

Gene flow and selection interact to promote adaptive divergence in regions of low recombination

Abstract

Adaptation to new environments often occurs in the face of gene flow. Under these conditions, gene flow and recombination can impede adaptation by breaking down linkage disequilibrium between locally adapted alleles. Theory predicts that this decay can be halted or slowed if adaptive alleles are tightly linked in regions of low recombination, potentially favoring divergence and adaptive evolution in these regions over others. Here, we compiled a global genomic dataset of over 1300 individual threespine stickleback from 52 populations and compared the tendency for adaptive alleles to occur in regions of low recombination between populations that diverged with or without gene flow. In support of theory, we found that adaptive alleles (F_{ST} and d_{XY} outliers) tend to occur more often in regions of low recombination in populations where divergent selection and gene flow jointly occur. This result remained significant when we: employed different genomic window sizes; controlled for the effects of mutation rate and gene density; controlled for overall genetic differentiation; varied the genetic map used to estimate recombination and used a continuous (rather than discrete) measure of geographic distance as proxy for gene flow/shared ancestry. We argue that our study provides the first statistical evidence that gene flow per se shapes genomic patterns of differentiation by biasing where divergence occurs in the genome.

Introduction

Understanding the genetic basis of adaptation is a fundamental goal of evolutionary biology. Yet, we still know little about the myriad interacting factors that determine the number, genomic location and effect size of loci underlying adaptive traits. Recent work suggests that interactions between two common evolutionary forces, natural selection and gene flow, may profoundly shape where adaptation occurs in the genome (Kirkpatrick & Barton 2006; Noor & Feder 2006; Yeaman & Whitlock 2011; Nachman & Payseur 2012; Aeschbacher *et al.* 2016). When divergent selection and gene flow co-occur (hereafter ‘DS-GF’), hybridization between migrant and local individuals breaks down positive linkage disequilibrium (LD) between sets of locally adapted alleles, impeding adaptation (Kirkpatrick & Barton 2006; Nachman & Payseur 2012; Sousa & Hey 2013). This decay of positive LD can be slowed if locally adapted alleles are tightly genetically linked, e.g. physically

32 close on the same chromosome, or occurring together in a region of low recombination (Rieseberg
33 2001; Noor *et al.* 2001a; Navarro & Barton 2003; Yeaman & Whitlock 2011). Accordingly, theory
34 predicts that DS-GF will drive a tendency for locally adapted alleles to be tightly linked in the
35 genome, either by physical proximity or by co-localization in regions of low recombination (Yeaman
36 & Whitlock 2011; Bürger & Akerman 2011; Aeschbacher *et al.* 2016).

37 Recent studies have offered mixed support for this prediction. Roesti *et al.* (2013) and
38 Marques *et al.* (2016) both report that parapatric pairs of stickleback ecotypes exhibit elevated
39 divergence in region of low recombination (suggesting that gene flow and selection may interact as
40 predicted), while Renaut *et al.* (2013) and Burri *et al.* (2015) found no relationship between gene
41 flow, selection and recombination in sunflowers and flycatchers respectively.

42 However, definitively testing the prediction that gene flow and selection interact to promote
43 divergence in regions of low recombination requires a system in which we can carry out replicated
44 comparisons of the genomic distribution of adaptive alleles between populations with and without
45 gene flow, and populations with and without divergent selection. This has not yet been possible, as
46 previous studies have focused on individual populations or several pairs of populations (Roesti *et al.*
47 2013; Renaut *et al.* 2013; Marques *et al.* 2016). It is also necessary to disentangle the effects of
48 selection and gene flow from other processes that can generate clustering of adaptive alleles. For
49 example, linked selection – hitchhiking and background selection – is widely known to cause
50 clustering of diverged loci (e.g. a single adaptive allele and surrounding linked neutral alleles), an
51 effect that is amplified in regions of low recombination even in the absence of gene flow
52 (Charlesworth 2012; Cutter & Payseur 2013). In addition, recombination may itself be mutagenic,
53 which would result in decreased rates of divergence in regions of low recombination (Hairston *et al.*
54 2005; Nachman & Payseur 2012). Isolating the effects of these various processes has thus far proved
55 challenging (Renaut *et al.* 2013; Burri *et al.* 2015).

56 To approach this problem, we assembled a large population genomic dataset of threespine
57 sticklebacks (*Gasterosteus aculeatus*) from across the northern hemisphere (Figure S1, Table S1).
58 Threespine sticklebacks are a holarctic species of fish that have evolved into a variety of unique
59 forms over the last 10,000 years (McKinnon & Rundle 2002). Notably, the various forms of
60 stickleback have evolved repeatedly in the presence and absence of gene flow (McKinnon & Rundle
61 2002). This allows for statistical comparisons of the genomic distribution of adaptive alleles among
62 groups of population pairs experiencing varying levels of divergent selection and gene flow. Here,
63 we focused on comparing population pairs in which divergent selection occurs in the face of gene

64 flow to population pairs experiencing selection alone, gene flow alone, or neither. Using this
65 approach, we tested the theoretical prediction that when divergent selection and gene flow co-occur,
66 adaptive alleles are more likely to fix in regions of low recombination and/or occur in tightly linked
67 clusters throughout the genome.

69 **Results**

70 **Population genomic dataset**

71 We obtained DNA sequences from databases and generated new genomic data for 20
72 populations. The combined dataset included genomic data from 1356 individuals from 52 unique
73 populations, each belonging to one of seven described ecotypes: oceanic, lake, stream, benthic,
74 limnetic, white, and Sea of Japan (Figure S1, Table S1). The genomic data were a mixture of
75 Restriction Amplified Digest (RAD), Genotyping-By-Sequencing (GBS), and whole genome re-
76 sequencing datasets. We used a single bioinformatics pipeline to standardize the identification of
77 single nucleotide polymorphisms (SNPs) across all study populations (see Methods). Using a variety
78 of criteria (see Methods), we classified each pair of populations into four discrete “evolutionary
79 regimes”: divergent selection with gene flow (DS-GF), divergence selection in allopatry (DS-Allo),
80 parallel selection with gene flow (PS-GF), and parallel selection in allopatry (PS-Allo).

82 **Localizing candidates for adaptive divergence**

83 In accordance with previous work, we found a general pattern of divergence being higher in
84 regions of low recombination (Figure 1). We identified adaptively differentiated regions of the
85 genome by separately locating SNPs and 75 kilobase pair windows that exhibited unusually high
86 levels of genetic divergence in each pair-wise comparison. For all loci (SNPs or windows), we used
87 two metrics of divergence: F_{ST} and d_{XY} , each analyzed independently. We considered loci with
88 divergence scores larger than the 95th percentile of the total distribution to be putatively adaptive
89 loci. While other forces may have caused divergence at these loci, loci subject to divergent selection
90 should be enriched in this set (Narum & Hess 2011). For convenience, we refer to these hereafter as
91 ‘outlier SNPs’ and ‘outlier windows’. For each window, we also estimated mutation rates using a
92 phylogenetic approach, and obtained estimates of gene density for each window from the
93 ENSEMBL database.

