



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/120727/>

Version: Accepted Version

Article:

Ravinet, M., Faria, R., Butlin, R.K. et al. (2017) Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30 (8). pp. 1450-1477. ISSN: 1010-061X

<https://doi.org/10.1111/jeb.13047>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 **Target review**

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

4
5 Mark Ravinet^{1,2*}, Rui Faria^{3,4,5}, Roger K. Butlin^{5,6}, Juan Galindo⁷, Nicolas Bierne⁸, Marina
6 Rafajlović⁹, Mohamed A. F. Noor¹⁰, Bernhard Mehlig⁹ & Anja M Westram⁵

7

8 ¹Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway.

9 ²National Institute of Genetics, Mishima, Shizuoka, Japan

10 ³CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório
11 Associado, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

12 ⁴IBE, Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health
13 Sciences, Pompeu Fabra University, Doctor Aiguader 88, 08003 Barcelona, Spain

14 ⁵Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK,

15 ⁶Centre for Marine Evolutionary Biology, Department of Marine Sciences, University of
16 Gothenburg, Gothenburg, Sweden.

17 ⁷Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain

18 ⁸CNRS, Université Montpellier, ISEM, Station Marine Sète, France

19 ⁹Department of Physics, University of Gothenburg, Sweden

20 ¹⁰Biology Department, Duke University, Durham, North Carolina, USA

21

22

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

24

25

26 Running head: The genomic landscape of speciation

27

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

30 **Abstract**

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

49

50

51

52

53

54

55

56

57

58

59

60

61

62 Introduction

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

78

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

92

93 Despite progress in documenting patterns, interpreting the peaks and troughs of differentiation
94 in genome scan data has not been as straightforward as initially assumed (Fig. 1). This has

95 caused problems for researchers hoping to use genome scans to identify signatures of local
96 adaptation (Hoban *et al.*, 2016) and barriers to gene flow during speciation (Noor & Bennett,
97 2009; Cruickshank & Hahn, 2014). There are several reasons for these difficulties. Firstly, peaks
98 of high differentiation are produced in diverging populations without gene flow as a result of
99 background selection and selective sweeps after isolation (Charlesworth *et al.*, 1993; Noor &
100 Bennett, 2009; Cruickshank & Hahn, 2014; Burri *et al.*, 2015). Although some of these peaks
101 may indicate loci that become barrier loci after contact, many other peaks do not. Instead they
102 may reflect sweeps of universally adaptive alleles, genomic conflict, sexual selection, or drift.
103 Therefore, the effects of barrier loci can be clearly identified only when they have actually
104 recently acted to prevent gene flow in Nature (Harrison & Larson, 2016; Marques *et al.*, 2016;
105 McGee *et al.*, 2016). Allopatric divergence remains important for understanding genome-wide
106 heterogeneity in the absence of gene flow (Noor & Bennett, 2009); however barrier loci can be
107 identified by hybridization either in the field or in the lab. Tests for on-going or recent gene flow
108 are therefore a crucial prerequisite for the identification of barrier loci from genome scans.
109 Secondly, patterns of F_{ST} (or other differentiation and divergence measures) are influenced by
110 multiple factors that vary across the genome, including mutation, demographic history, genetic
111 drift, selection, gene flow, recombination, gene density, and genome architecture; and some of
112 these factors are expected to change during different stages of speciation (Fig. 1). From the
113 speciation perspective, the principal objective is to infer the number, distribution and strength of
114 barriers to gene flow, as well as their influence on other genomic regions. However, extracting
115 this signal from genome scan data in the presence of so many other processes remains
116 challenging.

117

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

129 landscape into context by combining them with other sources of data (e.g. experiments) to gain
130 further insight into the speciation process.

131

132 **Section 1: Barriers to gene flow in the genomic** 133 **landscape**

134 **Barrier loci and barrier effects**

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

152

153 In order for an allele at a barrier locus to under divergent selection spread and contribute to a
154 barrier effect in the long term, selection locally favouring this allele must be strong enough to
155 overcome the opposing effect of gene flow (Haldane, 1930; Slatkin *et al.*, 1985; Slatkin, 1987).
156 In small populations the efficacy of selection is reduced by greater drift, and stronger selection is
157 sometimes needed to reach a given degree of differentiation (Yeaman & Otto, 2011). The
158 distribution of barrier locus effect sizes in a given case study is therefore likely to depend on
159 both effective population size (N_e) and migration (m). For large populations with strong extrinsic
160 barriers to the exchange of individuals, barrier effect sizes should vary over a wide range,

161 whereas in small populations exchanging many migrants, only large-effect barrier loci are
162 expected (Yeaman & Whitlock, 2011). The distribution is also expected to vary with progression
163 towards speciation and demographic history; small effect alleles may be more common during
164 periods of geographical isolation than during contact, and late rather than early speciation,
165 although these scenarios need to be investigated more thoroughly. The effect-size distribution
166 of barrier loci remains elusive because although theoretical work shows that even alleles under
167 very weak selection may temporarily contribute to phenotypic divergence (Yeaman, 2015), loci
168 with small fitness effect sizes are difficult to identify from empirical data. The same is true for
169 phenotypic effect sizes; loci of large effect are easier to detect (Rockman, 2012). Empirical work
170 often focuses on loci with large phenotypic and fitness effects, e.g. stickleback plate armour
171 (Colosimo *et al.*, 2005) and pelvic spine reduction (Shapiro *et al.*, 2004; Chan *et al.*, 2010), but
172 the general pattern remains unclear (e.g. Seehausen *et al.*, 2014).

173

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

190

191 Even if an outlier scan correctly identifies a genomic region containing a barrier locus, narrowing
192 the region down to the barrier locus itself may be difficult. This is partly because measures of
193 differentiation are noisy, due to stochasticity in coalescence as well as sampling (Fig. 3), but
194 also due to the resolution of the scan and the chromosomal scale influenced by the barrier

195 effect: large blocks of linkage disequilibrium can occur in some species. Given the complexity
196 and cost of dealing with whole genome data, particularly in non-model organisms, the vast
197 majority of genome scan studies still make use of reduced representation sequencing
198 approaches (Davey *et al.*, 2011; Andrews *et al.*, 2016). In these cases, outlier markers may
199 frequently show high differentiation because they are linked to a barrier locus, rather than being
200 the direct target of selection. For genomic regions under selection, multiple SNPs may often
201 show elevated differentiation (hence the island concept – see Box 1), although there may be
202 variance among sites because of drift-related stochasticity. For this reason, differentiation in
203 whole-genome data is usually calculated across a window spanning multiple variants rather
204 than using single nucleotides. However the resolution of this approach might mean
205 differentiated regions are missed, especially at the start of the speciation continuum when
206 genetic differentiation decays rapidly with genomic distance (Hoban *et al.* 2016).

207

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

219

220 **Loci linked to barrier loci**

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

228 needing to be under selection themselves, markers may indicate the presence of barrier loci by
229 showing elevated F_{ST} due to linkage.

230

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

245

246 One important departure from equilibrium happens during and after a selective sweep (Box 1),
247 where an adaptive allele increases in frequency. In Fig. 2 we demonstrate the development of
248 F_{ST} from a transient state towards equilibrium for a soft sweep under continuous gene flow (Fig.
249 2A), a hard sweep under continuous gene flow (Fig. 2B), and a hard sweep in allopatry followed
250 by secondary contact for comparison (Fig. 2C). However, this figure shows differentiation
251 averaged over a 5000 independent evolutionary histories, meaning the results obtained are in
252 approximate agreement with the theoretical expectations (not shown). In Fig. 3 we show the
253 outcome of a single evolutionary history to illustrate effects of stochasticity, demonstrating that
254 deviations from the expectation are possible (see Supplementary Material for more details on
255 the simulations run and parameters used to generate these illustrations).

256

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

262 (Przeworski, 2002; Hermisson & Pennings, 2005; Pennings & Hermisson, 2006; Messer &
263 Petrov, 2013). However, the genomic region where average F_{ST} is increased is relatively small
264 immediately after the sweep, and grows towards equilibrium (i.e. from left to right in Figs. 2A &
265 B). This is because the haplotype (or haplotypes) sweeping to high frequencies contain
266 common alleles at most loci, initially generating little differentiation. Therefore, F_{ST} may initially
267 remain low even in genomic regions where m_e is reduced due to linkage. However, over time,
268 this reduced m_e allows for an accumulation of allele frequency differences due to both drift and
269 new mutations. These patterns indicate that barrier loci that have undergone sweeps in the face
270 of gene flow may be more easily detectable when they are closer to equilibrium, because the
271 proportion of surrounding loci showing elevated differentiation increases with time after the
272 sweep (Fig 2 & B). However, it is unclear how quickly equilibrium is approached (Wood & Miller,
273 2006; Bierne, 2010; Yeaman *et al.*, 2016). This approach may be slow because it requires both
274 mutation and rare recombination events, suggesting many loci in empirical studies are not at
275 equilibrium.

276

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

291

292 In some cases the contrast between F_{ST} and d_{XY} is likely to be helpful for distinguishing between
293 transient states and equilibrium, facilitating the correct interpretation of outlier loci (Cruickshank
294 & Hahn, 2014; Delmore *et al.*, 2015; Irwin *et al.*, 2016). Relative measures such as F_{ST} may
295 miss the distinct effects on diversity and divergence (Charlesworth *et al.* 1997), and peaks of

296 differentiation can be present for both recent local sweeps (transient) and in equilibrium (see
297 above and Fig. 2). Measures of absolute divergence such as d_{XY} in regions surrounding barrier
298 loci take longer to increase via the establishment of new mutations. Recent local sweeps should
299 be characterised by F_{ST} peaks lacking elevated d_{XY} , while in equilibrium both F_{ST} and d_{XY} are
300 expected to be higher in the vicinity of barrier loci because of the reduction of effective migration
301 rate (Fig. 4). Unfortunately, such distinct behaviour of F_{ST} and d_{XY} might not apply to more
302 complex scenarios involving secondary or intermittent contact. These scenarios need further
303 investigation.

304

305 The spread of barrier effects to linked neutral loci is uncontroversial. More contentious is the
306 effect of a barrier locus on divergence of linked loci that are also under divergent selection.
307 Some F_{ST} outlier analyses have identified loci that occur in proximity to QTL, e.g. at a distance
308 of ~ 10 cM in pea aphids (Via & West, 2008), and divergently selected loci may cluster in the
309 genome (Yeaman, 2013). Moreover, in some species, highly differentiated genomic regions
310 appear to increase in size along the speciation continuum (Feder *et al.*, 2012a; Renaut *et al.*,
311 2012). These findings suggest that further divergence might be more likely in the vicinity of
312 existing barrier loci, and that this might lead to a growth of highly differentiated genomic regions.
313 Conceptual thinking has focused on one potential explanatory mechanism, divergence
314 hitchhiking (Via & West, 2008; Feder *et al.*, 2012a; b). Under this framework, reduced m_e around
315 divergently selected loci may facilitate the establishment of new mutations under weak divergent
316 selection in their vicinity (Feder *et al.*, 2012a; Nosil & Feder, 2012a; Via, 2012), causing an
317 increase in size of differentiated regions (Feder *et al.*, 2012a; Via, 2012, see also next section).
318 However, using multi-locus simulations Feder and colleagues (2010; 2012b) demonstrated that
319 divergent selection facilitated the establishment of weakly adaptive mutations only under limited
320 conditions when selection is strong, N_e is small and migration is low. Furthermore, if divergence
321 hitchhiking does occur, Hill-Robertson interference may prevent weakly adaptive alleles from
322 establishing when they arise in habitats or genomes where they are maladaptive and they are
323 unable to escape via recombination (Feder *et al.*, 2012b; Yeaman, 2015). Clustering under high
324 migration load can be facilitated by chromosomal rearrangements or when linkage allows
325 multiple weakly selected alleles to surpass the m_e threshold preventing homogenisation
326 (Yeaman, 2013, 2015). Alternatively, clustering may occur when weak differentiation is better
327 protected from loss via drift when linkage to a strongly diverged locus is tight (Rafajlovic *et al.*
328 2016). However, if drift is strong enough to allow new adaptive loci to regularly replace those
329 lost via stochasticity, selection against recombinants can favour clustering without the need of

330 recombination modifiers (Yeaman, 2013; Yeaman *et al.*, 2016). Theoretical and empirical
331 evidence therefore suggests that selection against recombinants under high migration load may
332 facilitate the clustering but not necessarily the establishment of barrier loci.

333

334 **Barriers and genome-wide effects**

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

351

352 Selection on multiple traits, i.e. multifarious selection, is thought to be more likely to facilitate
353 speciation than strong selection on a single trait (Rice & Hostert, 1993; Nosil *et al.*, 2008; Nosil,
354 2013). Similarly, selection against migrants at multiple loci results in a stronger barrier to gene
355 flow, reducing effective migration rate across the genome when overall selection is the same
356 (Barton & Bengtsson, 1986; Feder *et al.*, 2012b). This allows new locally-adaptive mutations to
357 establish, independent of their genomic position, even if their effect size is relatively small; it
358 also facilitates an increase in genome-wide divergence at neutral regions due to drift (Feder *et al.*
359 *et al.*, 2012b). This process has been termed genome hitchhiking (Feder *et al.*, 2012a) and it
360 essentially describes the impact of multifarious divergent selection when coupling is strong.
361 Flaxman and colleagues (2014) used simulations to demonstrate that statistical associations
362 amongst a large number of genes combined with divergent selection can interact to drive a
363 rapid transition from local to genome-wide barrier effects. This genome-wide congealing (GWC)

364 is resembles the coupling transition predicted by Barton (Flaxman *et al.*, 2014; Tittes & Kane,
365 2014). During progression towards speciation in their model, numerous, weakly selected
366 mutations occur but are unable to generate differentiation due to the effects of gene flow.
367 Following a transition from local to genome-wide barriers, however, the contribution of these
368 mutations to reproductive isolation increases as the genome-wide m_e is reduced below a
369 threshold and LD increases (Tittes & Kane, 2014). Importantly, GWC does not require physical
370 linkage or periods of allopatry that might elevate LD amongst loci (Tittes & Kane, 2014).
371 Nonetheless, Flaxman *et al.* (2014) demonstrate that genomic features such as chromosome
372 length or clustering of adaptive loci on specific chromosomes, as well as periods of
373 geographical isolation, can drastically reduce the waiting time to GWC. Simulations show that
374 both genome hitchhiking and genome-wide congealing are able to occur under a wide range of
375 parameters provided there is selection on many loci (Feder & Nosil, 2010; Nosil & Feder,
376 2012b). However, as with divergence hitchhiking, empirical evidence showing that genome
377 hitchhiking allows weakly adaptive alleles to establish remains elusive.

378

379 **Section 2: Other factors modifying the genomic** 380 **landscape**

381

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

387

388 **Demographic and evolutionary history**

389 Understanding the demographic and evolutionary history of population and species pairs is
390 necessary to generate expected patterns of genomic differentiation. Fluctuations in effective
391 population size (N_e) can have a profound effect in this regard; *e.g.* when N_e is small, the effect of
392 drift is greater whereas selection is more efficient when N_e is large (Charlesworth *et al.*, 2003;
393 Charlesworth, 2009; Charlesworth & Charlesworth, 2010). Pronounced changes in N_e such as
394 bottlenecks can shift the mean and variance of baseline genomic differentiation, making it
395 difficult to identify highly differentiated regions (Ferchaud & Hansen, 2016). N_e is an important

396 parameter to estimate because, as well as determining the effectiveness of selection, it
397 influences scaled mutation and recombination rates; for example, scaled mutation rate, $N_e\mu$
398 determines the rate at which adaptive mutations enter a population (Hartl & Clark, 2007;
399 Charlesworth & Charlesworth, 2010).

