the within-population heterozygosities $\Pi_{\mathrm{S},\tau}^{(1)}$, and $\Pi_{\mathrm{S},\tau}^{(2)}$ for the first and second population, respectively. Here index “S” stands for “subpopulation”, and τ denotes a generation (measured since the start of divergent selection) when the corresponding measure is taken. In a majority of the simulations we take measures every 5th generation. The within- and total-population heterozygosities allow for computing the statistics $F_{\mathrm{ST},\tau}$ ad $d_{\mathrm{XY},\tau}$ as follows (Cruickshank and Hahn, 2014):

$$F_{\mathrm{ST},\tau} = \frac{\Pi_{\mathrm{T},\tau} - \frac{\Pi^{(1)}_{\mathrm{S},\tau} + \Pi^{(2)}_{\mathrm{S},\tau}}{2}}{\Pi_{\mathrm{T},\tau}} \ ,$$

$$d_{\mathrm{XY},\tau} = 2\Pi_{\mathrm{T},\tau} - \frac{\Pi^{(1)}_{\mathrm{S},\tau} + \Pi^{(2)}_{\mathrm{S},\tau}}{2} \ . \tag{S1}$$

Note that for the model used here (symmetric two-allele mutation model), the number of differences between haplotypes sampled from the different populations at a given distance from the selected locus is either 0 or 1. Therefore, $d_{\mathrm{XY},\tau}$ is equal to the probability that an allele sampled randomly from one population is different from an allele sampled randomly from the opposite population.

Appendix S2. PARAMETER CHOICES

The parameter values used in the simulations are listed in Table S1. We run the model for 10^4 generations after the neutral-evolution phase. We perform 5000 independent realisations for each case modelled (unless stated otherwise).

The results obtained under the model are shown and discussed in the main text.

Tables

TABLE 1 Parameters of the model, their explanations, and the values used in our computer simulations.

Parameter	Explanation	Values
N	Number of individuals in each population	500
m	Migration rate	0.004
s	Selection coefficient	0.2
r	Recombination rate between a pair of adjacent loci[a]	$0, 0.001, 0.01,\ 0.5$
L_{linked}	Number of loci linked to the selected locus	61
L_{unlinked}	Number of loci unlinked to the selected locus	10

[a]: One neural locus is fully linked to the selected locus ($r = 0$). For the next 50 neutral loci, the recombination rate between a pair of adjacent loci is set to 0.001. Then, for the next 10 neutral loci, the corresponding recombination rate is 0.01. The recombination rate between any locus and an unlinked locus is 0.5.

References

Cruickshank, T. E., Hahn, M. W., 2014. Re-analysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23, 3133–3157.