95 **Divergence in regions of low recombination**

96 For each pairwise comparison we used logistic regression to fit outlier status of windows
97 (outlier vs. non-outlier) to their estimated rates of recombination, while controlling for mutation rate
98 and gene density. The slopes of these regressions were then compared among the four gene
99 flow/selection regimes using a permutation test (see Methods)

100 In agreement with previous work (Noor & Bennett 2009; Roesti *et al.* 2013; Renaut *et al.* 2013;
101 Marques *et al.* 2016), we found that F_{ST} outlier windows occurred most often in regions of low
102 recombination, even between allopatric populations and between populations inhabiting similar
103 environments (Figure 2). However, as predicted, this tendency was significantly more extreme in
104 DS-GF comparisons compared to other evolutionary regimes (Figure 2; Figure S2, permutation test
105 on difference in correlation coefficients between regimes: two-sided $p = 0.0002$). The result
106 remained significant after re-analysis using a window size of 150kb (permutation test, $p < 0.0002$)
107 and when recombination rates were estimated using a genetic map derived from North American
108 stickleback populations (Glazer *et al.* 2015; permutation test, $p < 0.0024$).

109 d_{XY} outliers also showed a tendency (albeit non-significant) to occur most often in regions of
110 low recombination (Figure S2; permutation test: two-sided $p = 0.475$). That said, our estimates of
111 d_{XY} from GBS/RAD dataset had considerable levels of noise, likely due to low marker density in the
112 75kb windows. We thus repeated the d_{XY} analysis, but restricted the analysis to whole genome
113 datasets (see Methods). Using this reduced dataset and 75 kb windows, we found that the
114 relationship between d_{XY} (both outlier status and mean d_{XY}) and recombination was negative in DS-
115 GF comparison and positive in DS-Allo comparisons (Figure 3). This difference in slopes between
116 regimes was highly significant (likelihood ratio test: $\chi^2_{22} = 28.85$, $p = 5.41 \times 10^{-5}$). Thus, DS-GF
117 comparisons exhibited unusually high levels of both relative and absolute divergence in regions of
118 low recombination.

119 **Ruling out potential sources of bias**

120 *Discretization of geographic distance*

121 The use of a continuous measure of geographic distance led to qualitatively similar results for both
122 F_{ST} and d_{XY} (Figure S5). The tendency for outliers of any type to occur in regions of low
123 recombination was inversely correlated with geographic distance, but only when populations
124 exhibited divergent adaptation (Figure S5; permutation test on differences in divergent vs. parallel
125 slopes: two-sided $p = 0.0002$).
126
127

128 *Differences in genome-wide F_{ST}*

129 Previous studies have reported that the relationship between divergence and recombination might
130 scale with genome-wide divergence (Lowry *et al.* 2008; Burri *et al.* 2015). However, we found that the
131 tendency for F_{ST} outlier windows to occur in regions of low-recombination was negatively associated
132 with genome-wide F_{ST} (Figure 4, permutation test on correlation, two-sided $p = 0.0001$). This
133 suggests that the correlation between geography (as a proxy for gene flow) and F_{ST} in our dataset
134 likely biased our results in the *opposite* direction of our findings: as a regime, DS-GF had the greatest
135 number of low- F_{ST} comparisons (Figure 4, red points). Further, we found that if we restricted our
136 analyses in Figure 2 to comparisons in which genome-wide F_{ST} is in the range shared across all
137 regimes (0.185 – 0.675), the tendency for DS-GF comparisons to exhibit more F_{ST} outliers in
138 regions of low recombination remained significant (Figure S4, permutation test: two-sided $p =$
139 0.0002). Moreover, when analysed in a similar fashion, the enrichment of dxy outliers in regions of
140 low recombination in DS-GF populations was also significant (Figure S4, permutation test: two-
141 sided $p = 0.0002$).

142

143 *Differences in heterozygosity vs. recombination among regimes*

144 Intra-population heterozygosity (H_s) was generally lower in regions of low recombination (as
145 expected from linked selection in general), but DS-GF comparisons did not exhibit unusually low
146 levels of heterozygosity these regions (Figure S2; permutation test: two-sided $p = 0.755$). This
147 suggests that the tendency for outliers to occur more often in regions of low recombination in DS-
148 GF comparisons is not an artifact of reduced diversity in those specific comparisons.

149

150 **Clustering of outlier SNPs**

151 In addition to our windowed analyses, we performed a separate analysis to test if individual
152 outlier SNPs from DS-GF comparisons were more clustered than outlier SNPs in other regimes. To
153 do this, we calculated (a) the nearest neighbor distance in centimorgans (cM) between outlier SNPs
154 relative to nearest neighbor distance between all SNPs; and (b) the coefficient of variation of genetic
155 distances (in cM) between outlier SNPs. Importantly, these clustering metrics control for variation in
156 SNP density among genomic regions, and thus are not biased by differences in sequencing coverage.

157 DS-GF population pairs showed more clustering of F_{ST} outlier SNPs than population pairs in
158 other gene flow/selection regimes (Figure S4). Specifically, DS-GF outlier SNPs were on average
159 approximately one standard deviation closer together in map distance than expected on the basis of

160 overall SNP density (Figure S4, permutation test: two-sided $p < 0.0001$). Coefficients of variation
161 for the distance between F_{ST} outlier SNPs showed similar results (Figure S4, permutation test: two-
162 sided $p < 0.0001$), again indicating the highest levels of clustering in DS-GF comparisons.

163

164 **Discussion**

165 The role of gene flow in shaping the course of evolution remains a key topic in modern
166 evolutionary genetics. Here, we found that in stickleback populations experiencing divergent
167 selection in the face of gene flow (DS-GF), signatures of adaptation are unusually frequent in
168 regions of low recombination. This finding is consistent with theory predicting that maladaptive
169 gene flow favors genetic clustering of adaptive alleles (Yeaman & Whitlock 2011; Bürger &
170 Akerman 2011; Aeschbacher *et al.* 2016).

171 This finding has several key implications for our understanding of the genetics of adaptation.
172 First, we provide key support for theoretical predictions (Navarro & Barton 2003; Yeaman &
173 Whitlock 2011; Nachman & Payseur 2012; Aeschbacher *et al.* 2016) that DS-GF should exhibit
174 unique patterns of genomic divergence. Testing these predictions has been a major challenge,
175 because it is difficult to control for, or rule out the effects of other evolutionary processes –
176 divergent selection *per se* being the most important (see below). Given that gene flow and selection
177 often co-occur in nature, and our results imply that the relative strengths of these processes are likely
178 an important determinate of the genomic architecture of adaptation in general (Schluter & Rambaut
179 1996; Nosil *et al.* 2009; Feder *et al.* 2012). Secondly, our results suggest that by constraining where
180 divergence can occur, gene flow may cause the “usable area” of the genome to become effectively
181 smaller. This may represent a general constraint on adaptation, and could be an important
182 contribution to our ability to explain and predict where adaptation occurs in the genome. Another
183 key implication of this constraint is that by limiting the useable areas of the genome, gene flow may
184 indirectly increase the probability that the same loci will be reused during phenotypic evolution in
185 general. Thus, we might predict that pairs of DS-GF populations (perhaps even ones where selective
186 pressures are different) should display unusual levels of concordance in the loci involved in
187 divergence, and that these loci will occur in regions of low recombination. Interestingly, many QTLs
188 involved in parallel adaptation in sticklebacks localize to regions of low recombination in the
189 genome (Noor *et al.* 2001b; Peichel & Marques 2017)

190 Note that the analyses presented here were not designed to detect changes in genome
191 structure or the modification of recombination rate among populations. We assume that

192 recombination rates are highly conserved between threespine stickleback populations. This is likely a
193 reasonable assumption given that (a) recombination maps are highly similar among threespine
194 stickleback populations from Europe and the United States (Roesti *et al.* 2013; Glazer *et al.* 2015) and
195 (b) homologous chromosomes in the distantly-related ninespine stickleback show very similar
196 patterns of recombination (Rastas *et al.* 2016). While modification of recombination can be
197 important in some systems, our results pertain to the (likely far more common) scenario in which
198 many loci with potentially varying linkage relationships underlie adaptation and DS-GF favors
199 genetic architectures in which adaptive alleles are tightly linked over other architectures (Yeaman &
200 Whitlock 2011). Future studies could extend our framework to study how gene flow shapes the
201 evolution of recombination rate and genome structure.