400

401 We have emphasised the need to test for gene flow (see Box 3) to better appreciate the relative
402 role of alternative processes explaining a landscape of heterogeneous genomic differentiation
403 (see Box 2). When populations or species meet, the landscape may point to barrier loci resistant
404 to gene flow (Harrison & Larson, 2016), but without accounting for divergence history, it is not
405 clear whether these populations have diverged *in situ* or have resisted genome-wide
406 homogenization upon secondary contact between divergent lineages (Bierne *et al.*, 2013; Feder
407 *et al.*, 2013). First, with recent secondary contact, peaks of differentiation may just reflect loci
408 that differentiated due to drift during allopatry, and have yet to be homogenised by gene flow.
409 Such spurious outliers may obscure or hinder the detection of true barrier loci. Second, the
410 genomic signatures of selected loci may also differ between primary divergence and secondary
411 contact (Fig. 2). With primary divergence, during and immediately after a local selective sweep,
412 transient high differentiation peaks will occur at large distances from the selected locus but
413 these are eroded by recombination and migration (Fig. 3). In contrast, during allopatry, this
414 erosion does not happen, generating large regions of high differentiation, which will be
415 maintained for some time after secondary contact. Therefore, for local sweeps of comparable
416 age, differentiated regions will often be much larger (and therefore potentially easier to detect) in
417 secondary compared to primary divergence as migration has had less time to act. Recent
418 studies have explicitly tested for primary vs. secondary contact, allowing for a more accurate
419 interpretation of genome scan data; a wide array of tools is available for this sort of approach
420 (Sousa & Hey, 2013; see also Box 3). Some have provided support for primary divergence
421 (Nosil *et al.*, 2012; Butlin *et al.*, 2014) whereas others indicate that secondary contact after a
422 period of isolation best explains heterogeneous differentiation (Tine *et al.*, 2014; Martin *et al.*,
423 2015b; Roesti *et al.*, 2015; Rougemont *et al.*, 2016).

424

425 Even sophisticated statistical frameworks for testing divergence hypotheses only consider a
426 small proportion of the 'universe of potential historical scenarios' (Knowles, 2009). Divergence
427 history varies across the genome due to non-uniformity in effective migration rate, effective
428 population size and recombination (Maddison, 1997; Roux *et al.*, 2014, 2016; Mallet *et al.*,
429 2016). Gene-tree vs. species-tree discordance can occur because of introgression (Maddison,

430 1997; Knowles & Maddison, 2002; Geneva *et al.*, 2015; Rosenzweig *et al.*, 2016), but also
431 because of incomplete lineage sorting (ILS) (Hobolth *et al.*, 2011b; Dutheil & Hobolth, 2012).
432 Described as ‘deep coalescence’ by Maddison (1997), ILS occurs when the most-recent
433 common ancestor for a genealogy exists before speciation begins, resulting in counter-intuitive
434 three taxa phylogenies (Sally *et al.*, 2012) or distortions of divergence time estimates between
435 two species (Leaché *et al.*, 2013). ILS therefore increases the variance of genomic divergence
436 estimates, making it difficult to identify true outliers and also potentially introducing false
437 positives. ILS affects a greater proportion of the genome when speciation events occur close in
438 time and the ancestral effective population size is large (Barton, 2006; Hobolth *et al.*, 2011b).
439 This presents an obvious challenge to studies of multiple species pairs or adaptive radiations
440 (Mallet *et al.*, 2016). Furthermore, stochasticity in divergence times and ILS at neutral loci can
441 generate false signals of both genomic divergence and gene flow between species pairs
442 (Barton, 2006; Pease & Hahn, 2013; Cruickshank & Hahn, 2014). Incorporating demographic
443 history in tests for selection is difficult as incorrect specification of the history, potentially
444 generated by ILS patterns, increases error rates (Lotterhos & Whitlock, 2014; Aeschbacher *et al.*
445 *et al.*, 2016; Fraïsse *et al.*, 2016a; Hoban *et al.*, 2016; Le Moan *et al.*, 2016). Approaches that do
446 not use demographic models may be preferable in some cases although these too are prone to
447 bias (Hoban *et al.*, 2016).

448
449 Speciation is undoubtedly complex, unfolding in space and time with populations overlapping,
450 contracting and re-expanding (Butlin *et al.*, 2008; Abbott *et al.*, 2013; Seehausen *et al.*, 2014).
451 This complexity suggests that most species have probably evolved with gene flow occurring at
452 some point in their evolutionary history (Smadja & Butlin, 2011) and that the process cannot
453 easily be delineated into primary vs. secondary contact or with vs. without gene flow (Bierne *et al.*
454 *et al.*, 2013; Cruickshank & Hahn, 2014). A genic perspective on speciation predicts that
455 divergence history will vary across the genome (Wu, 2001; Wu & Ting, 2004) therefore the
456 history of barrier loci might not necessarily reflect the history of populations, as *Heliconius*
457 butterflies, *Anopheles* mosquitoes and marine-freshwater sticklebacks appear to show (Bierne
458 *et al.*, 2013; Mallet *et al.*, 2016). Adaptive alleles may evolve during a period of geographical
459 isolation but introgress between divergent lineages via hybridisation and only act as barrier loci
460 in a later phase of *in-situ* divergence between populations (Bierne *et al.*, 2011, 2013). Ancient
461 divergence times for adaptive variants in several systems also suggest that these alleles are
462 maintained as standing variation and spread between populations as a result of gene flow,
463 repeatedly becoming involved in divergence (Colosimo *et al.*, 2005; Lamichhaney *et al.*, 2015;

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

471

472 **Mutation rate variation**

473 In the absence of gene flow and selection, neutral diversity within and divergence between
474 populations scales with mutation rate. In the human genome, for example, nucleotide diversity is
475 positively correlated with *de novo* mutation rate, which in turn accounts for a third of sequence
476 divergence variation between humans and chimpanzees (Francioli *et al.*, 2015). Mutation rate
477 variation amongst species, populations and individuals and the implications of this for
478 evolutionary inference are relatively well understood (Drummond *et al.*, 2006; Ho & Larson,
479 2006; Hodgkinson & Eyre-Walker, 2011). However, absolute mutation rates (i.e. the number of
480 mutations per site and generation) are also non-uniform across the genome (Hodgkinson &
481 Eyre-Walker, 2011; Ness *et al.*, 2015). Mutation probability is influenced by G:C bases and
482 neighbouring base identity (Hodgkinson & Eyre-Walker, 2011; Ness *et al.*, 2015). Replication
483 timing also has an effect, with longer exposure to mutagens during transcription in late
484 replicating regions (Hodgkinson & Eyre-Walker, 2011; Francioli *et al.*, 2015). Mutation rate is
485 often higher on Y-chromosomes than the X or autosomes because 100% of Y chromosomes
486 occur in males, experiencing higher mutation rates due to spermatogenesis (Hodgkinson &
487 Eyre-Walker, 2011). Despite knowledge of mechanisms causing mutation rate variation, it
488 remains contentious whether systematic genome-wide variation occurs at a scale that might
489 bias genome scans. For example, while Ness *et al.* (2015) detected fine-scale heterogeneity in
490 mutation rate, they found no clear variation amongst 200 kbp genome windows (Hodgkinson &
491 Eyre-Walker, 2011), suggesting that the extent of any bias in genome scans will also differ with
492 the scale of the analyses.

493

494 Irrespective of the scale at which it varies, mutation rate is an important population genetic
495 parameter used to scale estimates of parameters such as effective population size (N_e) and
496 divergence time (t) derived from genomic data. Since N_e is typically estimated from θ ($4N_e\mu$ -
497 scaled mutation rate on autosomes, where μ = absolute mutation rate), assuming a uniform

498 mutation rate will inflate estimates of N_e for mutational hotspots, obscuring the extent to which
499 drift or selection contributes to divergence in these genomic regions (Charlesworth, 2009).
500 Furthermore, given the importance of estimating demographic parameters for determining how
501 and when speciation has occurred (see *Demographic and evolutionary history*), uniform
502 mutation rates incorrectly applied across the genome may obscure the history of barrier loci and
503 the speciation process (Scally & Durbin, 2012). Mutation rate variation also has implications for
504 genomic differentiation; high mutation rate at some genomic regions may downwardly bias local
505 measures of relative differentiation, e.g. F_{ST} , obscuring loci putatively under selection (Foll &
506 Gaggiotti, 2008). Absolute divergence measures such as d_{XY} are also subject to bias due to
507 mutation rate variation; a low mutation rate will result in low levels of divergence, potentially
508 giving a false impression of constraint or introgression (Geneva *et al.*, 2015; Rosenzweig *et al.*,
509 2016).

510

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

518

519 **Background selection and selective sweeps at non-barrier loci**

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

529

530 Like the other processes described in this section, the extent of background selection varies
531 across the genome. Evidence from *Drosophila melanogaster* suggests it is highest on

532 autosomes, accounting for 58% of the observed variation in nucleotide diversity across 100 kbp
533 windows (Comeron, 2014). Simulations based on theoretical approximations show the effects of
534 background selection on patterns of diversity are greatest when deleterious mutation rate is high
535 and recombination rate is low, i.e. when linked neutral sites are unable to escape via
536 recombination from new mutations entering a population (Charlesworth *et al.*, 1993;
537 Charlesworth, 2012). Background selection should also be higher in genomic regions with a
538 high density of coding sequence, where mutations are more likely to have deleterious effects;
539 this is supported by lower diversity in these regions (Lohmueller *et al.*, 2011; Cutter & Payseur,
540 2013; Enard *et al.*, 2014). Whether or not mutations are deleterious within a coding region may
541 vary with proximity to optimum fitness on a adaptive landscape; when a population is close to
542 maximum fitness, a greater proportion of mutations will be deleterious, causing a shift away
543 from the optimum (Orr, 1998; Cutter & Payseur, 2013). On a genome-wide level, drastic
544 reductions in effective population size can limit background selection as the frequencies of new
545 deleterious mutations are more strongly influenced by drift (Charlesworth, 2012).

546

547 Despite being different processes, background and positive selection may produce similar
548 patterns of reduced intraspecific diversity and increased interspecific genomic differentiation in
549 genome scans using relative measures like F_{ST} (Noor & Bennett, 2009; Cruickshank & Hahn,
550 2014). Distinguishing between them is important in order to identify barrier loci under divergent
551 selection and rule out false positives; ideally, positive selection should be tested against a null-
552 evolutionary model that incorporates background selection (Cutter & Payseur, 2013; Comeron,
553 2014; Zeng & Corcoran, 2015; Elyashiv *et al.*, 2016). Predictive models incorporating
554 background selection are able to estimate the contribution of the process to differentiation
555 (Lohmueller *et al.*, 2011; Comeron, 2014; Zeng & Corcoran, 2015; Elyashiv *et al.*, 2016).
556 Similarly, outlier analyses and demographic inferences that account for signatures of
557 background selection are more robust, with fewer false positives (Ewing & Jensen, 2016; Huber
558 *et al.*, 2016). To-date however, only a few studies have attempted to account for background
559 selection in the context of speciation and barrier loci (e.g. Roesti *et al.*, 2013; Burri *et al.*, 2015;
560 Delmore *et al.*, 2015; Feulner *et al.*, 2015).

561

562 Global selective sweeps of universally adaptive alleles (see Box 1), i.e. those adaptive in both
563 diverging populations, may also generate signatures similar to barrier loci. Divergence history
564 may involve phases of allopatric isolation, during which universally adaptive mutations can
565 become fixed in only one subpopulation because gene flow is absent. This generates a peak of

566 differentiation that will decay with the introgression of the adaptive allele to the other
567 subpopulation when contact and gene flow are restored. However, homogenisation of allele
568 frequencies after secondary contact does not occur instantaneously, and peaks of differentiation
569 will be maintained during early phases of gene flow, potentially being misinterpreted as
570 indicating barrier loci (see Fig 1). Similar effects may occur at loci that do not contribute to
571 adaptation to the environment or speciation at all, but that are subject to sexual
572 selection, genomic conflict or drift occurring independently in geographically isolated
573 subpopulations.

574

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

588

589 **Recombination rate variation**

590 With a uniform recombination rate across the genome, the width of a genomic region of
591 differentiation surrounding a barrier locus is directly proportional to the strength of the barrier
592 effect (Barton & Bengtsson, 1986). In reality, however, recombination rate varies widely across
593 the genome of most species studied (Jensen-Seaman & Furey, 2004). This may be associated
594 with chromosome type (i.e. sex chromosomes vs. autosomes), distance to the centromere, GC
595 content, CpG motifs, transposable elements, polyA and polyT sequences, gene density and
596 recombination modifiers (Butlin, 2005; Smukowski & Noor, 2011 and references therein), or, on
597 a fine scale, with recombination hotspots (Myers *et al.*, 2010; Massy, 2013). Since many of
598 these factors are associated, determining the true cause of recombination rate variation is
599 difficult but its effects on genomic variation are more predictable. A barrier locus will influence a

600 larger genomic region when it occurs in a low-recombination region compared to a high-
601 recombination region (Stephan, 2010; Cutter & Payseur, 2013). Therefore, it might be easier to
602 detect in a genome scan, but harder to narrow down to small functional regions or individual
603 nucleotides. This alone is justification enough to account for recombination rate variation when
604 interpreting patterns of differentiation across the genome (Nachman & Payseur, 2012; Roesti *et*
605 *al.*, 2012). However, a strong correlation between recombination rate and nucleotide diversity
606 (Begun & Aquadro, 1992; Cutter & Payseur, 2013) suggests that recombination rate variation
607 can confound interpretation of the genomic landscape in other ways too.

608

609 While recombination rate has a mutagenic effect, this does not appear to be correlated with
610 genomic divergence (Noor, 2008; Charlesworth & Campos, 2014). Indeed, controlling for
611 mutation rate variation shows recombination determines the extent human-chimpanzee
612 divergence in other ways (Francioli *et al.*, 2015). Background selection reducing genetic
613 diversity in regions of low recombination is a compelling explanation for these patterns
614 (Charlesworth *et al.*, 1993; Cutter & Payseur, 2013). Neutral alleles in low recombination
615 regions are more frequently in LD with deleterious mutations and so experience a stronger
616 purging effect (Charlesworth *et al.*, 1993; Charlesworth, 2012). This leads to a reduction in
617 within-population diversity, while measures of absolute divergence (d_{XY}) remain largely
618 unaffected, provided gene flow is sufficiently low (Charlesworth *et al.*, 1997; Noor & Bennett,
619 2009; Cruickshank & Hahn, 2014; Zeng & Corcoran, 2015; but see also Phung *et al.*, 2016).
620 However, measures of relative differentiation (F_{ST}) will be inflated and some regions may appear
621 as outliers. High differentiation between species has indeed been observed in low
622 recombination regions, e.g. close to centromeres (e.g. Nachman & Payseur, 2012; Roesti *et al.*,
623 2012). Nonetheless, it remains unclear how low gene flow between populations must be for
624 background selection in recombination cold spots to cause false positive signals of
625 differentiation.

626

627 Importantly, recombination can influence selection beyond its signature in genome scans. High
628 recombination allows the independent evolution of individual selected positions, counteracting
629 Hill-Robertson interference (Stephan, 2010; Gossmann *et al.*, 2014). The more efficient fixation
630 of adaptive mutations can potentially lead to greater divergence in high recombination regions
631 (Bullaughay *et al.*, 2008). Conversely, when recombination is absent, lower efficacy of selection
632 at eliminating weakly deleterious mutations can lead to increased non-synonymous divergence
633 (Haddrill *et al.*, 2007; Bullaughay *et al.*, 2008). However, the effects of gene flow on the

634 relationship between recombination and efficacy of selection have not been studied.
635 Additionally, regions of reduced recombination may allow existing barrier loci to shield closely
636 linked, newly established barrier loci under weaker selection from stochastic loss (Rafajlovic *et*
637 *al.*, 2016). Clusters of barrier loci may be more likely to evolve in low recombination regions and
638 it is possible that recombination suppressors evolve because they enhance clustering effects
639 (Yeaman, 2013).