202

203 *The costs of low recombination*

204 By definition, loci in regions of low recombination have increased physical linkage to all
205 nearby loci. We have argued this linkage can facilitate the formation of clusters of adaptive alleles,
206 which are more likely to persist in the face of gene flow. However, low recombination also makes it
207 more difficult to (a) establish LD between adaptive alleles that arise on different backgrounds and
208 (b) break down LD among adaptive alleles and deleterious alleles that happen to arise nearby (the
209 Hill-Robertson effect, (Barton 2010). What then, is happening in the case of DS-GF populations?
210 One possibility is that recombination is still sufficiently common in regions of low recombination to
211 mitigate Hill-Robertson effects. This would imply that the extent of adaptation in regions of low
212 recombination is a complex balance between selection, migration, recombination and the rate of
213 deleterious mutation (Yeaman & Whitlock 2011; Bürger & Akerman 2011; Marques *et al.* 2016).
214 Another possibility is that the cumulative selective effects of a block of linked adaptive alleles are
215 large enough to negate all but the strongest deleterious mutations. This latter scenario would imply
216 that the (putatively adaptive) clusters of linked alleles are gradually accumulating weakly deleterious
217 alleles, and thus may eventually decay (Kirkpatrick 2016).

218

219 *Heterogenous genomic divergence*

220 Our findings also suggest that the patterns of heterogenous genomic divergence observed in
221 many speciation studies (Marko & Hart 2011; Feder *et al.* 2012) may be partly a product of the
222 interaction between gene flow and selection. Explaining this phenomenon has become a major
223 question in speciation genetics, and many recent studies have shown that patterns of heterogenous

224 divergence in the genome are correlated with recombination rate (Roesti *et al.* 2013; Renaut *et al.*
225 2013; Burri *et al.* 2015). The association between diversity, divergence and recombination is widely
226 thought to be the result of linked selection, i.e. background selection and hitchhiking (Charlesworth
227 2012). Our results suggest that there is a general negative association between recombination rate
228 and both diversity and divergence (probably generated by background selection) and this
229 relationship can be further shaped by the effects of selection (hitchhiking) and gene flow (decay of
230 divergence in regions of high recombination and/or favoring linkage between adaptive alleles).

231 Interestingly, previous work (Renaut *et al.* 2013; Burri *et al.* 2015) found no relationship
232 between gene flow and patterns of genomic divergence. One reason for this may simply be power:
233 our dataset had many individuals and populations, and included pairs of populations across the
234 speciation continuum (in terms of magnitude and time of divergence, geography and type of
235 selection). In the case of Burri *et al.* (2015), there also appears to be limited amounts of actual
236 introgression between flycatcher populations (although hybridization occurs), weakening any
237 potential pattern. Another possible explanation is that statistically detectable clustered genetic
238 architectures may require long temporal scales and/or recurrent bouts of gene flow to develop.
239 Although most stickleback populations are less than 10 000 years old, the stickleback
240 metapopulation has repeatedly cycled between adapting to freshwater environments during
241 interglacial periods, followed by extinction of these populations during glacial periods (Taylor &
242 McPhail 2000; Hendry *et al.* 2009). However, gene flow between freshwater and marine populations
243 has likely allowed ancient freshwater haplotypes to persist in marine populations throughout this
244 process (Schluter & Conte 2009). This recurrent process coupled with large effective population
245 sizes of marine stickleback may have increased the opportunity for clustered sets of co-selected
246 alleles to arise and persist.

247

248 *The effect of divergent selection*

249 Divergent selection alone is predicted to generate a correlation between recombination rate
250 and genomic divergence across the genome (Barton 2010). This effect is particularly apparent in
251 reduced representation datasets, such as the RAD and GBS datasets we analyzed here (Lowry *et al.*
252 2016). Our data support this prediction: all “divergent selection” comparisons (DS-GF and DS-Allo)
253 show increased divergence in regions of low recombination (e.g. Figure 2B, red and yellow lines).
254 However, the divergence-recombination correlation is significantly more negative in DS-GF
255 populations, which we interpret as a unique joint effect of gene flow and divergent selection. Note

256 that this pattern held when the analysis was restricted to whole-genome data (Figure 3), suggesting
257 that low marker density is not the sole source of the low-recombination bias (although undoubtedly
258 a contributor). Interestingly, gene flow alone (e.g. parallel selection + gene flow, blue lines in Figures
259 2 and 4) appears to not be sufficient to generate a bias for divergence in regions of low
260 recombination.

261 A potential alternate explanation for the increase in outlier density in regions of low
262 recombination in DS-GF comparisons is that maladaptive gene flow *per se* increases the strength of
263 divergent selection (Lenormand 2002). Stronger selection magnifies the scale of linked selection (i.e.
264 the number of loci influenced), and this in turn could increase the negative correlation between
265 recombination and divergence (Barton 2010). We cannot completely rule out this alternative.
266 However, several facts suggest that variation in the strength of selection is not the sole explanation
267 for our results. For one, the increased clustering of divergence in regions of low recombination we
268 observe is partly generated by a deficit of highly-diverged loci in regions of high recombination (e.g.
269 high recombination regions in Figure 2A). Stronger selection *per se* should not result in fewer
270 divergent loci in regions of high recombination (Barton 2010; Cutter & Payseur 2013). Gene flow,
271 on the other hand, is predicted to cause such a deficit, particularly when divergent selection is also
272 acting (Yeaman & Whitlock 2011; Aeschbacher *et al.* 2016). Secondly, because we took an “all-
273 pairwise” approach for our F_{ST} analyses, populations experiencing unusually strong directional
274 selection are also included in DS-Allopatry comparisons. Thus, any population-specific effects were
275 balanced between comparisons of regimes. Finally, it should be noted that the connection between
276 gene flow and the strength of selection is by no means well characterized – indeed under some
277 circumstances, gene flow may actually decrease the strength of divergent selection (Rolshausen *et al.*
278 2015).

279

280 *Caveats*

281 The main strength of the approach we applied here was that it allowed for replication within each
282 gene-flow/selection regime, which is necessary for examining statistical differences between regimes
283 in their recombination bias. However, the number of comparisons involved (1000+) also created
284 serious computational bottlenecks, which precluded using more sophisticated methods for detecting
285 natural selection and gene flow (Aeschbacher *et al.* 2016). Further, we do not have detailed
286 knowledge of the demographic history and historical rates of introgression between any of the
287 populations studied here. Both of these factors are known to affect patterns of divergence, and can

288 potentially alter the relationship between divergence and recombination (Tine *et al.* 2014). It is
289 possible that the more extreme recombination vs. divergence bias we observed in DS-GF
290 populations was a result of an unusual demographic or introgression history that was somehow
291 confounded with the contemporary “DS-GF” classification. For example, these comparisons may be
292 enriched for populations that have experienced a period of allopatry, followed by the resumption of
293 gene flow (secondary contact). However, this would still imply that divergent selection and gene
294 flow interact to generate a low-recombination bias, as loci not involved in divergent selection should
295 still flow freely between populations. Thus, while the mechanistic details behind the patterns we
296 describe here are still unclear, we hope our study stimulates further studies of the relationship
297 between gene flow, selection and recombination in shaping patterns of divergence.

298

299 **Acknowledgements**

300 We are very grateful to the Semiahmoo (BC) and Waycobah (NS) First Nations for granting us land
301 access for the collection of fish used directly or indirectly in this study. We also are indebted to the
302 threespine stickleback research community, whose body of work made this study possible. A.L.
303 Ferchaud, M. Roesti, M. Ravinets, J. Kitano, and T. Veen helped in obtaining the data sets used in
304 this study. G. Blackburn, M. Whitlock, L. Rieseberg, S. Yeaman, A. Geraldes, M. Roesti, M. Noor
305 and K. Ostevik and six anonymous reviewers provided key comments on ideas presented here.
306 WestGrid (Compute Canada) provided computational resources used in this project. This work was
307 supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant to
308 DS. KS, GO, DR and KD were additionally supported by NSERC graduate doctoral scholarships.

309

310 The authors made the following contributions to the work presented here. Project conception and
311 development: KS, KD, SM, GO, DR, DS; Genomic pipeline: KS, KD, SM, GO, DR; Field and lab
312 work for new data sets: KS, DR, GO; Statistical analysis: KS, GO, DS with input from KD, SM and
313 DR; Wrote the paper: KS with input from DS and the other authors. Correspondence and requests
314 for material should be addressed to KS (ksamuk@gmail.com).

315

316 Sequenced reads for the two new datasets provided here are deposited on the NCBI Sequence Read
317 Archive (accession #, to be made available before publication).

318

319 **Methods**

320

321 **Github Repository**

322 The code used to generate our dataset and perform the analyses described here is available on
323 Github at https://github.com/ksamuk/gene_flow_linkage. Additional raw data is also hosted on
324 Dryad (Dryad accession, to be made available). All scripts were written in Perl or R 3.2.2 (Team
325 2015).

326

327 **Data Sources**

328 The stickleback population genomic datasets used in this study came from two sources:
329 online databases, and new data from two of the authors. During the period from May to July 2014,
330 we periodically searched the Short Read Archive (SRA), the European Nucleotide Archive (ENA)
331 and the Databank of Japan Sequence Read Archive (DRA) for “threespined/three-
332 spined/threespine/three-spine stickleback”, “stickleback”, “Gasterosteus aculeatus”. We also
333 searched for stickleback population genetic studies on Google Scholar using the same terms as
334 above, with the inclusion of “genomic”, “genome scan”, “population genetic”, and “genetics”, and
335 examined them for SRA/ENA/DRA accession numbers. Detail information for all the populations
336 included in the study is shown in Table S1 (Hohenlohe *et al.* 2010; Roesti *et al.* 2012; Catchen *et al.*
337 2013; Yoshida *et al.* 2014; Chain *et al.* 2014; Feulner *et al.* 2015).

338 In addition to previously published data, we prepared three new datasets from
339 benthic/limnetic, freshwater lake, and white/marine populations from various locations in Canada.
340 The libraries for these datasets were prepared using a mix of Genotyping-by-Sequencing method of
341 (Elshire *et al.* 2011) and whole-genome genomic DNA (TruSeq DNA PCR-Free Library Preparation
342 Kit, Illumina, California). The collection locations and sequencing methods are listed in Table S1.
343 The resultant GBS libraries were sequenced at the University of British Columbia Biodiversity
344 Sequencing Centre, and the whole-genome libraries were sent for sequencing at Genome Quebec.
345 Sequencing was performed on an Illumina Hi-Seq 2000 at both facilities. These datasets are available
346 on the SRA (accessions # to be made available).

347

348 **Variant identification and processing**

349 We identified variants using a standard, reference-based bioinformatics pipeline (see Github
350 code repository for details). After demultiplexing, we used Trimmomatic v0.32 (Bolger *et al.* 2014) to

351 filter low quality sequences and adapter contamination. We then aligned reads to the stickleback
352 reference genome (BROAD S1, (Jones *et al.* 2012) using BWA v0.7.10 (Li & Durbin 2010), followed
353 by realignment with STAMPY v1.0.23 (Lunter & Goodson 2011). We then followed the GATK
354 v3.3.0 (Cachat *et al.* 2010) best practices workflow except that we skipped the MarkDuplicates step
355 when reads were derived from reduced representation libraries (RAD and GBS). We realigned reads
356 around indels using RealignTargetCreator, and IndelRealigner, identified variants in individuals using
357 the HaplotypeCaller, and each dataset using GenotypeGVCFs. The results were sent to a VCF file
358 containing all variant and invariant sites and converted to tabular format. All datasets were
359 combined for processing.

360

361 **Calculation of divergence metrics**

362 Our final dataset included individuals from 56 unique populations. As there was no *a priori*
363 reason to select only a subset pairs of populations in the analysis, we instead performed all possible
364 pairwise comparisons. We employ an unbiased significance testing method to overcome redundant
365 use of populations in multiple pairs (see permutation test).

366 For each of the 1128 pairwise comparisons, we calculated two divergence metrics: Weir and
367 Cockerham's F_{ST} (Weir & Cockerham 1984) and Nei's d_{XY} (Nei 1987). We calculated F_{ST} at two
368 scales: first, at each individual shared SNP; and second, averaged across 75 kilobase pair (kbp)
369 windows. For all SNPs, we required: a minor allele frequency of at least 0.05, coverage in at least 5
370 individuals per population. For windowed analysis, we required that windows contain at least 3
371 variable sites genotyped in at least 5 individuals per population. The distribution of total sequenced
372 and total variable sites for all the comparisons is shown in Figure S10.

373 Window-averaged F_{ST} values were calculated by dividing the sum of the numerators of all
374 SNP-wise F_{ST} estimates within a given window by the sum of their denominators. We calculated d_{XY}
375 in 75-kbp windows, including all shared variant and invariant sites in the window. We required d_{XY}
376 windows to contain more than 500 shared sequenced sites (i.e. nucleotides with a genotype call in
377 both populations), because we found that the variance in d_{XY} greatly increases below this threshold.
378 After calculating F_{ST} or d_{XY} , we classified SNPs and windows exhibiting extreme values as 'outliers',
379 defined as those in the 95th percentile or higher of F_{ST} or d_{XY} . Note, only d_{XY} window 'outliers' were
380 used because individual site d_{XY} scores are uninformative. All calculations were performed using
381 custom Perl and R scripts (see code repository).

382

383 **Classification of Populations**

384 For populations with multiple individuals (48 of the 56), we classified all pair-wise
385 comparisons between our 48 populations ($n = 1128$ comparisons) along two axes: ecology and gene
386 flow. We scored populations as ecologically “divergent” or “parallel” based on whether they (a)
387 inhabited different ecosystems or ecological niches and/or (b) had been directly identified by
388 previous authors as ecologically divergent (Figure S1, see Table S1 for details). The correlation
389 between divergent selection and ecology in stickleback is extremely well-supported (Schluter 1993;
390 McKinnon & Rundle 2002; Hendry *et al.* 2009) and while the strength of divergent selection may
391 vary among comparisons, we believe this is a reasonable proxy.