640

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

652

653 **Gene density**

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

668

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

685

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

697

698 The grouping of genes with related functions can also be expected to influence large-scale
699 mechanisms in the speciation process when gene flow is occurring, e.g. the evolution of
700 inversion polymorphisms or divergence hitchhiking (see *Loci linked to barrier nucleotides*).
701 Functional grouping means multiple loci affecting the same divergently selected trait or suite of

702 traits may be physically linked (Hurst *et al.*, 2004; Al-Shahrour *et al.*, 2010). Inversions are
703 mainly adaptive if they capture multiple barrier loci (Kirkpatrick & Barton, 2006; Faria & Navarro,
704 2010), and the potential for capturing multiple barrier loci in an inversion when gene flow is
705 occurring is higher if they are grouped. Divergence hitchhiking occurs when adaptive mutations
706 arise close to an established barrier locus and are shielded from gene flow. When functionally
707 related genes are closely linked, new mutations occurring in the same genomic region are more
708 likely to be adaptive than if genes are randomly distributed, this increases the potential for
709 divergence hitchhiking. Similarly, new adaptive mutations would also be better protected against
710 stochastic loss (Rafajlovic *et al.* 2016).

711

712 **Section 3: A roadmap for the genomic landscape**

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

726 With so many modifying factors, interacting in complex ways, the prospects for disentangling the
727 genomic landscape might seem bleak. We believe this conclusion premature; in this section, we
728 outline a roadmap for future research in speciation genomics to overcome the issues faced by
729 the field. Our roadmap will not be feasible in all study systems, but it should represent a
730 guideline for researchers to work with. Over the last 15 years, since the publication of Wu's
731 (2001) 'genic view', a huge number of empirical studies have provided previously unimagined
732 insight into how speciation has progressed, and this number is still increasing. We argue that
733 with a carefully considered approach, ongoing speciation research will provide us with an even
734 greater understanding of the "mystery of mysteries".

735 **Step 1: Know the study system**

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

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

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

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

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

785 However, reduced representation sequencing may not always be the ideal choice for identifying
786 barrier loci because of their relatively low genome coverage (e.g. 0.45% of 0.4 Gb three-spine
787 stickleback genome; Hohenlohe *et al.*, 2010). Markers will rarely be the direct targets of
788 selection. In low recombination regions, physical distance between barrier loci and markers that
789 are outliers is likely to be large; in high-recombination regions, barrier loci are less likely to be
790 detected in the first place as the scale of LD is small. Furthermore, these methods may bias
791 studies in favour of identifying barrier loci with single nucleotide substitutions, overlooking
792 structural variants, rearrangements and changes in genome organization that can only be
793 detected reliably by using long insert mate pair libraries (Jones *et al.*, 2012). Most importantly,
794 users should be aware of the pitfalls and biases unique to each different reduced representation
795 method that may ultimately distort the picture of the genomic landscape; e.g. null alleles and
796 sequence length bias in RAD-seq (Davey *et al.*, 2013; Gautier *et al.*, 2013; Ravinet *et al.*, 2016)
797 and bias towards conserved genic regions or overexpressed alleles in RNA-seq (Hoban *et al.*,
798 2016).

799 Whole-genome re-sequencing is becoming increasingly affordable as an alternative to reduced
800 representation approaches and has been used successfully in multiple taxa (see Table 1 for

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

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

817 Measuring factors influencing the genomic landscape is difficult, but not insurmountable.
818 Genome-wide recombination rate variation can be documented by mapping in experimental
819 crosses (Roesti *et al.*, 2013) or pedigrees (Kong *et al.*, 2002; Kawakami *et al.*, 2014). LD-based
820 methods using population genetic data are also able to estimate average realised recombination
821 across the population and over time (Tine *et al.*, 2014), which may be more relevant in the
822 landscape context (Smukowski & Noor, 2011). Whichever approach is used, high-density
823 genomic markers and large numbers of individuals are essential since it is clear that
824 recombination rate can vary on a small genomic scale (Roesti *et al.*, 2013; Kawakami *et al.*,
825 2014). Furthermore, if possible, a comparative recombination mapping approach, i.e. using all
826 taxa studied, should be taken to account for differences between closely-related species
827 (Renaut *et al.*, 2013).

828 Directly measuring genome-wide variation in mutation rate is likely to be more difficult,
829 especially in non-model organisms with long generation times. Estimates at putatively neutral
830 sites using phylogenetic methods remain valuable (Kondrashov & Kondrashov, 2010; Scally &
831 Durbin, 2012). However these estimates are prone to bias depending on the timescale over
832 which they are estimated (Ho *et al.*, 2005; Ho, 2014), and they do not incorporate deleterious or
833 weakly deleterious mutations: i.e. they are substitution, not mutation rates. If possible, whole-

834 genome sequencing within families using parent-offspring trios provides a direct measurement
835 of genome-wide mutation rate heterogeneity and also allows classification of mutations, as
836 adaptive, deleterious or neutral (Francioli *et al.*, 2015). Mutation accumulation lines offer an
837 experimental approach in lab-based populations; natural selection is reduced and mutations are
838 allowed to accumulate even if they would otherwise have negative fitness consequences (Ness
839 *et al.*, 2015).

840 Estimating gene density relies on a high quality reference genome and precise annotation; with
841 accurate annotation, gene density is relatively easy to quantify (Hurst *et al.*, 2002; Al-Shahrour
842 *et al.*, 2010). Precise genome annotation, aided with transcriptomic data, should also mean that
843 measures of gene density are feasible for most organisms following genome assembly.
844 However, greater effort needs to be made to better annotate regions that are not protein-coding
845 but still play a functional role, e.g. regulatory regions. Importantly, measuring gene density via
846 annotation may also provide insight into other confounding factors influencing the genomic
847 landscape, potentially overcoming limitations for non-model organisms. For example,
848 recombination hotspots may be predicted by identifying transposons and sequence motifs
849 recognised by recombination modifier genes (Myers *et al.*, 2010). Similarly, models using the
850 spatial distribution of CpG dinucleotides, flanking sequence and other mutation rate modifiers
851 could potentially be used to estimate mutation rate variation (Francioli *et al.*, 2015; Ness *et al.*,
852 2015).

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

854 To properly identify the signature of selection properly, controlling for factors that modify or
855 mimic the signature of barriers to gene flow is essential. Previous work has attempted to do this,
856 at least in part, e.g. removing the effects of recombination rate variation by either correcting
857 local estimates of differentiation for regional differentiation (Roesti *et al.*, 2012), correlating
858 differentiation with recombination rate (Renaut *et al.*, 2013) or focusing on barrier loci in high-
859 recombination regions (Marques *et al.*, 2016). Clearly much of the focus to-date has been on
860 recombination rate variation although mutation rate has been tentatively linked to genomic
861 differentiation using indirect measures such as synonymous divergence (d_S - Renaut *et al.*,
862 2014). Human-chimpanzee sequence divergence models incorporating both mutation and
863 recombination rate variation also show promise in partitioning these effects (Francioli *et al.*,
864 2015).

865 Ultimately, the aim should be to infer selection using models that account for variation in
866 multiple confounding factors. It is now possible to detect hard selective sweeps in a single

867 population by including fixed differences with an outgroup to account for mutation rate variation
868 and by scaling the site frequency spectrum by estimates of background selection derived from
869 mutation and recombination rate variation and genome annotation data (Huber *et al.*, 2016).
870 However, this has yet to be extended to cases of divergence with gene flow. Methods using
871 genome-wide measures of recombination rate variation and nucleotide diversity in order to
872 estimate the intensity and timing of selection and gene flow are also now available and can be
873 extended to include background selection (Aeschbacher *et al.*, 2016). Such methods can only
874 be used if independent measurements of these factors (see Step 4) are combined with genome
875 scan data. Modelling the genomic landscape with local estimates of recombination rate,
876 mutation rate and gene density, can then be used to ask whether we need to invoke divergent
877 selection and gene flow to explain peaks of high differentiation (Cruickshank & Hahn, 2014).

878 Systems of parallel divergence or speciation may also be helpful in separating the effects of
879 various factors (Irwin *et al.*, 2016). For example, when recombination rate variation is correlated
880 among closely related taxa, high differentiation in low-recombination regions that appear in
881 multiple species pairs is more likely to have arisen due to background selection (Burri *et al.*,
882 2015). This is especially true if contrasts involve different types of barriers to gene flow, and if
883 the same highly differentiated regions occur in comparisons with and without gene flow.
884 However, as a caveat, differentiated regions shared amongst contrasts may sometimes still be
885 due to the same loci under divergent selection. Disentangling these explanations is only
886 possible with information on gene density, mutation rate, the types of barriers involved, and the
887 history of gene flow. Nonetheless, even with these data we can still only identify candidate
888 barrier regions: experimental and functional approaches are necessary to identify barrier loci
889 unequivocally.

890 **Step 6: Independent evidence for barrier loci**

891 Crucially, genomic data alone cannot provide conclusive evidence of barrier loci. Disentangling
892 effects is difficult precisely because some modifying factors (e.g. demographic history) are
893 estimated from data used to measure the landscape of differentiation. Even with good genomic
894 evidence of selection on a candidate region, other processes, such as local adaptation following
895 or unrelated to speciation, can be invoked (Cruickshank & Hahn, 2014). For this reason, the
896 search for evidence of selection should extend beyond the genome scan. In principle, there are
897 two ways of obtaining independent evidence for selection; we can either directly test for
898 signatures of selection on a given locus; or we can test for a link between the genotype and the
899 phenotype, and separately test for selection on the phenotype (Table 2). The advantage of the

900 former is that it provides a more direct test of selection; the advantage of the latter is that
901 knowing the associated phenotypic change allows for a complete "story" and a better
902 understanding of the system.

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

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

922 Nonetheless, the majority of these approaches stop short of directly demonstrating how a
923 barrier allele alters the function to produce phenotypic consequences and ultimately results in
924 reproductive isolation (Seehausen *et al.*, 2014). In some cases, molecular assays of protein
925 function are possible; but often conclusive evidence is only really possible using transgenic or
926 gene interference methods which to-date have largely been limited to model organisms such as
927 *Drosophila* (Thomae *et al.*, 2013; Satyaki *et al.*, 2014; Phadnis *et al.*, 2015). With the rapid
928 adoption of CRISPR, a method applicable to a much wider range of organisms, transgenic
929 experiments are likely to become an important part of speciation research (Bono *et al.*, 2015).
930 Gene insertion, knockouts and reciprocal transplant experiments, for example, will be able to
931 provide direct evidence of barrier nucleotide function in non-model organisms (Bono *et al.*,
932 2015).

933 **Concluding remarks**

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

949

950 **Box 1 – Clearer definitions**

951 A wealth of technical terms, often without clear definition, makes an attempt to understand the
952 literature on speciation genomics a daunting task (Harrison, 2012). In this review, we argue for
953 the importance of identifying **barrier loci**, positions in the genome that contribute to barriers to
954 gene flow between populations. These include loci under divergent ecological selection, but also
955 loci involved in other barriers, e.g. mate choice, or intrinsic postzygotic isolation. When a **locally**
956 **beneficial allele**, adaptive in a single population, arises, divergent positive selection will cause
957 it to increase in frequency, resulting in a **local selective sweep**. **The barrier effect** is a
958 reduction of effective migration rate relative to actual migration between populations that occurs
959 at the barrier locus (i.e. the direct effect) but can also extend beyond it (i.e. the indirect effect). In
960 surrounding genomic regions, the barrier effect will initially allow a build-up of **genomic**
961 **differentiation**, i.e. a difference in allele frequency, between populations, typically documented
962 using a relative measure such as F_{ST} . Over time, the barrier effect will allow neutral mutations to
963 establish, resulting in **genomic divergence** between populations, typically measured using d_{XY} .
964 In contrast to barrier loci, when **globally beneficial alleles** arise they will increase in frequency

965 due to positive (but crucially, not divergent) selection and spread amongst populations in
966 contact. Both globally and locally adaptive alleles may undergo **hard sweeps**, i.e. from de novo
967 mutation or introgression, or **soft sweeps**, i.e. from standing genetic variation. **Genome scans**,
968 comparisons between pairs of populations or species at multiple loci across the genome
969 (typically thousands of loci nowadays), can quantify the genomic landscape of differentiation
970 and divergence when placed on a physical or genetic map. These are used to identify **outliers**,
971 i.e. loci or regions that fall outside the expected equilibrium neutral distribution of differentiation
972 or divergence, which may be influenced by barrier effects.

973

974 **Box 2 – Searching for islands in a sea of metaphors**

975 Genomic differentiation may be heterogeneous during much of the speciation process (Nosil
976 2012; Table 1). Under the genic view of speciation, the genome is porous to gene flow while
977 reproductive isolation is incomplete (Wu, 2001; Wu & Ting, 2004). A large number of genome
978 scans have identified distinct genomic regions (“islands”) of greater differentiation than the
979 putatively neutral genomic background (“sea level”) that tends toward homogenization by gene
980 flow (Nosil *et al.*, 2009). First described as “genomic islands of speciation” in *Anopheles*
981 mosquitoes, these regions were assumed to harbour loci underlying reproductive isolation
982 (Turner *et al.*, 2005). The genomic island metaphor has proved popular and has been valuable
983 for driving empirical progress; a wide array of studies searching for “speciation islands” in
984 multiple taxa has been published in the last decade.

985

986 Other terms have also been coined to describe genomic heterogeneity. These may not explicitly
987 invoke speciation, e.g. “genomic islands of differentiation” (Harr, 2006) or “genomic islands of
988 divergence” (Nosil *et al.*, 2009). Large differentiation regions, potentially containing multiple
989 speciation genes have been referred to as “continents of divergence” (Michel *et al.*, 2010; Egan
990 *et al.*, 2015). These metaphors have led to conceptual frameworks, such as Feder *et al.*'s
991 (2012) four-phase model, which incorporate processes such as divergence and genome
992 hitchhiking (see main text) to explain how differentiation across the genome evolves. Although
993 the metaphors have proved useful for describing observed patterns and communicating a
994 complex concept to a wider audience, introducing attractive terminology runs the risk of
995 encouraging ambiguity (Harrison, 2012). For example, differentiation is more likely to vary
996 continuously during speciation rather than showing clearly defined “islands” or “continents”.
997 Metaphors also lead to arbitrary and unproductive discussions on how to define them: what

998 level of differentiation defines an island and when or at what length does an “island” become a
999 “continent”?

1000

1001 Although genomic regions of high differentiation undoubtedly exist (Table 1), they are not
1002 necessarily caused by the interplay between gene flow and divergent selection; they may in fact
1003 be “incidental islands” that emerge when gene flow is absent (Noor & Bennett, 2009; Turner &
1004 Hahn, 2010; Cruickshank & Hahn, 2014). Divergent and indirect selection (i.e. hitchhiking and
1005 background selection) can reduce within-population diversity in geographically isolated and
1006 potentially locally adapted populations, leading to high- F_{ST} regions that may not be related to
1007 speciation, while much of the genome remains undifferentiated due to incomplete lineage
1008 sorting. This process results in a specific genomic signature with high levels of differentiation
1009 (F_{ST} , a relative measure) and low levels of absolute divergence (d_{XY}) at loci affected by local
1010 adaptation or background selection. In this case, divergence due to direct and indirect selection
1011 occurs in the absence of gene flow, potentially after speciation is completed or even just while
1012 local adaptation is occurring. It is necessary to rule out this alternative explanation before
1013 interpreting regions of elevated differentiation as barrier loci. For that, it is crucial to test for
1014 ongoing or recent gene flow (Box 3).

1015

1016 Importantly, even if gene flow does occur, elevated divergence/differentiation alone is not
1017 sufficient to identify barrier loci; additional evidence is necessary (see Roadmap). Given that
1018 “islands” may not be involved in speciation at all, we suggest avoiding any terminology linking
1019 highly differentiated genomic regions to speciation unless further evidence suggests this is, in
1020 fact, the case.

1021

1022 **Box 3 – Inferring and quantifying gene flow**

1023 Barrier effects can only be detected in the presence of recent or ongoing gene flow. Inferring
1024 gene flow, outside of a genome scan and preferably in the context of evolutionary history, is an
1025 important first step for interpreting the genomic landscape of speciation. However, given the
1026 complexity of speciation history and the high probability that, in many cases, gene flow is not
1027 constant over time, this presents a major difficulty for speciation research.