392 Secondly, we scored whether there has been opportunity for gene flow between populations
393 (“gene flow” / “allopatry”), based on geographic distance and barriers. This is a common
394 assumption in comparative studies, and there is strong empirical evidence that this is a reasonable
395 assumption for threespine sticklebacks. Extensive previous work suggests that nearby stickleback
396 populations often interbreed (Hendry *et al.* 2009; Marques *et al.* 2016). This interbreeding leads to
397 gene flow, as complete reproductive isolation is extremely rare among stickleback populations
398 (McKinnon & Rundle 2002; Hendry *et al.* 2009). Indeed, even the most highly differentiated
399 populations (e.g. benthic to limnetic) experience ongoing gene flow (Gow *et al.* 2006). In some cases,
400 gene flow between nearby populations becomes opposed by divergent selection, limiting the number
401 of loci affected by gene flow, although still allowing substantial gene flow in much of the genome
402 (Roesti *et al.* 2012; Jones *et al.* 2012). Thus, the use of geographic isolation as a proxy for the
403 opportunity (past or present) for gene flow is likely highly reasonable for this species.

404 We thus considered any populations within 500km of one another as having the potential
405 for gene flow. We calculated geographic distance (great circle distance) between all pairs of
406 populations using the function “earth.dist” from the R package *fossil* (Vavrek 2011). Note that this
407 classifier is conservative, as it likely causes populations that are largely allopatric (DS-Allopatry) to be
408 classified as DS-GF, decreasing the power to detect a difference between regimes.

409 Note that for both classification schemes, we are not assuming a perfect, discrete mapping
410 of selection and gene flow onto individual populations. We only assume that when considered
411 together, populations in each category will tend to exhibit greater (or less) gene flow and/or
412 divergent selection. In total, our classification scheme resulted in the following number of
413 comparisons: 130 divergent selection with gene flow, 31 parallel selection with gene flow, 113
414 parallel selection with gene flow, and 821 divergent selection in allopatry.

415

416 **Addition of Genomic Variables**

417 We measured three genomic variables in each 75-kbp window in the divergence dataset with:
418 recombination rate, mutation rate and gene density. Recombination rates (cM/MB) were obtained
419 from a previously published high-density genetic map (Roesti *et al.* 2013). Where windows
420 overlapped regions with different estimates of recombination rate, we assigned them an average of
421 the two rates weighted by the degree of overlap.

422 We obtained estimates of mutation rate by estimating the synonymous substitution rate (ds)
423 in a phylogenetic framework. For neutral sites, ds is an estimator of the primary mutation
424 rate (Wielgoss *et al.* 2011). To do this, we used the R (version 3.2.2) package *biomaRt* to obtain a list
425 of all annotated *G. aculeatus* coding DNA sequences (CDS) from ENSEMBL. For each *G. aculeatus*
426 CDS, we queried ENSEMBL for all homologous CDS from three other fish species: *Xiphophorus*
427 *maculatus*, *Poecilia formosa*, and *Oreochromis niloticus*. These species all have identical estimated
428 divergence times from *G. aculeatus* (150 MYR). We aligned each set of homologous coding sequences
429 using PRANK (Löytynoja & Goldman 2008) and analyzed the output using PAML (Branch model
430 2) to estimate ds trees. We excluded trees with fewer than three species, in order to ensure that
431 lineage-specific artefacts did not bias ds estimates. We also excluded any individual branches where
432 ds exceeded 5 standard deviations of the distribution of the ds values from all branches of every tree
433 (values exceeding this threshold were categorically the result of bad alignments). After filtering ds
434 trees, we used the R package *ape* (Paradis *et al.* 2004) to calculate the mean pairwise branch distance
435 between *G. aculeatus* and each other species in the tree. Because the other three species all have
436 identical divergence times from *G. aculeatus*, this results in a single normalized value of ds for each
437 coding sequence. After obtaining all the mutation rate estimates, we assigned them to 75 kbp
438 windows in the divergence datasets by averaging the ds estimates for genes in each window (if any),
439 weighted by the degree of overlap for each gene.

440 Estimates of gene density (number of genes overlapping the window) were calculated by
441 querying ENSEMBL (Kautt *et al.* 2012) for the physical position of all genes in the stickleback
442 genome using *biomaRt* (Yang 2007). We then wrote a custom R script (see Github repository) to
443 count the number of genes in each 75-kbp window along the reference genome.

444

445 **Tendency for adaptive divergence in regions of low recombination**

446 To quantify the tendency for outliers to occur in regions of low recombination in each
447 comparison, we employed a linear modeling approach. Using the 75-kbp windows as data points, we
448 fit a logistic regression model to each comparison dataset using the following form: Outlier status =
449 Recombination rate + mutation rate + gene density, where outlier status is 1 if a window is an
450 outlier (>95th percentile) and 0 otherwise. We performed separate model fits for F_{ST} and d_{XY} outliers.
451 We also fit models of the same type using mean intra-population heterozygosity (H_S) as the response
452 variable in order to assess its role in driving any patterns of increased divergence.

453 We fit these models in R (version 3.2.2) using the generalized linear model function “glm”.
454 Prior to model fitting, we filtered out pairwise population comparisons with fewer than 100 75-kbp
455 windows represented to ensure convergence of the linear models. To assess statistical significance of
456 the model fits, we extracted the regression coefficient for the recombination rate term from each
457 model, representing the slope of the relationship between outlier occurrence and recombination rate.
458 The steepness of the slope coefficients estimates the tendency for outliers to occur in regions of low
459 recombination, controlling for the effects of mutation rate and gene density.

460

461 **Permutation tests**

462 To test the hypothesis that adaptation with gene flow favors divergence in regions of low
463 recombination, we employed a permutation test to assess whether the slopes from the models
464 described above differed significantly between populations differing in divergent selection and gene
465 flow. To do this, we randomly shuffled regime assignments of all the populations and estimated the
466 mean low recombination outlier tendency (the grouped mean of the regression coefficients from
467 above) for each regime in 10,000 permutations. This generated a null distribution of mean slopes for
468 each regime, accounting for sample size differences between categories (Figure S2). We then
469 calculated a two-sided P value for each empirical mean by the computing the fraction of samples in
470 the null distribution greater than the observed value and multiplying by two. Note this method of
471 analysis also employed elsewhere throughout the paper (referred to as “permutation test” wherever
472 it was applied).

473

474 **Clustering vs. geographic distance and overall divergence**

475 To ensure our results were not influenced by our discrete geographic categorization scheme,
476 we examined how the tendency for F_{ST} outliers to occur in regions of low-recombination varied with
477 pairwise geographic distance. To do this, we regressed the low recombination outlier tendency

478 (regression coefficients from above) on geographic distance between populations using the R
479 function “lm”. The linear model was of the form recombination bias = distance + ecology +
480 distance * ecology (interaction). We then assessed significance of the model terms using a
481 permutation test similar to the one previously described (see code supplement)

482 The results of (Burri *et al.* 2015) and (Roesti *et al.* 2013) suggest that the tendency for F_{ST}
483 outliers to occur in regions of low recombination may be highest at intermediate levels of overall
484 genetic divergence
485 ($F_{ST} = 0.3-0.5$). Overall F_{ST} thus represents a potential source of bias, as our use of geographic
486 distance as a proxy for gene flow is naturally confounded with overall F_{ST} – with isolation by
487 distance, more distant populations will have higher divergence, all else being equal. To test if this
488 may have influenced our results, we examined the correlation between low-recombination clustering
489 tendency and overall F_{ST} . To obtain overall F_{ST} estimates between each pair of populations, we
490 divided the sum of the numerator terms by the sum of the denominator terms of all locus-specific
491 F_{ST} values for each pair (Weir & Cockerham 1984). This yielded a single average F_{ST} value for each
492 pair of populations. We then employed the same approach as the analysis of distance, with a linear
493 model the form recombination bias = F_{ST} + ecology + F_{ST} * ecology (interaction). We assess the
494 significance of this difference again via permutation test (see code supplement).

495

496 **Increased clustering of outlier SNPs**

497 To test the hypothesis that adaptation with gene flow favors clustering (reduced genetic map
498 distance) between outlier SNPs, we used two metrics of clustering: nearest neighbor map distance
499 between outliers (NND) and the coefficient of variation in map distance between consecutive
500 outliers. Both of these metrics were calculated using the SNP-level data.

501 We first asked: do map distances between nearest-neighbour outlier loci differ significantly
502 from the expected map distances of identical numbers of nearest-neighbour SNPs? This approach
503 was designed to account for disparities in SNP density that might occur due to differences in
504 sequencing outcomes between our various datasets. To do this, we first partitioned each SNP data
505 set by chromosome. Then, for each chromosome we identified the number of outlier loci using the
506 previously described method. We then drew 10,000 samples of random SNPs from each
507 chromosome equal to the number of outliers on that chromosome, and calculated the mean map
508 distance between each SNP and its nearest neighbor in the random sample. We then compared the
509 empirical mean nearest neighbor map distance of outliers to this null distribution for each

510 chromosome within each individual comparison dataset. We then used permutation tests to
511 compare (a) the proportion of chromosomes that were significantly over-clustered and (b) the
512 difference between the average NND between outliers and the average NND expected between
513 SNPs, in units of standard deviations, between the four selection and gene flow regimes.

514 In addition to the re-sampled approach, we also computed a coefficient of variation: the
515 ratio of the standard deviation in map distances between consecutive SNP on the chromosome
516 divided by the mean distance. Values exceeding one are indicative of over-dispersion (clustering),
517 whereas values below one suggest under-dispersion (uniformity of distances). We calculated the
518 coefficient of variation for outliers on each chromosome, and computed the mean for all
519 chromosomes containing outliers for each comparison. We then used a permutation test (as
520 described above) to compare the means of this quantity among gene flow/selection regimes.

521

522 **Whole genome data collection**

523 We obtained whole-genome sequences from single individuals from a total of nine stickleback
524 populations. One of these is the reference genome, derived from a marine-like individual from Bear
525 Paw Lake, Alaska (Jones *et al.* 2012). Four were individuals collected from two pairs of populations
526 that have diverged into benthic and limnetic ecotypes from Paxton and Priest Lake on Texada Island
527 in BC, Canada. These two pairs of populations (one limnetic and one benthic in each lake) have
528 diverged from each other in the face of gene flow (Taylor & McPhail 2000), making them “DS-GF”
529 populations in our classification scheme. The remaining five were collected from freshwater lakes
530 with a single, non-diverged stickleback population – Hoggan, Bullock, Trout, Cranby and Stowell
531 lakes (Miller). These latter populations diverged from the marine ancestor in allopatry – i.e. they are
532 “DS-Allopatry” populations in our scheme. DNA from these individuals was extracted via phenol-
533 chloroform extraction, and whole-genome library preparation carried out using Nextera DNA
534 Library Prep Kits (Illumina Inc.). All populations were sequenced on an Illumina HiSeq 2000 in the
535 University of British Columbia Biodiversity Sequencing Facility.

536

537 **Whole genome dxv calculation and analysis**

538 We used the GATK best practices workflow described above to call variants on the eight
539 populations above (not including the reference). We emitted VCF files containing all variant and
540 invariant sites for each population. We then computed dxv in 75,000 base pair windows using the
541 method described previously (see “Calculation of Divergence Metrics” above; code available in

542 repository). For the two pairs of DS-GF populations (Paxton and Priest), we computed d_{xy} between
543 sympatric populations within each lake. For the remaining DS-Allopatry populations, we computed
544 d_{xy} between each population and a marine population (Bear Paw Lake, i.e. the reference genome).
545 We allowed for missing sites, and for windows with no variable sites. Prior to analysis, we inspected
546 relationships between the number of genotyped sites in each window and d_{xy} . We found that the
547 variance in d_{xy} was highly inflated in windows containing fewer than 7500 genotyped sites (variant
548 and invariant). We thus excluded all windows with less than 7500 sites (out of 75,000) from the
549 analysis. As before, we classified windows with d_{xy} values exceeding the 95th percentile as “outlier
550 windows”.

551 We used a generalized linear mixed model (GLMM) to test if the relationship between d_{xy}
552 outlier status (0,1) and recombination differed between DS-GF pairs and DS-Allo pairs. We used the
553 function “glmer” in the R package *lme4* (Bates *et al.* 2015) fit a GLMM of the following form: d_{xy}
554 outlier status = recombination rate + regime + comparison (random effect). Outlier status was a
555 binary variable, and we thus used a binomial error function (i.e. a logistic regression). We then refit
556 the model, but included an interaction term: recombination rate \times regime. We then compared the fit
557 of the latter model to the simpler model using a likelihood ratio test, implemented via the R function
558 “anova”.

559

560

561

562 References

563

- 564 Aeschbacher S, Selby JP, Willis JH, Coop G (2016) *Population-genomic inference of the strength and timing of*
565 *selection against gene flow*. Cold Spring Harbor Labs Journals.
- 566 Barton NH (2010) Genetic linkage and natural selection. *Philosophical transactions of the Royal Society of*
567 *London. Series B, Biological sciences*, **365**, 2559–2569.
- 568 Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4.
569 *Journal of Statistical Software*, **67**.
- 570 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data.
571 *Bioinformatics*, **30**, 2114–2120.
- 572 Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive
573 the evolution of the genomic landscape of differentiation across the speciation continuum of
574 *Ficedula* flycatchers. *Genome Research*, **25**, 1656–1665.
- 575 Bürger R, Akerman A (2011) The effects of linkage and gene flow on local adaptation: a two-locus
576 continent-island model. *Theoretical Population Biology*, **80**, 272–288.
- 577 Cachat JM, Stewart A, Utterback E *et al.* (2010) Deconstructing Adult Zebrafish Behavior with Swim
578 Trace Visualizations. In: *Zebrafish Ecology and Behaviour*, pp. 191–201. Humana Press, Totowa, NJ.
- 579 Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization

580 history of Oregon threespine stickleback determined using restriction-site associated DNA-
581 sequencing. *Molecular Ecology*, **22**, 2864–2883.

582 Chain FJ, Feulner PGD, Panchal M *et al.* (2014) Extensive copy-number variation of young genes
583 across stickleback populations. (J Zhang, Ed.). *PLoS Genetics*, **10**, e1004830.

584 Charlesworth B (2012) The Role of Background Selection in Shaping Patterns of Molecular
585 Evolution and Variation: Evidence from Variability on the Drosophila X Chromosome. *Genetics*,
586 **191**, 233–246.

587 Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity
588 among species. *Nature Reviews Genetics*, **14**, 262–274.

589 Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS)
590 approach for high diversity species. (L Orban, Ed.). *PLoS ONE*, **6**, e19379.

591 Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**,
592 342–350.

593 Feulner PGD, Chain FJ, Panchal M *et al.* (2015) Genomics of divergence along a continuum of
594 parapatric population differentiation. (J Zhang, Ed.). *PLoS Genetics*, **11**, e1004966.