1028

1029 Identifying recent gene flow using population clustering methods that reliably detect F1, F2 and
1030 backcross hybrids is relatively straightforward (Pritchard *et al.*, 2000; Anderson & Thompson,

1031 2002; Falush *et al.*, 2003). Emphasis should be placed on identifying introgression over several
1032 generations: i.e. on the presence of backcrossed individuals. Clinal analysis of allele
1033 frequencies across hybrid zones or across the genome overcomes a significant current
1034 disadvantage of clustering techniques as it allows for reliable migration estimates (Barton, 1983;
1035 Barton & Hewitt, 1985; Gompert & Buerkle, 2011). Other evidence for recent or ongoing gene
1036 flow makes use of the biogeographical distributions of species, e.g. asking if genetic
1037 differentiation is lower in sympatry compared to allopatry (Noor & Bennett, 2009; Marques *et al.*,
1038 2016). For example, *Heliconius* butterfly studies show greater divergence between allopatric
1039 races than between those in sympatry or parapatry, suggesting ongoing gene flow (Nadeau *et al.*,
1040 2012, 2013; Martin *et al.*, 2013). Similarly, very recently diverged populations (i.e. hundreds
1041 of generations) with documented hybridization events suggest low genomic differentiation is
1042 maintained, at least in part, by gene flow (Lescak *et al.*, 2015; Marques *et al.*, 2016). Finally,
1043 non-genetic evidence of migration or potential migration between populations using mark-
1044 recapture experiments (Bolnick *et al.*, 2009), mate-choice experiments (Nosil *et al.*, 2002;
1045 McKinnon *et al.*, 2004) and phenotypic variation (Lescak *et al.*, 2015) can bolster the argument
1046 that low background differentiation in a genome scan is due to ongoing gene flow.

1047
1048 Several key approaches incorporate demographic history, making it possible to infer both gene
1049 flow and mechanisms of divergence (Sousa & Hey, 2013). Site frequency spectrum (SFS)
1050 methods can rapidly approximate the joint allele frequency distribution between populations,
1051 allowing comparisons of divergence with and without gene flow and the estimation of migration
1052 rate (Gutenkunst *et al.*, 2009; Excoffier *et al.*, 2013). Isolation-with-Migration (IM) models have
1053 also recently been extended to incorporate whole-genome data and overcome some simplifying
1054 assumptions such as absence of recombination (Hobolth *et al.*, 2011a; Mailund *et al.*, 2012).
1055 Approximate Bayesian Computation (ABC) is more computationally expensive but can
1056 incorporate thousands of loci resulting in high precision parameter estimation (Robinson *et al.*,
1057 2014; Shafer *et al.*, 2015). ABC is flexible, allowing variation in migration rates amongst loci to
1058 be incorporated (Roux *et al.*, 2013, 2014) or the inclusion of haplotype-based statistics for
1059 estimating gene flow (Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010). However, we note that model-
1060 based inference is limited to distinguishing amongst the models tested; parameter estimates are
1061 therefore meaningful only in the context of a specific model. Since these are generally
1062 simplifications, the results must be interpreted with caution.

1063

1064 Modelling approaches typically perform poorly when estimating gene flow timing (Roux *et al.*,
1065 2013), but this may be possible to overcome when there is sufficient biogeographical and
1066 phylogenetic information to resolve periods of contact between populations (Garrigan *et al.*,
1067 2012; Nadachowska-Brzyska *et al.*, 2013). This is the rationale behind comparative statistics
1068 such as ABBA-BABA that test for an excess of derived alleles at positions across the genome
1069 (Green *et al.*, 2010; Durand *et al.*, 2011; Martin *et al.*, 2014). By incorporating different taxa with
1070 known divergence times, it is possible to infer the time interval when introgression may have
1071 occurred (Martin *et al.*, 2013; Eaton *et al.*, 2015). Methods comparing the size of introgressed
1072 haplotypes ('migrant tracts') to an expected distribution under migration within T generations
1073 may provide relatively accurate estimates of the timing of gene flow (Pool & Nielsen, 2009).
1074 However, this requires accurate haplotype phasing and has very little power to date admixture
1075 more than 1000 generations in the past (Pool & Nielsen, 2009; Liang & Nielsen, 2014). Identity-
1076 by-state tracts, i.e. the distance between polymorphisms on a haplotype, also require phased
1077 data to calculate but provide a promising means for estimating the timing and extent of gene
1078 flow, as well as other demographic parameters (Harris & Nielsen, 2013). An extension of the
1079 Markov coalescent approach for estimating effective population size as a function of time can
1080 now use haplotype data from multiple individuals to determine cross-coalescence rate (i.e.
1081 coalescent events within and between populations) providing accurate estimates of the timing
1082 and rate of last migration without a specified demographic model (Li & Durbin, 2011; Schiffels &
1083 Durbin, 2014).

1084

1085 **Acknowledgments**

1086 Many of the ideas for this review were first formulated from discussions between co-authors
1087 during our symposium on "The genomic landscape of speciation" at ESEB 2015, Lausanne,
1088 Switzerland. We were kindly sponsored by Floragenex, Oregon, USA, and also by Stab Vida,
1089 Portugal. Mark Ravinet was funded by a JSPS Postdoctoral Fellowship for Foreign Researchers.
1090 RF was funded by FCT under the Programa Operacional Potencial Humano – Quadro de
1091 Referência Estratégico Nacional from the European Social Fund and the Portuguese Ministério
1092 da Educação e Ciência through the postdoctoral fellowship SFRH/BPD/89313/2012 and
1093 research project PTDC/BIA-EVF/113805/2009 and FCOMP-01-0124-FEDER-014272. JG was
1094 funded by a postdoctoral fellowship from Xunta de Galicia (Modalidade B). AMW and RKB are
1095 funded by NERC.

1096

1097 **References**

- 1098 Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., *et al.* 2013.
1099 Hybridization and speciation. *J. Evol. Biol.* **26**: 229–46.
- 1100 Aeschbacher, S., Selby, J.P., Willis, J.H. & Coop, G. 2016. Population-genomic inference of the
1101 strength and timing of selection against gene flow. *bioRxiv* 1–21.
- 1102 Al-Shahrour, F., Minguéz, P., Marques-Bonet, T., Gazave, E., Navarro, A. & Dopazo, J. 2010.
1103 Selection upon genome architecture: Conservation of functional neighborhoods with
1104 changing genes. *PLoS Comput. Biol.* **6**.
- 1105 Anderson, E.C. & Thompson, E.A. 2002. A model-based method for identifying species hybrids
1106 using multilocus genetic data. *Genetics* **160**: 1217–1229.
- 1107 Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. & Hohenlohe, P.A. 2016. Harnessing the
1108 power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **in press**.
1109 Nature Publishing Group.
- 1110 Arnegard, M.E., McGee, M.D., Matthews, B., Marchinko, K.B., Conte, G.L., Kabir, S., *et al.*
1111 2014. Genetics of ecological divergence during speciation. *Nature*, doi:
1112 10.1038/nature13301. Nature Publishing Group.
- 1113 Barrett, R.D.H. & Hoekstra, H.E. 2011. Molecular spandrels: tests of adaptation at the genetic
1114 level. *Nat. Rev. Genet.* **12**: 767–80.
- 1115 Barrett, R.D.H., Rogers, S.M. & Schluter, D. 2008. Natural selection on a major armor gene in
1116 threespine stickleback. *Science (80-.)*. **322**: 255–257.
- 1117 Barton, N. & Bengtsson, B.O. 1986. The barrier to genetic exchange between hybridising
1118 populations. *Heredity (Edinb)*. **57**: 357–376.
- 1119 Barton, N. & Hewitt, G. 1985. Analysis of Hybrid Zones. *Annu. Rev. Ecol. Syst.* **16**: 113–148.
- 1120 Barton, N.H. 2006. Evolutionary Biology: How did the human species form? *Curr. Biol.* **16**: 648–
1121 650.
- 1122 Barton, N.H. 1983. Multilocus Clines. *Evolution (N. Y)*. **37**: 454–471.
- 1123 Barton, N.H. & de Cara, M.A.R. 2009. The evolution of strong reproductive isolation. *Evolution*
1124 **63**: 1171–90.
- 1125 Begun, D.J. & Aquadro, C.F. 1992. Levels of naturally occurring DNA polymorphism correlate
1126 with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- 1127 Berner, D., Moser, D., Roesti, M., Buescher, H. & Salzburger, W. 2014. Genetic architecture of
1128 skeletal evolution in European lake and stream stickleback. *Evolution* **68**: 1792–805.
- 1129 Bertorelle, G., Benazzo, A. & Mona, S. 2010. ABC as a flexible framework to estimate
1130 demography over space and time: some cons, many pros. *Mol. Ecol.* **19**: 2609–2625.

- 1131 Bierne, N. 2010. The distinctive footprints of local hitchhiking in a varied environment and global
1132 hitchhiking in a subdivided population. *Evolution (N. Y.)* **64**: 3254–3272.
- 1133 Bierne, N., Gagnaire, P. & David, P. 2013. The geography of introgression in a patchy
1134 environment and the thorn in the side of ecological speciation. **59**: 72–86.
- 1135 Bierne, N., Welch, J., Loire, E., Bonhomme, F. & David, P. 2011. The coupling hypothesis: why
1136 genome scans may fail to map local adaptation genes. *Mol. Ecol.* **20**: 2044–72.
- 1137 Bilandžija, H., Ma, L., Parkhurst, A. & Jeffery, W.R. 2013. A potential benefit of albinism in
1138 *Astyanax* cavefish: Downregulation of the *oca2* gene increases tyrosine and catecholamine
1139 levels as an alternative to melanin synthesis. *PLoS One* **8**: 1–14.
- 1140 Bolnick, D.I., Snowberg, L.K., Patenia, C., Stutz, W.E., Ingram, T. & Lau, O.L. 2009. Phenotype-
1141 dependent native habitat preference facilitates divergence between parapatric lake and
1142 stream stickleback. *Evolution (N. Y.)* **63–8**: 2004–2016.
- 1143 Bono, J.M., Olesnick, E.C. & Matzkin, L.M. 2015. Connecting genotypes, phenotypes and
1144 fitness: Harnessing the power of CRISPR/Cas9 genome editing. *Mol. Ecol.* **24**: 3810–3822.
- 1145 Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., *et al.* 2014. The genomic
1146 substrate for adaptive radiation in African cichlid fish. *Nature* **513**: 375–381.
- 1147 Bullaughey, K., Przeworski, M. & Coop, G. 2008. No effect of recombination on the efficacy of
1148 natural selection in primates. *Genome Res.* **18**: 544–554.
- 1149 Burri, R., Nater, A., Kawakami, T., Mugal, C.F., Olason, P.I., Smeds, L., *et al.* 2015. Linked
1150 selection and recombination rate variation drive the evolution of the genomic landscape of
1151 differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**:
1152 1656–1665.
- 1153 Butlin, R. 1987. Speciation by reinforcement. *Trends Ecol. Evol.* **2**: 8–13.
- 1154 Butlin, R.K. 2005. Recombination and speciation. *Mol. Ecol.* **14**: 2621–2635.
- 1155 Butlin, R.K., Galindo, J. & Grahame, J.W. 2008. Sympatric, parapatric or allopatric: the most
1156 important way to classify speciation? *Philos. Trans. R. Soc. London Ser. B* **363**: 2997–
1157 3007.
- 1158 Butlin, R.K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., *et al.* 2014. Parallel
1159 evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*
1160 **68**: 935–49.
- 1161 Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F.W., Alfoldi, J., Barrio, a. M., *et al.* 2014.
1162 Rabbit genome analysis reveals a polygenic basis for phenotypic change during
1163 domestication. *Science (80-.)* **345**: 1074–1079.
- 1164 Chan, Y.F., Marks, M.E., Jones, F.C., Villareal, G., Shapiro, M.D., Brady, S.D., *et al.* 2010.

1165 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1
1166 enhancer. *Science (80-.)*. **327**: 302–305.

1167 Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and
1168 variation. *Nat. Rev. Genet.* **10**: 195–205.

1169 Charlesworth, B. 2012. The role of background selection in shaping patterns of molecular
1170 evolution and variation: Evidence from variability on the Drosophila X chromosome.
1171 *Genetics* **191**: 233–246.

1172 Charlesworth, B. & Campos, J.L. 2014. The relations between recombination rate and patterns
1173 of molecular variation and evolution in Drosophila. *Annu. Rev. Genet.* **48**: 383–403.

1174 Charlesworth, B. & Charlesworth, D. 2010. *Elements of Evolutionary Genetics*. Roberts &
1175 Company Publishers, Greenwood Village, Colorado.

1176 Charlesworth, B., Charlesworth, D. & Barton, N.H. 2003. The effects of genetic and geographic
1177 structure on neutral variation. *Annu. Rev. Ecol. Syst.* **34**: 99–125.

1178 Charlesworth, B., Morgan, M.T. & Charlesworth, D. 1993. The effect of deleterious mutations on
1179 neutral molecular variation. *Genetics* **134**: 1289–303.

1180 Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997. The effects of local selection,
1181 balanced polymorphism and background selection on equilibrium patterns of genetic
1182 diversity in subdivided populations. *Genet. Res.* **70**: 155–74.

1183 Christe, C., Stölting, K.N., Paris, M., Fraïsse, C., Bierne, N. & Lexer, C. 2016. Adaptive
1184 evolution and segregating load contribute to the genomic landscape of divergence in two
1185 tree species connected by episodic gene flow. *Mol. Ecol.* **in press**.

1186 Colosimo, P.F., Hoseman, K.E., Balabhadra, S., Villareal, G., Dickson, M., Grimwood, J., *et al.*
1187 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin
1188 alleles. *Science (80-.)*. **307**: 1928–1933.

1189 Comeron, J.M. 2014. Background Selection as Baseline for Nucleotide Variation across the
1190 Drosophila Genome. *PLoS Genet.* **10**.

1191 Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., *et al.* 2013. Multiplex Genome
1192 Engineering Using CRISPR/Cas System. *Science (80-.)*. **339**: 819–824.

1193 Corbett-Detig, R., Zhou, J. & Clark, A. 2013. Genetic incompatibilities are widespread within
1194 species. *Nature* **504**: 135–137. Nature Publishing Group.

1195 Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sinauer, New York.

1196 Cruickshank, T.E. & Hahn, M.W. 2014. Reanalysis suggests that genomic islands of speciation
1197 are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**: 3133–3157.

1198 Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O., Csillery, K. & François, O. 2010.

1199 Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25**: 410–418.

1200 Cutter, A.D. & Payseur, B. a. 2013. Genomic signatures of selection at linked sites: unifying the
1201 disparity among species. *Nat. Rev. Genet.* **14**: 262–274. Nature Publishing Group.

1202 Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L. 2013. Special
1203 features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* **22**: 3151–64.

1204 Davey, J.W., Hohenlohe, P. a, Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L. 2011.
1205 Genome-wide genetic marker discovery and genotyping using next-generation sequencing.
1206 *Nat. Rev. Genet.* **12**: 499–510. Nature Publishing Group.

1207 Delmore, K.E., Hübner, S., Kane, N.C., Schuster, R., Andrew, R.L., Câmara, F., *et al.* 2015.
1208 Genomic analysis of a migratory divide reveals candidate genes for migration and
1209 implicates selective sweeps in generating islands of differentiation. *Mol. Ecol.* **24**: 1873–
1210 1888.

1211 Drummond, A.J., Ho, S.Y.W., Phillips, M. & Rambaut, A. 2006. Relaxed phylogenetics and
1212 dating with confidence. *Plos Biol.* **4**: e88.

1213 Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. 2011. Testing for ancient admixture
1214 between closely related populations. *Mol. Biol. Evol.* **28**: 2239–52.

1215 Duret, L. & Arndt, P.F. 2008. The impact of recombination on nucleotide substitutions in the
1216 human genome. *PLoS Genet.* **4**.