595 Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT (2015) Genome Assembly
596 Improvement and Mapping Convergent Evolutionary Traits in Sticklebacks with
597 Genotyping-by-Sequencing. *G3: Genes | Genomes | Genetics*, **5**, 1463–1472.

598 Gow JL, Peichel CL, Taylor EB (2006) Contrasting hybridization rates between sympatric three-
599 spined sticklebacks highlight the fragility of reproductive barriers between evolutionarily young
600 species. *Molecular Ecology*, **15**, 739–752.

601 Hairston NG Jr, Ellner SP, Geber MA, Yoshida T, Fox JA (2005) Rapid evolution and the
602 convergence of ecological and evolutionary time. *Ecology Letters*, **8**, 1114–1127.

603 Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in
604 sticklebacks. *Journal of Fish Biology*, **75**, 2000–2036.

605 Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in
606 Threespine Stickleback using Sequenced RAD Tags (DJ Begun, Ed.). *PLoS Genetics*, **6**.

607 Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in
608 threespine sticklebacks. *Nature*, **484**, 55–61.

609 Kautt AF, Elmer KR, MEYER A (2012) Genomic signatures of divergent selection and speciation
610 patterns in a “natural experiment,” the young parallel radiations of Nicaraguan crater lake cichlid
611 fishes. *Molecular Ecology*, **21**, 4770–4786.

612 Kirkpatrick M (2016) The Evolution of Genome Structure by Natural and Sexual Selection. *The*
613 *Journal of heredity*, esw041.

614 Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics*,
615 **173**, 419–434.

616 Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, **17**,
617 183–189.

618 Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform.
619 *Bioinformatics*, **26**, 589–595.

620 Lowry DB, Hoban S, Kelley JL *et al.* (2016) Breaking RAD: An evaluation of the utility of restriction
621 site associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*.

622 Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis
623 of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B:*
624 *Biological Sciences*, **363**, 3009–3021.

625 Löytynoja A, Goldman N (2008) Phylogeny-Aware Gap Placement Prevents Errors in Sequence
626 Alignment and Evolutionary Analysis. *Science*, **320**, 1632–1635.

627 Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of

628 Illumina sequence reads. *Genome Research*, **21**, 936–939.

629 Marko PB, Hart MW (2011) The complex analytical landscape of gene flow inference. *Trends in*
630 *Ecology & Evolution*, **26**, 448–456.

631 Marques DA, Lucek K, Meier JI *et al.* (2016) Genomics of Rapid Incipient Speciation in Sympatric
632 Threespine Stickleback. *PLoS Genetics*, **12**, e1005887.

633 McKinnon JS, Rundle HD (2002) Speciation in nature: the threespine stickleback model systems.
634 *Trends in Ecology & Evolution*, **17**, 480–487.

635 Miller SE Intraguild predation is a mechanism of divergent selection in the threespine stickleback.
636 University of British Columbia, Vancouver.

637 Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical
638 predictions and empirical results from rabbits and mice. *Philosophical transactions of the Royal Society*
639 *of London. Series B, Biological sciences*, **367**, 409–421.

640 Narum SR, Hess JE (2011) Comparison of F(ST) outlier tests for SNP loci under selection. *Molecular*
641 *Ecology Resources*, **11 Suppl 1**, 184–194.

642 Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: A new twist
643 on chromosomal speciation. *Evolution*, **57**, 447–459.

644 Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press.

645 Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of
646 restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

647 Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Reviews Genetics*, **7**,
648 851–861.

649 Noor MA, Grams KL, Bertucci LA, Reiland J (2001a) Chromosomal inversions and the
650 reproductive isolation of species. *Proceedings of the National Academy of Sciences*, **98**, 12084–12088.

651 Noor M, Cunningham AL, Larkin JC (2001b) Consequences of Recombination Rate Variation on
652 Quantitative Trait Locus Mapping Studies: Simulations Based on the *Drosophila melanogaster*
653 Genome. *Genetics*, **159**.

654 Nosil P, Harmon LJ, Seehausen O (2009) Ecological explanations for (incomplete) speciation. *Trends*
655 *in Ecology & Evolution*, **24**, 145–156.

656 Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R
657 language. *Bioinformatics*, **20**, 289–290.

658 Peichel CL, Marques DA (2017) The genetic and molecular architecture of phenotypic diversity in
659 sticklebacks. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **372**,
660 20150486.

661 Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J (2016) Construction of Ultradense Linkage
662 Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. *Genome biology and*
663 *evolution*, **8**, 78–93.

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

666 Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**,
667 351–358.

668 Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary
669 diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*,
670 **21**, 2852–2862.

671 Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome-patterns
672 and consequences. *Molecular Ecology*, **22**, 3014–3027.

673 Rolshausen G, Muttalib S, Kaeuffer R *et al.* (2015) When maladaptive gene flow does not increase
674 selection. *Evolution*, **69**, 2289–2302.

675 Schluter D (1993) Adaptive Radiation in Sticklebacks: Size, Shape, and Habitat Use Efficiency.

- 676 *Ecology*, **74**, 699.
- 677 Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of*
678 *Sciences of the United States of America*, **106 Suppl 1**, 9955–9962.
- 679 Schluter D, Rambaut A (1996) Ecological Speciation in Postglacial Fishes [and Discussion].
680 *Philosophical Transactions of the Royal Society B: Biological Sciences*, **351**, 807–814.
- 681 Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene
682 flow. *Nature Reviews Genetics*, **14**, 404–414.
- 683 Taylor EB, McPhail JD (2000) Historical contingency and ecological determinism interact to prime
684 speciation in sticklebacks, *Gasterosteus*. *Proceedings of the Royal Society B: Biological Sciences*, **267**,
685 2375–2384.
- 686 Team RC (2015) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical
687 *Computing, Vienna, 2012*). URL: <http://www.R-project.org>.
- 688 Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide
689 insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.
- 690 Vavrek MJ (2011) fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia*
691 *Electronica*, **14**.
- 692 Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure.
693 *Evolution*, **38**, 1358.
- 694 Wielgoss S, Barrick JE, Tenaillon O *et al.* (2011) Mutation Rate Inferred From Synonymous
695 Substitutions in a Long-Term Evolution Experiment With *Escherichia coli* (BJ Andrews, Ed.).
696 *G3: Genes | Genomes | Genetics*, **1**, 183–186.
- 697 Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and*
698 *Evolution*, **24**, 1586–1591.
- 699 Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection
700 balance. *Evolution*, **65**, 1897–1911.
- 701 Yoshida K, Makino T, Yamaguchi K *et al.* (2014) Sex Chromosome Turnover Contributes to
702 Genomic Divergence between Incipient Stickleback Species (J Zhang, Ed.). *PLoS Genetics*, **10**,
703 e1004223.

704 **Data Accessibility**

706 **Published genomic datasets:** The original study references and accession numbers are listed in
707 Table S1. **New genomic datasets:** All new datasets will be made available on the SRA. **Analysis**
708 **code and processed data:** https://github.com/ksamuk/gene_flow_linkage.