1217 Dutheil, J.Y. & Hobolth, A. 2012. Ancestral Population Genomics. In: *Evolutionary Genomics:*
1218 *Statistical and Computational Methods. Volume 2* (M. Anisimova, ed). Springer, New York,
1219 USA.

1220 Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A. & Cavender-Bares, J. 2015. Historical
1221 introgression among the American live oaks and the comparative nature of tests for
1222 introgression. *Evolution (N. Y.)*. **69**: 2587–2601.

1223 Egan, S.P., Ragland, G.J., Assour, L., Powell, T.H.Q., Hood, G.R., Emrich, S., *et al.* 2015.
1224 Experimental evidence of genome-wide impact of ecological selection during early stages
1225 of speciation-with-gene-flow. *Ecol. Lett.* **18**: 817–825.

1226 Elyashiv, E., Sattath, S., Hu, T.T., Strustovsky, A., McVicker, G., Andolfatto, P., *et al.* 2016. A
1227 genomic map of the effects of linked selection in *Drosophila*. *PLoS Genet.* 1–38.

1228 Enard, D., Messer, P.W. & Petrov, D. a. 2014. Genome-wide signals of positive selection in
1229 human evolution. *Genome Res.* **24**: 885–95.

1230 English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., *et al.* 2012. Mind the Gap:
1231 Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology.
1232 *PLoS One* **7**: e47768.

1233 Ewing, G.B. & Jensen, J.D. 2016. The consequences of not accounting for background
1234 selection in demographic inference. *Mol. Ecol.* **25**: 135–141.

1235 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. 2013. Robust
1236 demographic inference from genomic and SNP data. *PLoS Genet.* **9**: e1003905.

1237 Eyre-Walker, A. & Keightley, P.D. 2007. The distribution of fitness effects of new mutations. *Nat*
1238 *Rev Genet.* **8**: 610–8.

1239 Falush, D., Stephens, M. & Pritchard, J.K. 2003. Inference of population structure: Extensions to
1240 linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

1241 Faria, R. & Navarro, A. 2010. Chromosomal speciation revisited: rearranging theory with pieces
1242 of evidence. *Trends Ecol. Evol.* **25**: 660–9. Elsevier Ltd.

1243 Feder, J.L., Egan, S.P. & Nosil, P. 2012a. The genomics of speciation-with-gene-flow. *Trends*
1244 *Genet.* 1–9. Elsevier Ltd.

1245 Feder, J.L., Flaxman, S.M., Egan, S.P., Comeault, A. a. & Nosil, P. 2013. Geographic Mode of
1246 Speciation and Genomic Divergence. *Annu. Rev. Ecol. Evol. Syst.* **44**: 73–97.

1247 Feder, J.L., Gejji, R., Yeaman, S. & Nosil, P. 2012b. Establishment of new mutations under
1248 divergence and genome hitchhiking. *Philos. Trans. R. Soc. B Biol. Sci.* **367**: 461–474.

1249 Feder, J.L. & Nosil, P. 2010. The efficacy of divergence hitchhiking in generating genomic
1250 islands during ecological speciation. *Evolution (N. Y.)*. **64–5**: 1729–1747.

1251 Ferchaud, A.L. & Hansen, M.M. 2016. The impact of selection, gene flow and demographic
1252 history on heterogeneous genomic divergence: Three-spine sticklebacks in divergent
1253 environments. *Mol. Ecol.* **25**: 238–259.

1254 Feulner, P.G.D., Chain, F.J.J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., *et al.* 2015.
1255 Genomics of Divergence along a Continuum of Parapatric Population Differentiation. *PLOS*
1256 *Genet.* **11**: e1004966.

1257 Flaxman, S.M., Wacholder, A.C., Feder, J.L. & Nosil, P. 2014. Theoretical models of the
1258 influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* 4074–4088.

1259 Flowers, J.M., Molina, J., Rubinstein, S., Huang, P., Schaal, B.A. & Purugganan, M.D. 2012.
1260 Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in
1261 wild and domesticated rice. *Mol. Biol. Evol.* **29**: 675–687.

1262 Foll, M. & Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for
1263 both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–93.

1264 Fraïsse, C., Belkhir, K., Welch, J.J. & Bierne, N. 2016a. Local interspecies introgression is the
1265 main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* **25**: 269–
1266 286.

1267 Fraïsse, C., Belkhir, K., Welch, J.J. & Bierne, N. 2016b. Local interspecies introgression is the
1268 main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* **25**: 269–
1269 286.

1270 Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., *et al.* 2015.
1271 Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**:
1272 822–826. Nature Publishing Group.

1273 Gagnaire, P.A., Pavey, S.A., Normandeau, E. & Bernatchez, L. 2013. The genetic architecture
1274 of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs
1275 assessed by rad sequencing. *Evolution (N. Y.)*. **67**: 2483–2497.

1276 Garrigan, D., Kingan, S.B., Geneva, A.J., Andolfatto, P., Clark, A.G., Thornton, K.R., *et al.* 2012.
1277 Genome sequencing reveals complex speciation in the *Drosophila simulans* clade.
1278 *Genome Res.* **22**: 1499–511.

1279 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., *et al.* 2013. The effect
1280 of RAD allele dropout on the estimation of genetic variation within and between
1281 populations. *Mol. Ecol.* 3165–3178.

1282 Geneva, A.J., Kingan, S.B. & Garrigan, D. 2015. A new method to scan genomes for
1283 introgression in a secondary contact model. *PLoS One* 1–15.

1284 Glazer, A.M., Killingbeck, E.E., Mitros, T., Rokhsar, D.S. & Miller, C.T. 2015. Genome assembly
1285 improvement and mapping convergently evolved skeletal traits in sticklebacks with
1286 genotyping-by-sequencing. *G3* **5**: 1463–72.

1287 Gompert, Z. & Buerkle, C.A. 2011. Bayesian estimation of genomic clines. *Mol. Ecol.* **20**: 2111–
1288 27.

1289 Gompert, Z., Comeault, A.A., Farkas, T.E., Feder, J.L., Parchman, T.L., Buerkle, C.A., *et al.*
1290 2014. Experimental evidence for ecological selection on genome variation in the wild. *Ecol.*
1291 *Lett.* **17**: 369–379.

1292 Gossmann, T.I., Santure, A.W., Sheldon, B.C., Slate, J. & Zeng, K. 2014. Highly variable
1293 recombinational landscape modulates efficacy of natural selection in birds. *Genome Biol.*
1294 *Evol.* **6**: 2061–2075.

1295 Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., *et al.* 2010. A draft
1296 sequence of the Neandertal genome. *Science* **328**: 710–22.

1297 Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009. Inferring the
1298 joint demographic history of multiple populations from multidimensional SNP frequency
1299 data. *PLoS Genet.* **5**: e1000695.

1300 Haddrill, P.R., Halligan, D.L., Tomaras, D. & Charlesworth, B. 2007. Reduced efficacy of

1301 selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8**:
1302 R18.

1303 Haldane, J.B.S. 1930. A mathematical theory of natural and artificial selection-I. *Math. Proc.*
1304 *Cambridge Philos. Soc.* **26**: 220–230.

1305 Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome*
1306 *Res.* **16**: 730–737.

1307 Harris, K. & Nielsen, R. 2013. Inferring demographic history from a spectrum of shared
1308 haplotype lengths. *PLoS Genet.* **9**: e1003521.

1309 Harrison, R.G. 1990. Hybrid zones: windows on evolutionary processes. In: *Oxford surveys in*
1310 *evolutionary biology*, pp. 69–128.

1311 Harrison, R.G. 2012. *The language of speciation*.

1312 Harrison, R.G. & Larson, E.L. 2016. Heterogeneous genome divergence, differential
1313 introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* **25**: 2454–2466.

1314 Hartl, D.L. & Clark, A.G. 2007. *Principles of population genetics*. Sinauer Associates,
1315 Sunderland, Massachusetts, USA.

1316 Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V., *et al.* 2010. The
1317 genome of the Western clawed frog *Xenopus tropicalis*. *Science* **328**: 633–6.

1318 Hermisson, J. & Pennings, P.S. 2005. Soft sweeps: molecular population genetics of adaptation
1319 from standing genetic variation. *Genetics* **169**: 2335–52.

1320 Hey, J. & Kliman, R.M. 2002. Interactions between natural selection, recombination and gene
1321 density in the genes of *Drosophila*. *Genetics* **160**: 595–608.

1322 Ho, S.Y.W. 2014. The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.* **29**:
1323 496–503. Elsevier Ltd.

1324 Ho, S.Y.W. & Larson, G. 2006. Molecular clocks: when times are a-changin'. *Trends Ecol. Evol.*
1325 **22**: 79–83.

1326 Ho, S.Y.W., Phillips, M.J., Cooper, A. & Drummond, A.J. 2005. Time dependency of molecular
1327 rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.*
1328 **22**: 1561–1568.

1329 Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., *et al.* 2016.
1330 Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future
1331 Directions. *Am. Nat.* **188**: 000–000.

1332 Hobolth, A., Andersen, L.N. & Mailund, T. 2011a. On computing the coalescence time density in
1333 an isolation-with-migration model with few samples. *Genetics* **187**: 1241–3.

1334 Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H. & Mailund, T. 2011b. Incomplete lineage

1335 sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan
1336 speciation and widespread selection. *Genome Res.* **21**: 349–56.

1337 Hodgkinson, A. & Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian
1338 genomes. *Nat. Rev. Genet.* **12**: 756–766. Nature Publishing Group.

1339 Hoekstra, H.E., Hirschmann, R.J., Bunday, R.A., Insel, P. a & Crossland, J.P. 2006. A single
1340 amino acid mutation contributes to adaptive beach mouse color pattern. *Science (80-.)*.
1341 **313**: 101–104.

1342 Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. 2010.
1343 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD
1344 Tags. *PLoS Genet.* **6**: e1000862.

1345 Huber, C.D., DeGiorgio, M., Hellmann, I. & Nielsen, R. 2016. Detecting recent selective sweeps
1346 while controlling for mutation rate and background selection. *Mol. Ecol.* **25**: 142–156.

1347 Hurst, L.D., Pál, C. & Lercher, M.J. 2004. The evolutionary dynamics of eukaryotic gene order.
1348 *Nat. Rev. Genet.* **5**: 299–310.

1349 Hurst, L.D., Williams, E.J.B. & Pal, C. 2002. Natural selection promotes the conservation of
1350 linkage of co-expressed genes. *Trends Genet.* **18**: 604–606.

1351 Irwin, D.E., Alcaide, M., Delmore, K.E., Irwin, J.H. & Owens, G.L. 2016. Recurrent selection
1352 explains genomic regions of high relative but low absolute differentiation in the greenish
1353 warbler ring species. *Mol. Ecol.* **25**: 4488–4507.

1354 Jensen-Seaman, M. & Furey, T. 2004. Comparative recombination rates in the rat, mouse, and
1355 human genomes. *Genome Res.* 528–538.

1356 Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., *et al.* 2012. The
1357 genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

1358 Kawakami, T., Smeds, L., Backstrom, N., Husby, A., Qvarnstrom, A., Mugal, C.F., *et al.* 2014. A
1359 high-density linkage map enables a second-generation collared flycatcher genome
1360 assembly and reveals the patterns of avian recombination rate variation and chromosomal
1361 evolution. *Mol. Ecol.* **23**: 4035–4058.

1362 Keller, I., Wagner, C.E., Greuter, L., Mwaiko, S., Selz, O.M., Sivasundar, a, *et al.* 2012.
1363 Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in
1364 the rapid radiation of Lake Victoria cichlid fishes. *Mol. Ecol.* 2848–2863.

1365 Kim, N. & Jinks-Robertson, S. 2012. Transcription as a source of genome instability. *Nat. Rev.*
1366 *Genet.* **13**: 204–214. Nature Publishing Group.

1367 Kirkpatrick, M. & Barton, N. 2006. Chromosome inversions, local adaptation and speciation.
1368 *Genetics* **173**: 419–34.

1369 Kitano, J., Ross, J.A., Mori, S., Kume, M., Jones, F.C., Chan, Y.F., *et al.* 2009. A role for neo-
1370 sex chromosomes in stickleback speciation. *Nature* **461**: 1079–1083.

1371 Knowles, L.L. 2009. Statistical Phylogeography. *Annu. Rev. Ecol. Syst.* **40**: 593–612.

1372 Knowles, L.L. & Maddison, W.P. 2002. Statistical phylogeography. *Mol. Ecol.* **11**: 2623–2635.

1373 Kondrashov, F.A. & Kondrashov, A.S. 2010. Measurements of spontaneous rates of mutations
1374 in the recent past and the near future. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365**: 1169–
1375 76.

1376 Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., *et*
1377 *al.* 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31**:
1378 241–7.

1379 Kruuk, L.E., Baird, S.J., Gale, K.S. & Barton, N.H. 1999. A comparison of multilocus clines
1380 maintained by environmental adaptation or by selection against hybrids. *Genetics* **153**:
1381 1959–71.

1382 Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., *et*
1383 *al.* 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing.
1384 *Nature* **518**: 371–375.

1385 Lamichhaney, S., Han, F., Berglund, J., Wang, C., Sallman, A.M., Webster, M.T., *et al.* 2016. A
1386 beak size locus in Darwin's finches facilitated character displacement during a drought.
1387 *Science (80-.)*. **6284**: 470–474.

1388 Leaché, A.D., Harris, R.B., Maliska, M.E. & Linkem, C.W. 2013. Comparative species
1389 divergence across eight triplets of spiny lizards (*Sceloporus*) using genomic sequence
1390 data. *Genome Biol. Evol.* **5**: 2410–9.

1391 Le Moan, A., Gagnaire, P. -A. & Bonhomme, F. 2016. Parallel genetic divergence among
1392 coastal-marine ecotype pairs of European anchovy explained by differential introgression
1393 after secondary contact. *Mol. Ecol.*, doi: 10.1111/mec.13627.

1394 Lescak, E.A., Bassham, S.L., Catchen, J., Gelmond, O., Sherbick, M.L., von Hippel, F.A., *et al.*
1395 2015. Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proc. Natl. Acad.*
1396 *Sci. U. S. A.* 201512020.

1397 Lewontin, R. & Krakauer, J. 1973. DISTRIBUTION OF GENE FREQUENCY AS A TEST OF
1398 THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS. *Genetics*
1399 175–195.

1400 Li, H. & Durbin, R. 2011. Inference of human population history from individual whole-genome
1401 sequences. *Nature* **475**: 493–6.

1402 Liang, M. & Nielsen, R. 2014. The lengths of admixture tracts. *Genetics* **197**: 953–967.

1403 Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., *et al.*
1404 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral
1405 sites across the human genome. *PLoS Genet.* **7**.

1406 Lotterhos, K.E. & Whitlock, M.C. 2014. Evaluation of demographic history and neutral
1407 parameterization on the performance of FST outlier tests. *Mol. Ecol.* **23**: 2178–2192.

1408 Luttikhuisen, P.C., Drent, J., Peijnenburg, K.T.C. a, van der Veer, H.W. & Johannesson, K.
1409 2012. Genetic architecture in a marine hybrid zone: comparing outlier detection and
1410 genomic clines analysis in the bivalve *Macoma balthica*. *Mol. Ecol.* **21**: 3048–61.

1411 Maddison, W.P. 1997. Gene trees in species trees. *Syst. Biol.* **46**: 523–536.

1412 Mailund, T., Halager, A.E., Westergaard, M., Dutheil, J.Y., Munch, K., Andersen, L.N., *et al.*
1413 2012. A new isolation with migration model along complete genomes infers very different
1414 divergence processes among closely related great ape species. *PLoS Genet.* **8**: e1003125.

1415 Mallet, J., Besansky, N. & Hahn, M.W. 2016. How reticulated are species? *BioEssays* **38**: 140–
1416 149.

1417 Marques, D.A., Lucek, K., Meier, J.I., Mwaiko, S., Wagner, C.E., Excoffier, L., *et al.* 2016.
1418 Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLOS Genet.*
1419 **12**: e1005887.

1420 Martin, C.H., Cutler, J.S., Friel, J.P., Dening Touokong, C., Coop, G. & Wainwright, P.C. 2015a.
1421 Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on
1422 one of the clearest examples of sympatric speciation. *Evolution (N. Y.)*. **69**: 1406–1422.

1423 Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., *et al.*
1424 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies.
1425 *Genome Res.* **23**: 1817–28.

1426 Martin, S.H., Davey, J.W. & Jiggins, C.D. 2014. Evaluating the use of ABBA-BABA statistics to
1427 locate introgressed loci. *Mol. Biol. Evol.* **32**: 244–257.

1428 Martin, S.H., Eriksson, A., Kozak, K.M., Manica, A. & Jiggins, C.D. 2015b. Speciation in
1429 *Heliconius* Butterflies : Minimal Contact Followed by Millions of Generations of
1430 Hybridisation. *BioRxiv* 1–24.

1431 Massy, B. De. 2013. Initiation of Meiotic Recombination : How and Where ? Conservation and
1432 Specificities Among Eukaryotes. *Annu. Rev. Genet.* **47**: 563–599.

1433 McGaugh, S.E. & Noor, M.A.F. 2012. Genomic impacts of chromosomal inversions in parapatric
1434 *Drosophila* species. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 422–9.

1435 McGee, M.D., Neches, R.Y. & Seehausen, O. 2016. Evaluating genomic divergence and

1436 parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. *Mol.*
1437 *Ecol.* **25**: 260–268.

1438 McKinnon, J.S., Mori, S., Blackman, B.K., David, L., Kingsley, D.M., Jamieson, L., *et al.* 2004.
1439 Evidence for ecology's role in speciation. *Nature* **429**: 294–298.

1440 McKinnon, J.S. & Rundle, H.D. 2002. Speciation in nature: the threespine stickleback model
1441 systems. *Trends Ecol. Evol.* **17**: 480–481.

1442 Messer, P.W. & Petrov, D. a. 2013. Population genomics of rapid adaptation by soft selective
1443 sweeps. *Trends Ecol. Evol.* **28**: 659–69. Elsevier Ltd.

1444 Michel, A., Sim, S., Powell, T.H.Q., Taylor, M.S., Nosil, P. & Feder, J.L. 2010. Widespread
1445 genomic divergence during sympatric speciation. *Proc. ...* **107**: 9724–9729.

1446 Myers, S., Bowden, R., Tumian, A., Bontrop, R., Freeman, C., MacFie, T., *et al.* 2010. Drive
1447 against hotspot motifs in primates implicates the PRDM9 Gene in Meiotic Recombination.
1448 *Science (80-)*. 876–879.

1449 Nachman, M.W. & Payseur, B. a. 2012. Recombination rate variation and speciation: theoretical
1450 predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. Lond. B.*
1451 *Biol. Sci.* **367**: 409–21.

1452 Nadachowska-Brzyska, K., Burri, R., Olason, P.I., Kawakami, T., Smeds, L. & Ellegren, H.
1453 2013. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred
1454 from Whole-Genome Re-sequencing Data. *PLoS Genet.* **9**: e1003942.

1455 Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., *et al.*
1456 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol.*
1457 *Ecol.* **22**: 814–826.

1458 Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., *et al.*
1459 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by
1460 large-scale targeted sequencing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 343–53.

1461 Neafsey, D.E., Lawniczak, M.K.N., Park, D.J., Redmond, S.N., Coulibaly, M.B., Traoré, S.F., *et*
1462 *al.* 2010. SNP genotyping defines complex gene-flow boundaries among African malaria
1463 vector mosquitoes. *Science* **330**: 514–7.

1464 Ness, R.W., Morgan, A.D., Vasanthakrishnan, R.B., Colegrave, N. & Keightley, P.D. 2015.
1465 Extensive de novo mutation rate variation between individuals and across the genome of
1466 *Chlamydomonas reinhardtii*. *Genome Res.* **25**: 1739–1749.

1467 Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., *et al.* 2015. A genetic
1468 mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* **47**: 1–7.
1469 Nature Publishing Group.

1470 Nobrega, M.A. 2003. Scanning Human Gene Deserts for Long-Range Enhancers. *Science* (80-
1471). **302**: 413–413.

1472 Noor, M.A.F. 2008. Mutagenesis from meiotic recombination is not a primary driver of sequence
1473 divergence between *Saccharomyces* species. *Mol. Biol. Evol.* **25**: 2439–2444.

1474 Noor, M.A.F. & Bennett, S.M. 2009. Islands of speciation or mirages in the desert? Examining
1475 the role of restricted recombination in maintaining species. *Heredity (Edinb)*. **103**: 439–444.

1476 Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., *et al.* 2005. The pattern
1477 of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**: 1289–1299.

1478 Nosil, P. 2013. Degree of sympatry affects reinforcement in drosophila. *Evolution (N. Y)*. **67**:
1479 868–872.

1480 Nosil, P. 2012. *Ecological Speciation*. Oxford University Press, Oxford, UK.

1481 Nosil, P., Crespi, B.J. & Sandoval, C.P. 2002. Host-plant adaptation drives the parallel evolution
1482 of reproductive isolation. *Nature* **417**: 440–443.

1483 Nosil, P. & Feder, J.L. 2012a. Genomic divergence during speciation: causes and
1484 consequences. *Philos. Trans. R. Soc. London Ser. B* **367**: 332–342.

1485 Nosil, P. & Feder, J.L. 2012b. Widespread yet heterogeneous genomic divergence. *Mol. Ecol.*
1486 **21**: 2829–32.

1487 Nosil, P., Funk, D.J. & Ortiz-Barrientos, D. 2009. Divergent selection and heterogeneous
1488 genomic divergence. *Mol. Ecol.* **18**: 375–402.

1489 Nosil, P., Gompert, Z., Farkas, T.E., Comeault, a. a., Feder, J.L., Buerkle, C. a., *et al.* 2012.
1490 Genomic consequences of multiple speciation processes in a stick insect. *Proc. R. Soc. B*
1491 *Biol. Sci.* 5058–5065.

1492 Nosil, P., Harmon, L.J. & Seehausen, O. 2008. Ecological explanations for (incomplete)
1493 speciation. *Trends Ecol. Evol.* **24**: 145–156.

1494 Ohta, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annu. Rev. Ecol. Syst.* **23**:
1495 263–286.

1496 Orr, H.A. 1998. The population genetics of adaptation: the distribution of factors fixed during
1497 adaptive evolution. *Evolution (N. Y)*. **52**: 935–949.

1498 Ortiz-Barrientos, D., Engelstädter, J. & Rieseberg, L.H. 2016. Recombination Rate Evolution
1499 and the Origin of Species. *Trends Ecol. Evol.* **31**: 226–236. Elsevier Ltd.

1500 Pease, J.B. & Hahn, M.W. 2013. More accurate phylogenies inferred from low-recombination
1501 regions in the presence of incomplete lineage sorting. *Evolution* **67**: 2376–84.

1502 Pennings, P.S. & Hermisson, J. 2006. Soft sweeps II - Molecular population genetics of
1503 adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**: 1076–1084.

1504 Phadnis, N., Baker, E.P., Cooper, J.C., Frizzell, K.A., Hsieh, E., de la Cruz, A.F.A., *et al.* 2015.
1505 An essential cell cycle regulation gene causes hybrid inviability in *Drosophila*. *Science* **350**:
1506 1552–5.

1507 Phadnis, N. & Orr, H.A. 2009. Sterility and Segregation Distortion in *Drosophila* Hybrids. **323**:
1508 376–379.

1509 Phifer-Rixey, M., Bomhoff, M. & Nachman, M.W. 2014. Genome-wide patterns of differentiation
1510 among house mouse subspecies. *Genetics* **198**: 283–297.

1511 Phung, T.N., Huber, C.D. & Lohmueller, K.E. 2016. Determining the Effect of Natural Selection
1512 on Linked Neutral Divergence across Species. *PLoS Genet* **12**: 1–27.

1513 Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., *et al.* 2014. The genomic
1514 landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* (80-.).
1515 **344**: 1410–4.

1516 Pool, J.E. & Nielsen, R. 2009. Inference of historical changes in migration rate from the lengths
1517 of migrant tracts. *Genetics* **181**: 711–9.

1518 Presgraves, D.C. 2007. Speciation genetics: epistasis, conflict and the origin of species. *Curr.*
1519 *Biol.* **17**: R125-7.

1520 Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using
1521 multilocus genotype data. *Genetics* **155**: 945–959.

1522 Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160**:
1523 1179–89.

1524 Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C. & Panova, M. 2016. Shared
1525 and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local
1526 scale. *Mol. Ecol.* **25**: 287–305.

1527 Renault, S., Grassa, C.J., Yeaman, S., Moyers, B.T., Lai, Z., Kane, N.C., *et al.* 2013. Genomic
1528 islands of divergence are not affected by geography of speciation in sunflowers. *Nat.*
1529 *Commun.* **4**: 1827. Nature Publishing Group.

1530 Renault, S., Maillet, N., Normandeau, E., Sauvage, C., Derome, N., Rogers, S.M., *et al.* 2012.
1531 Genome-wide patterns of divergence during speciation: the lake whitefish case study.
1532 *Philos. Trans. R. Soc. B* **367**: 354–363.

1533 Renault, S., Nolte, A.W., Rogers, S.M., Derome, N. & Bernatchez, L. 2010. SNP signatures of
1534 selection on standing genetic variation and their association with adaptive phenotypes
1535 along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.).
1536 *Mol. Ecol.* **20**: 545–559.

1537 Renault, S., Owens, G.L. & Rieseberg, L.H. 2014. Shared selective pressure and local genomic

1538 landscape lead to repeatable patterns of genomic divergence in sunflowers. *Mol. Ecol.* **23**:
1539 311–324.

1540 Rice, W. & Hostert, E.E. 1993. LABORATORY EXPERIMENTS ON SPECIATION - WHAT
1541 HAVE WE LEARNED IN 40 YEARS. *Evol. Heal. Dis.* **47**: 1637–1653.

1542 Robinson, J.D., Bunnefeld, L., Hearn, J., Stone, G.N. & Hickerson, M.J. 2014. ABC inference of
1543 multi-population divergence with admixture from un-phased population genomic data. *Mol.*
1544 *Ecol.* **23**: 4458–4471.

1545 Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: all that’s gold
1546 does not glitter. *Evolution* **66**: 1–17.

1547 Roda, F., Ambrose, L., Walter, G.M., Liu, H.L., Schaul, A., Lowe, A., *et al.* 2013. Genomic
1548 evidence for the parallel evolution of coastal forms in the *Senecio latus* complex. *Mol.*
1549 *Ecol.* **22**: 2941–52.

1550 Roesti, M., Gavrillets, S., Hendry, A.P., Salzburger, W. & Berner, D. 2014. The genomic
1551 signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**: 3944–3956.

1552 Roesti, M., Hendry, A.P., Salzburger, W. & Berner, D. 2012. Genome divergence during
1553 evolutionary diversification as revealed in replicate lake-stream stickleback population
1554 pairs. *Mol. Ecol.* **21**: 2852–2862.

1555 Roesti, M., Kueng, B., Moser, D. & Berner, D. 2015. The genomics of ecological vicariance in
1556 threespine stickleback fish. *Nat. Commun.* **41**: 1–14. Nature Publishing Group.

1557 Roesti, M., Moser, D. & Berner, D. 2013. Recombination in the threespine stickleback genome -
1558 Patterns and consequences. *Mol. Ecol.* **22**: 3014–3027.

1559 Rogers, S.M. & Bernatchez, L. 2005. Integrating QTL mapping and genome scans towards the
1560 characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus*
1561 *clupeaformis*). *Mol. Ecol.* **14**: 351–361.

1562 Rosenzweig, B.K., Pease, J.B., Besansky, N.J. & Hahn, M.W. 2016. Powerful methods for
1563 detecting introgressed regions from population genomic data. *Mol. Ecol.* 2387–2397.

1564 Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S., *et al.*
1565 2016. Inferring the demographic history underlying parallel genomic divergence among
1566 pairs of parasitic and non-parasitic lamprey ecotypes. *Mol. Ecol.*, doi: 10.1111/mec.13664.

1567 Roux, C., Fraïsse, C., Castric, V., Vekemans, X., Pogson, G.H. & Bierne, N. 2014. Can we
1568 continue to neglect genomic variation in introgression rates when inferring the history of
1569 speciation? A case study in a *Mytilus* hybrid zone. *J. Evol. Biol.* **27**: 1662–1675.

1570 Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N. & Bierne, N. 2016. Shedding light
1571 on the grey zone of speciation along a continuum of genomic divergence. *bioRxiv* 513–

1572 516.

1573 Roux, C., Tsagkogeorga, G., Bierne, N. & Galtier, N. 2013. Crossing the species barrier:
1574 genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species.
1575 *Mol. Biol. Evol.* **30**: 1574–87.

1576 Satyaki, P.R. V, Cuykendall, T.N., Wei, K.H.-C., Brideau, N.J., Kwak, H., Aruna, S., *et al.* 2014.
1577 The Hmr and Lhr hybrid incompatibility genes suppress a broad range of heterochromatic
1578 repeats. *PLoS Genet.* **10**: e1004240.

1579 Scally, A. & Durbin, R. 2012. Revising the human mutation rate: implications for understanding
1580 human evolution. *Nat. Rev. Genet.* **13**: 745–53.

1581 Scally, A., Duthheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., *et al.* 2012.
1582 Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.
1583 Nature Publishing Group.

1584 Schiffels, S. & Durbin, R. 2014. Inferring human population size and separation history from
1585 multiple genome sequences. *Nat. Genet.* **46**: 919–925. Nature Publishing Group.

1586 Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals —
1587 mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–
1588 763. Nature Publishing Group.

1589 Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P. a, *et al.*
1590 2014. Genomics and the origin of species. *Nat. Rev. Genet.* **15**: 176–92. Nature Publishing
1591 Group.

1592 Sémon, M. & Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters
1593 in mammals. *Mol. Biol. Evol.* **23**: 1715–1723.

1594 Servedio, M.R. & Noor, M. 2003. The role of reinforcement in speciation: theory and data. *Annu.*
1595 *Rev. Ecol. Evol. Syst.* **34**: 339–364.

1596 Servedio, M.R., Van Doorn, G.S., Kopp, M., Frame, A.M. & Nosil, P. 2011. Magic traits in
1597 speciation: “magic” but not rare? *Trends Ecol. Evol.* **26**: 389–97.

1598 Shafer, A.B. a., Gattepaille, L.M., Stewart, R.E. a. & Wolf, J.B.W. 2015. Demographic inferences
1599 using short-read genomic data in an approximate Bayesian computation framework: in
1600 silico evaluation of power, biases and proof of concept in Atlantic walrus. *Mol. Ecol.* 328–
1601 345.

1602 Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., *et al.*
1603 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine
1604 sticklebacks. *Nature* **428**: 717–723.

1605 Shuker, D.M., Underwood, K., King, T.M. & Butlin, R.K. 2005. Patterns of male sterility in a

1606 grasshopper hybrid zone imply accumulation of hybrid incompatibilities without selection.
1607 *Proc. Biol. Sci.* **272**: 2491–7.

1608 Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science* (80-).
1609 **236**: 787–792.

1610 Slatkin, M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res.* **58**: 167–175.
1611 Universitetsbiblioteket i Oslo.

1612 Slatkin, M., Reviews, A. & Review, A. 1985. Gene flow in natural populations. *Annu. Rev. Ecol.*
1613 *Syst.* **16**: 393–430.

1614 Smadja, C.M. & Butlin, R.K. 2011. A framework for comparing processes of speciation in the
1615 presence of gene flow. *Mol. Ecol. Early view*: 5123–40.

1616 Smukowski, C.S. & Noor, M.A.F. 2011. Recombination rate variation in closely related species.
1617 *Heredity (Edinb)*. **107**: 496–508. Nature Publishing Group.

1618 Soria-Carrasco, V., Gompert, Z., Comeault, a. a., Farkas, T.E., Parchman, T.L., Johnston, J.S.,
1619 *et al.* 2014. Stick Insect Genomes Reveal Natural Selection’s Role in Parallel Speciation.
1620 *Science* (80-). **344**: 738–742.

1621 Sousa, V. & Hey, J. 2013. Understanding the origin of species with genome-scale data:
1622 modelling gene flow. *Nat. Rev. Genet.* **14**: 404–14. Nature Publishing Group.

1623 Stephan, W. 2010. Genetic hitchhiking versus background selection: the controversy and its
1624 implications. *Philos. Trans. R. Soc. London Ser. B-Biological Sci.* **365**: 1245–1253.

1625 Stephan, W. 2016. Signatures of positive selection: From selective sweeps at individual loci to
1626 subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.* **25**: 79–88.

1627 Storz, J.F. 2005. Using genome scans of DNA polymorphism to infer adaptive population
1628 divergence. *Mol. Ecol.* **14**: 671–688.

1629 Storz, J.F. & Wheat, C.W. 2010. Integrating evolutionary and functional approaches to infer
1630 adaptation at specific loci. *Evolution (N. Y)*. **64**: 2489–2509.

1631 Terai, Y., Seehausen, O., Sasaki, T., Takahashi, K., Mizoiri, S., Sugawara, T., *et al.* 2006.
1632 Divergent Selection on Opsins Drives Incipient Speciation in Lake Victoria Cichlids. *PLOS*
1633 *Biol* **4**: e433.

1634 The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of
1635 mimicry adaptations among species. *Nature* **487**: 94–8. Nature Publishing Group.

1636 Thomae, A.W., Schade, G.O.M., Padeken, J., Borath, M., Vetter, I., Kremmer, E., *et al.* 2013. A
1637 pair of centromeric proteins mediates reproductive isolation in *Drosophila* species. *Dev.*
1638 *Cell* **27**: 412–24. Elsevier Inc.

1639 Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R.S.T., *et al.* 2014.

1640 European sea bass genome and its variation provide insights into adaptation to euryhalinity
1641 and speciation. *Nat. Commun.* **5**: 5770.

1642 Tittes, S. & Kane, N.C. 2014. The genomics of adaptation, divergence and speciation: A
1643 congealing theory. *Mol. Ecol.* **23**: 3938–3940.

1644 Trier, C.N., Hermansen, J.S., Sætre, G.-P. & Bailey, R.I. 2014. Evidence for mito-nuclear and
1645 sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species.
1646 *PLoS Genet.* **10**: e1004075.

1647 Turner, T.L. & Hahn, M.W. 2010. Genomic islands of speciation or genomic islands and
1648 speciation? *Mol. Ecol.* **19**: 848–50.

1649 Turner, T.L., Hahn, M.W. & Nuzhdin, S. V. 2005. Genomic islands of speciation in *Anopheles*
1650 *gambiae*. *PLoS Biol.* **3**: e285.

1651 Twyford, A.D. & Friedman, J. 2015. Adaptive divergence in the monkey flower *Mimulus guttatus*
1652 is maintained by a chromosomal inversion. *Evolution* 1–33.

1653 Via, S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological
1654 speciation-with-gene-flow. *Philos. Trans. R. Soc. London Ser. B* **366**: In press.

1655 Via, S. & West, J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological
1656 speciation. *Mol. Ecol.* **17**: 4334–4345.

1657 Vines, T.H., Köhler, S.C., Thiel, M., Ghira, I., Sands, T.R., MacCallum, C.J., *et al.* 2003. The
1658 maintenance of reproductive isolation in a mosaic hybrid zone between the fire-bellied
1659 toads *Bombina bombina* and *B. variegata*. *Evolution* **57**: 1876–1888.

1660 Wang, M., Beck, C.R., English, A.C., Meng, Q., Buhay, C., Han, Y., *et al.* 2015. PacBio-LITS: a
1661 large-insert targeted sequencing method for characterization of human disease-associated
1662 chromosomal structural variations. *BMC Genomics* **16**: 214.

1663 Whitlock, M.C. & McCauley, D.E. 1999. Indirect measures of gene flow and migration: F_{ST} not
1664 equal $1/(4Nm+1)$. *Heredity (Edinb)*. **82**: 117–125.

1665 Wood, B.P. & Miller, J.R. 2006. Linked selected and neutral loci in heterogeneous
1666 environments. *J. Math. Biol.* **53**: 939–975.

1667 Wu, C.-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* **14**: 851–865.

1668 Wu, C.-I. & Ting, C.-T. 2004. Genes and speciation. *Nat. Rev. Genet.* **5**: 114–22.

1669 Yeaman, S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive
1670 loci. *Proc. Natl. Acad. Sci. U. S. A.* **110**: E1743-51.

1671 Yeaman, S. 2015. Local Adaptation by Alleles of Small Effect. *Am. Nat.* **186**: S74–S89.

1672 Yeaman, S., Aeschbacher, S. & Bürger, R. 2016. The evolution of genomic islands by increased
1673 establishment probability of linked alleles. *Mol. Ecol.* 2542–2558.

- 1674 Yeaman, S. & Otto, S.P. 2011. Establishment and maintenance of adaptive genetic divergence
1675 under migration, selection, and drift. *Evolution* **65**: 2123–9.
- 1676 Yeaman, S. & Whitlock, M.C. 2011. The genetic architecture of adaptation under migration-
1677 selection balance. *Evolution* **65**: 1897–911.
- 1678 Zeng, K. & Corcoran, P. 2015. The effects of background and interference selection on patterns
1679 of genetic variation in subdivided populations. *Genetics* **201**: 1539–1554.

1680

1681 **Figures**

1682 Figure 1: Factors potentially shaping the genomic landscape. Different demographic histories,
1683 features of the genome and processes can produce apparently equivalent landscapes of
1684 differentiation. During primary divergence, barrier loci and their barrier effects increase
1685 differentiation. However, local selective sweeps not related to speciation may also produce
1686 peaks of divergence. Also regions of reduced recombination can give rise to such peaks. Under
1687 secondary contact, gene flow must eventually erode differentiation that has built up during
1688 isolation due to drift and potentially local adaptation. Yet mutation cold-spots may suggest that
1689 gene flow has recently occurred - when its effect in reality was negligible.

1690

1691 Figure 2. Relative differentiation F_{ST} averaged over 5000 independent evolutionary histories
1692 during different speciation scenarios. The figure shows an F_{ST} heatmap (see the colour bar for
1693 reference) as a function of time since the start of selection, and as a function of physical
1694 distance from the locus under selection in three different scenarios. Primary divergence with a
1695 hard sweep (A) and soft sweep (B) and secondary contact with a hard sweep during a period of
1696 isolation (C). Solid lines show the frequency of the allele that sweeps through the population
1697 where it is beneficial. Parameters: $N = 500$ individuals per deme. Migration rate per individual,
1698 deme, and generation: $m = 0.004$. Mutation rate per allele, locus, individual, and generation: $\mu =$
1699 4×10^{-5} . Selection coefficient: $s = 0.2$. In all cases, there was an initial phase of neutral
1700 evolution lasting for at least 2×10^4 generations (selection coefficient was set to $s = 0$ during this
1701 phase). In the case of soft sweep (panel B) the allele frequency at the selected locus in either
1702 population was conditioned to be between 30% and 70% when selection started. Approach to
1703 equilibrium is slow but the patterns obtained at the end of simulations are similar to those
1704 expected at equilibrium. Note the logarithmic timescale on the x-axis, and different spacing
1705 between neighbouring loci on the y-axis.

1706

1707

1708 Figure 3: Relative differentiation F_{ST} obtained in a single stochastic realisation of the model in
1709 the case of a hard sweep occurring in a primary contact. All parameters are same as in Figure
1710 2. Note that the range of F_{ST} values obtained in the individual realisation is larger than the range
1711 used in the colour bar (the highest F_{ST} values are >0.4 , very close to unity), but for simpler
1712 interpretation and comparison to the results shown in Fig. 2, all values here are truncated to the
1713 range used in Fig. 2.

1714

1715 Figure 4: Comparison of average relative differentiation F_{ST} and average absolute differentiation
1716 d_{XY} at the neutral locus at distance 500 kb from the locus under selection during a hard sweep in
1717 primary contact, as a function of time since the start of selection. Shown is the result obtained
1718 in. Blue: F_{ST} , red: d_{XY} . Averages are made over 10000 independent evolutionary histories. All
1719 other parameters are same as in Figure 2.

1720

1721

1722 **Tables**

1723 Table 1: Examples of systems where evidence of heterogeneous genomic differentiation or divergence has been identified using
 1724 genome-scan approaches. Note that this table is not intended to be an exhaustive summary.
 1725

Study system	NGS approach	Genome scan approach	Main findings	Reference
European rabbit subspecies: <i>Oryctolagus cuniculus cuniculus</i> & <i>O. c. algirus</i>	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 <i>et al.</i> , (2014)
Fruit fly subspecies: <i>Drosophila pseudoobscura pseudoobscura</i> & <i>D. persimilis</i>	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: <i>Helianthus annuus</i> & <i>H. petiolaris</i>	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 <i>et al.</i> , (2013)
Marine and freshwater three-spined stickleback ecotypes: <i>Gasterosteus aculeatus</i>	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 <i>et al.</i> , (2012)
Walking stick ecotypes: <i>Timema cristinae</i>	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 <i>et al.</i> , (2014)
House mouse subspecies:	RNA-seq	Sliding window estimates of	Higher differentiation on sex	Phifer-Rixey <i>et al.</i> ,

<i>Mus musculus musculus</i> & <i>M. m. domesticus</i>		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 maggot fly ecotypes: <i>Rhagoletis pomonella</i>	Microsatellites and allozymes	Point estimates of F_{ST}	Two genomic outlier regions on separate chromosomes, suggesting support for genomic island and continent hypotheses.	Michel <i>et al.</i> , (2010)
Normal benthic and dwarf limnetic whitefish ecotypes: <i>Coregonus clupeaformis</i>	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 <i>et al.</i> , (2013)
Annual and perennial Yellow Monkey Flower ecotypes: <i>Mimulus guttatus</i>	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: <i>Anopheles gambiae</i>	SNP-genotyping array	Sliding window estimates of nucleotide diversity and F_{ST}	Regions of high differentiation at centromeres, low nucleotide diversity in high recombination regions suggest recent sweeps and 'incidental islands'	Neafsey <i>et al.</i> , (2010)
Neotropical butterfly species: <i>Heliconius melpomne</i> , <i>H. cydno</i> & <i>H. timareta</i>	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 <i>et al.</i> , (2013)
Flycatchers: <i>Ficedula sp.</i>	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 <i>et al.</i> , (2015)

1726

1727

1728 Table 2: Examples of studies where alongside genome scan data, additional evidence besides genome-scan data has been used to
 1729 demonstrate that selection occurs at outliers. Here we delineate between studies that demonstrate a genotype-phenotype link (table
 1730 section in grey), which requires separate evidence of selection on the phenotype, and studies that show signatures of selection on
 1731 the genotype (unfilled table section). We note that in some cases, e.g. lateral plate armour in three-spined sticklebacks, there are
 1732 overlaps between these categories.

1733

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 <i>et al.</i> , 2014). Reduction in sperm number maps to sex chromosomes in Pacific Ocean-Japan Sea stickleback cross (Kitano <i>et al.</i> , 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 <i>et al.</i> , 2013). Flowering time genes divergent across latitudinal gradient in sunflowers (Renaut <i>et al.</i> , 2013).
Molecular assay	Functional assays of gene products using <i>in vitro</i> methods	Usually cannot be formed using study organism.	Cichlid opsin light absorbance (Terai <i>et al.</i> , 2006). Expression of Pocket mice

			Mc1r alleles in cultured cells (Hoekstra <i>et al.</i> , 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 <i>et al.</i> , 2005) and restoration of pelvic spine phenotype in sticklebacks (Chan <i>et al.</i> , 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 <i>et al.</i> , 2013) and <i>doublesex</i> gene controlling mimicry patterns in <i>Papilio</i> butterflies (Nishikawa <i>et al.</i> , 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 (Luttikhuisen <i>et al.</i> , 2012). Loci with steep clines at genes known to be involved in RI (Trier <i>et al.</i> , 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 <i>et al.</i> , 2012). Increased differentiation amongst stream populations flanking genomic regions involved in phenotypic differentiation in lake-stream sticklebacks due to propagating selective sweeps (Roesti <i>et al.</i> , 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 <i>Drosophila</i> 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 <i>et al.</i> , 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 <i>et al.</i> , 2014); transplant of marine sticklebacks with known lateral plate <i>Eda</i> genotype demonstrates reduced fitness and allele frequency shifts (Barrett <i>et al.</i> , 2008).

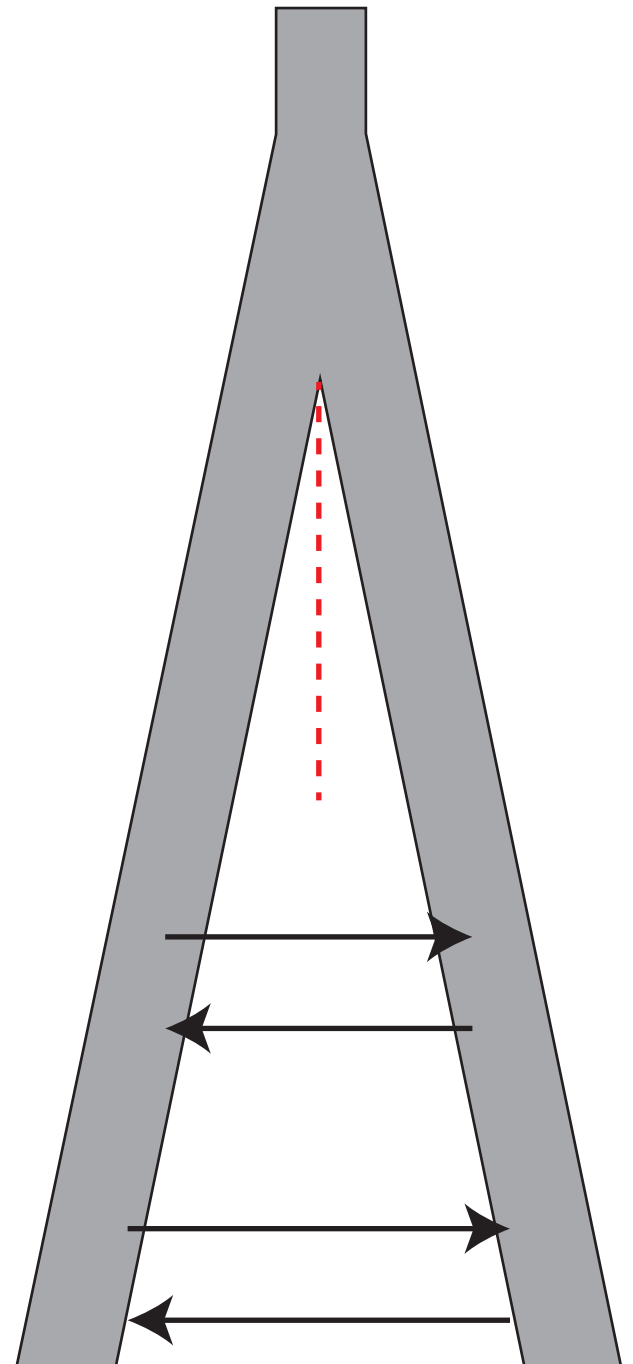
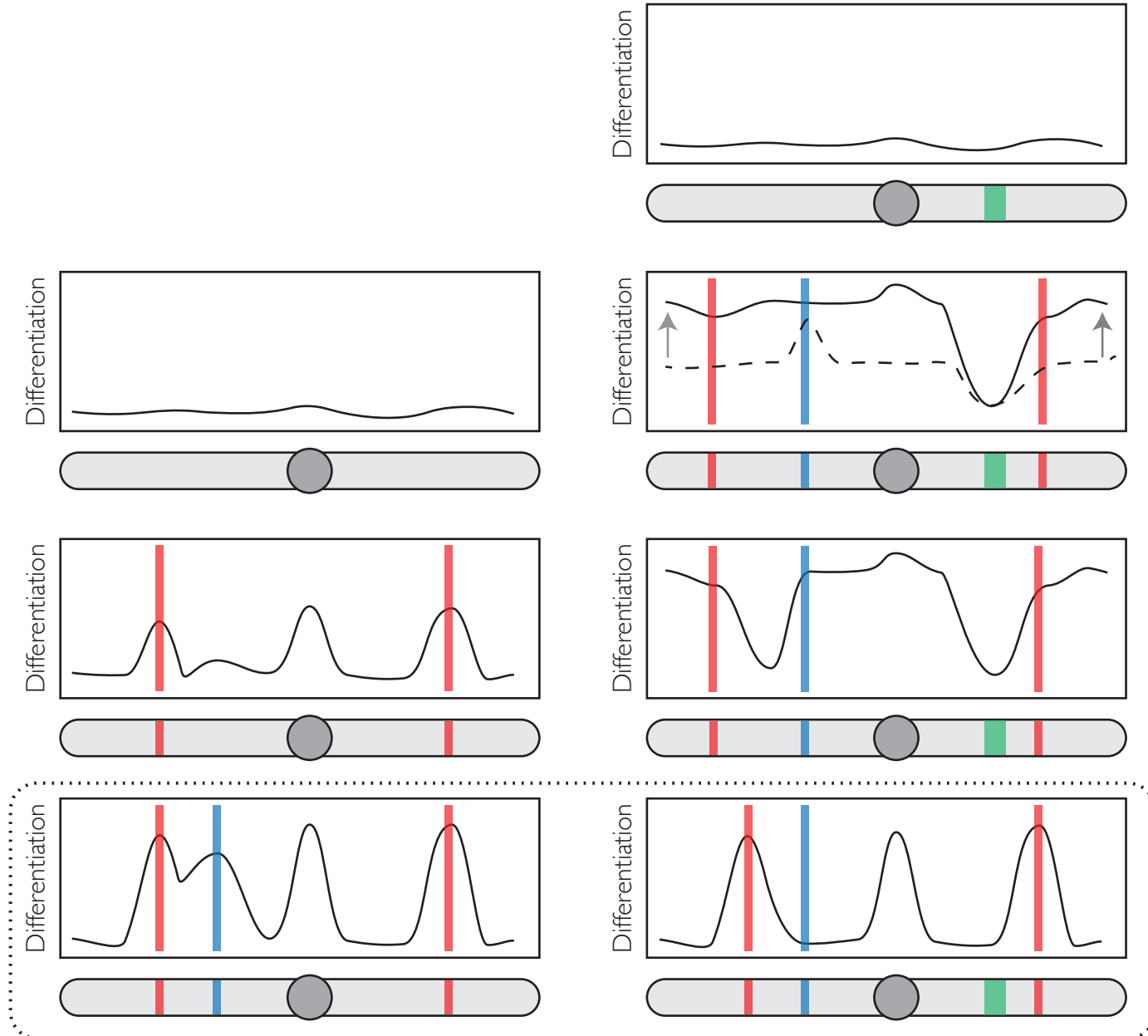
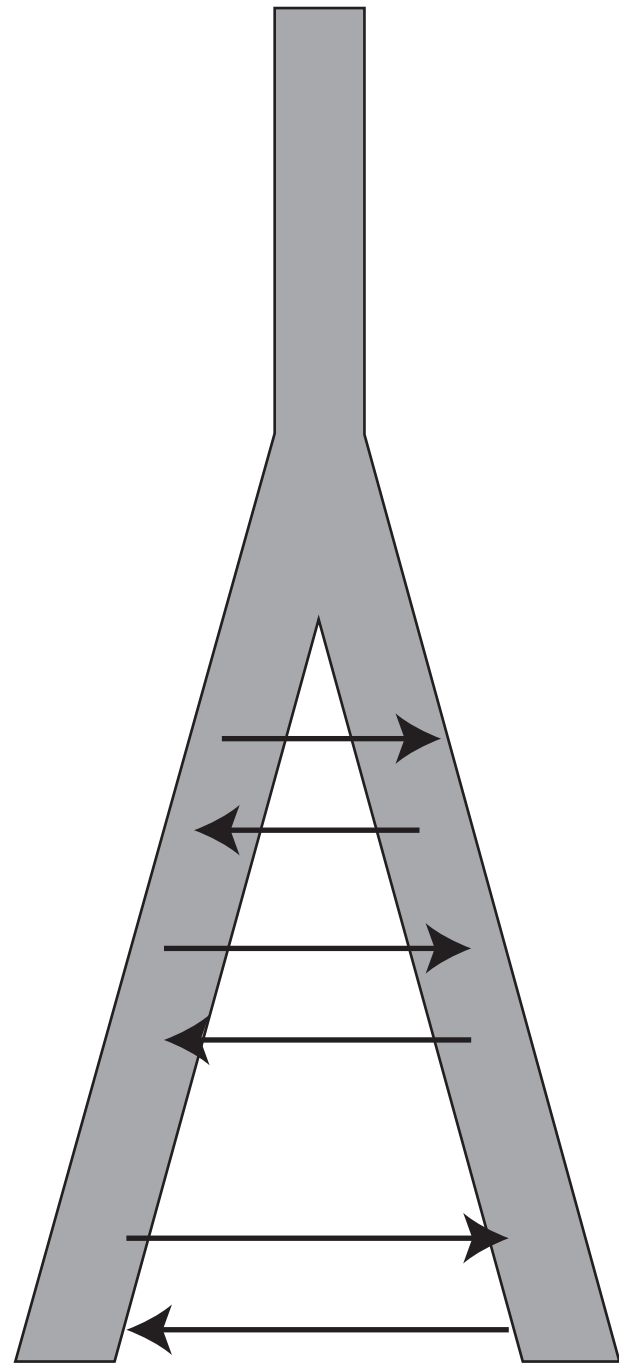
Primary divergence

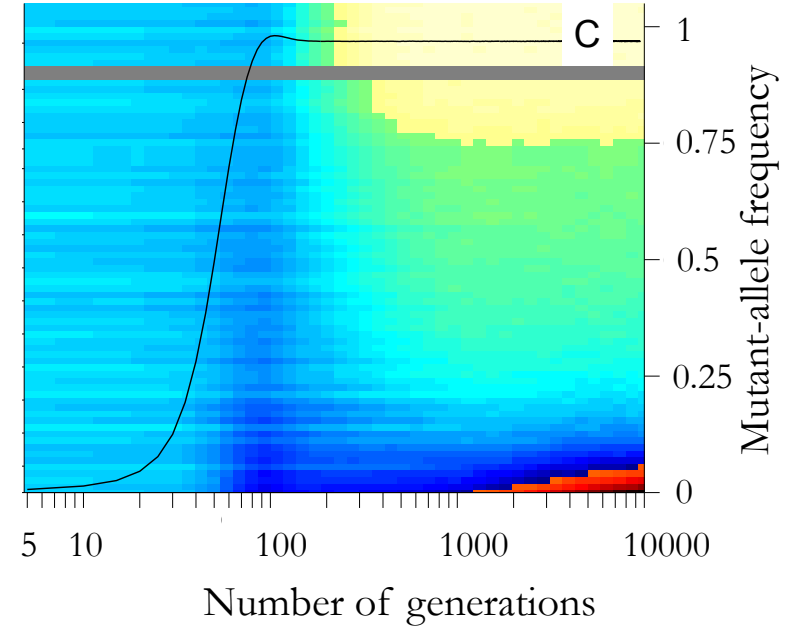
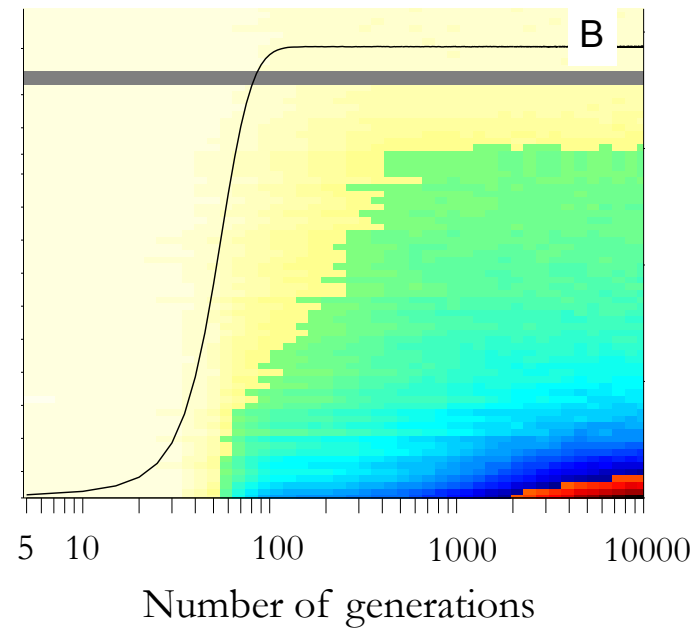
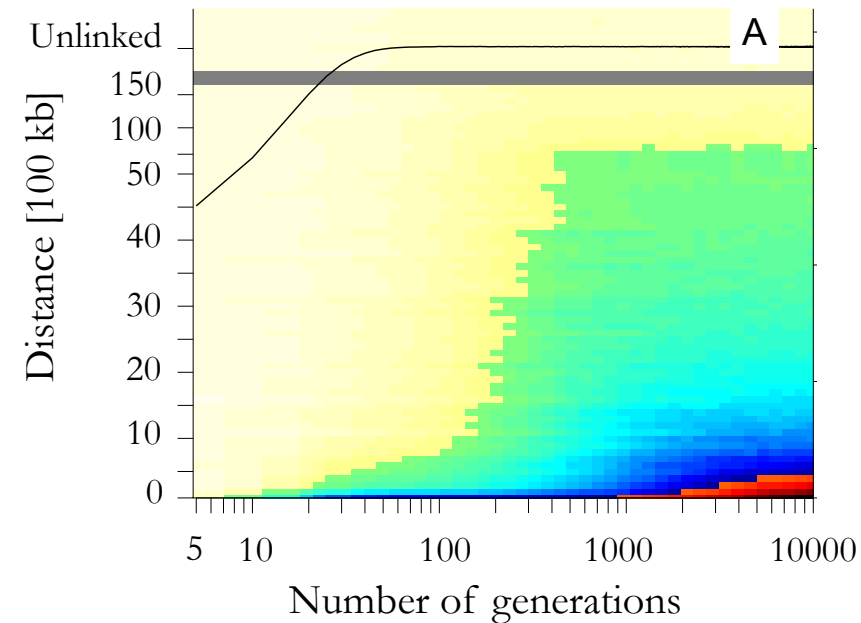
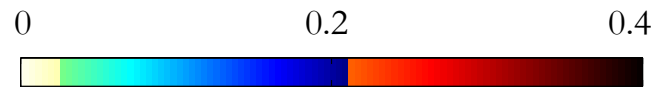
Secondary contact

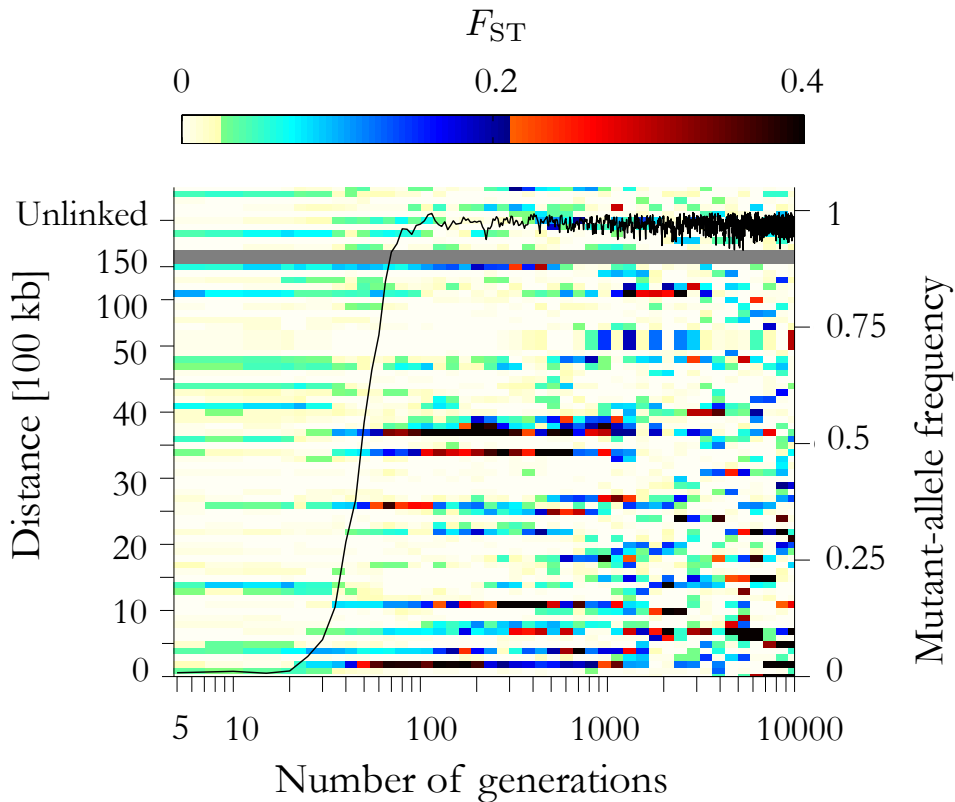
Time

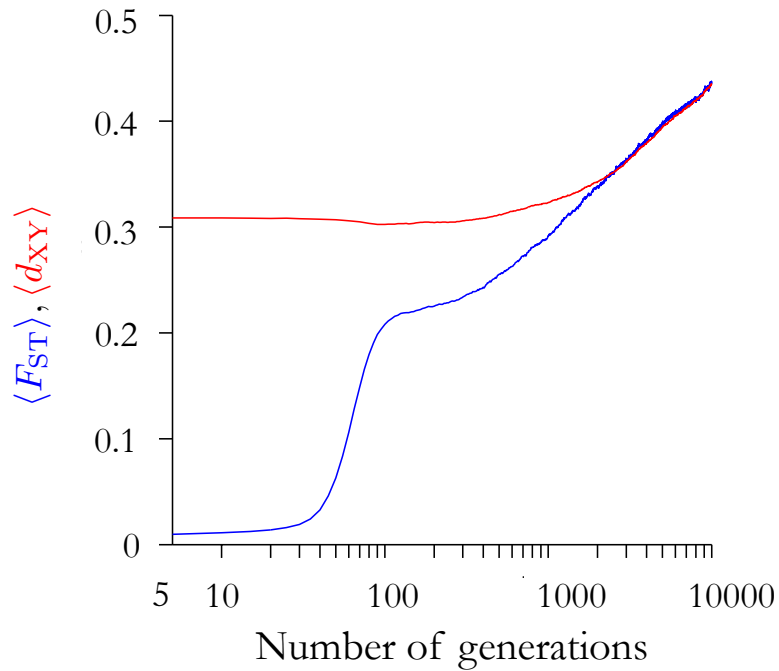
Legend:

- Red vertical bar: Barrier loci under divergent selection
- Blue vertical bar: Selective sweep due to local adaptation
- Green vertical bar: Mutation coldspot
- Grey rounded rectangle: Chromosome
- Grey circle: Centromere (reduced recombination)
- Red dashed line: Geographical isolation
- Black arrow: Gene flow



$\langle F_{ST} \rangle$ 





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 $\Pi_{T,\tau}$, as well as the within-population heterozygosities $\Pi_{S,\tau}^{(1)}$, and $\Pi_{S,\tau}^{(2)}$ 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 $F_{ST,\tau}$ and $d_{XY,\tau}$ as follows (Cruickshank and Hahn, 2014):

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

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

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, $d_{XY,\tau}$ 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 neutral 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.