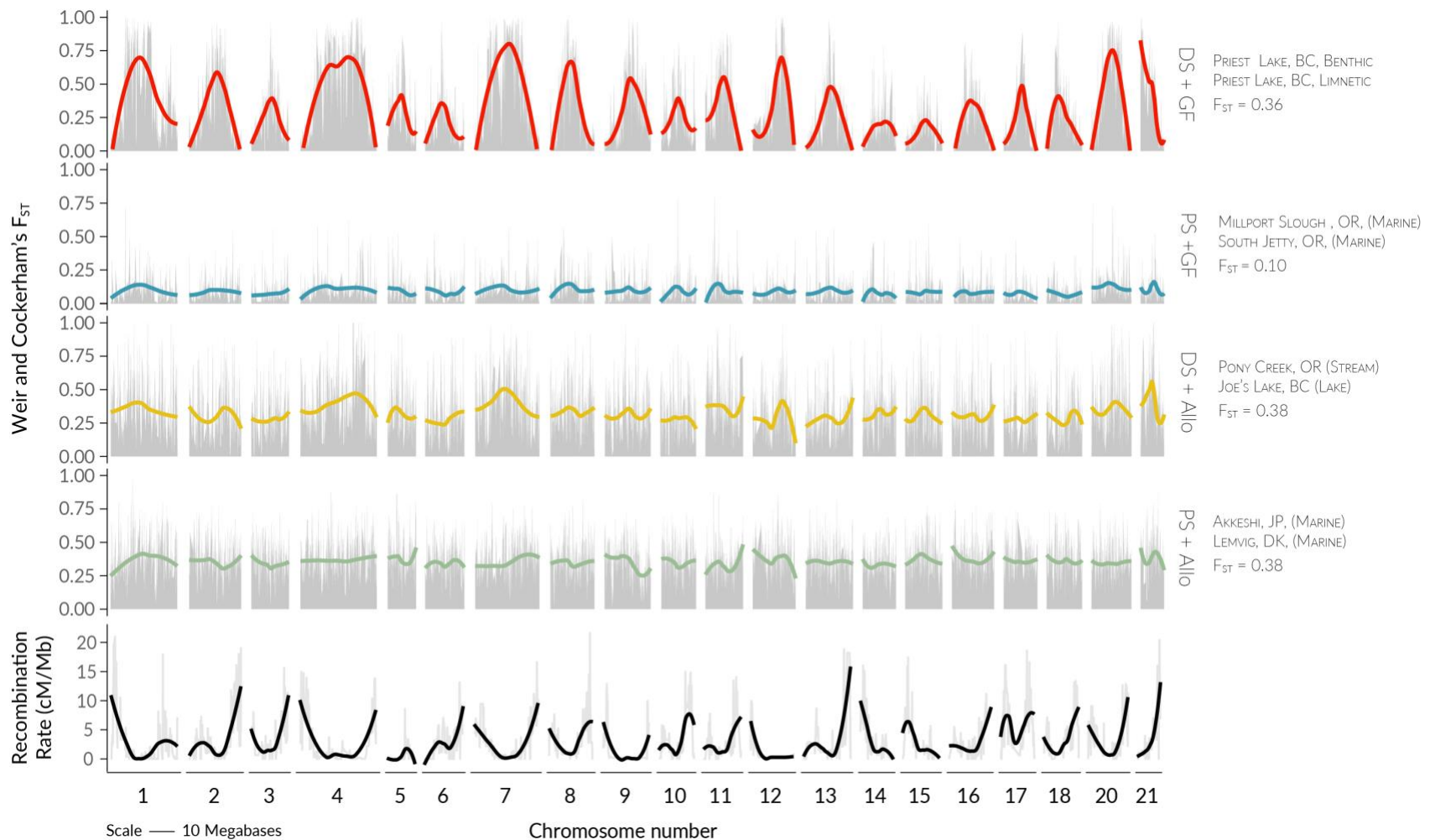


Figure 1 | Representative plots of genome-wide F_{ST} between single pairs of populations from four gene-flow and selection regimes. Each coloured line represents a loess smooth of F_{ST} vs. chromosomal position for a single chromosome (numbered along bottom). Raw F_{ST} (calculated in 75000 base-pair windows) is depicted in grey behind each smoothed line. Line color corresponds to gene flow and selection regime (labeled on the right side of the plot). Below the main plots, recombination rate estimates from Roesti et al. (2013) (black lines) are shown for each chromosome. Population pairs were chosen on the basis of similarity in overall F_{ST} and coverage of genomic data. Detailed additional statistics (diversity, d_{xy} , d_S , etc.) for each representative comparison are provided in supplemental figures S6-S9.

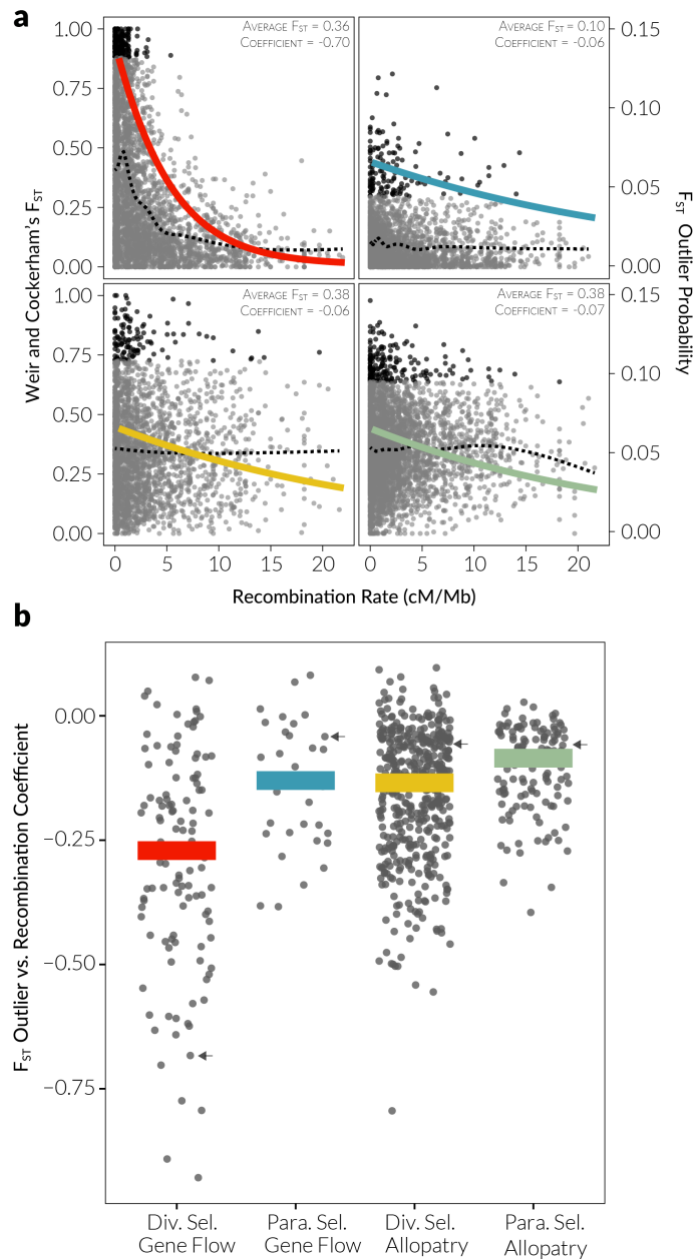


Figure 2 | Patterns of low recombination bias among the four gene flow and selection regimes. (a) Representative logistic regressions of outlier status against recombination rate. Each panel corresponds to a population shown in Figure 1. Regressions are corrected for variation in mutation rate and gene density. **(b)** Individual logistic regression coefficients for all pairwise comparisons (points) in each gene flow / selection regime. Colored horizontal lines indicate means. Increasingly negative coefficients indicate a stronger bias for outliers to occur in the regions of low recombination. Black arrows indicate the coefficient of each representative comparison used in Figure 1 and panel (a) above.

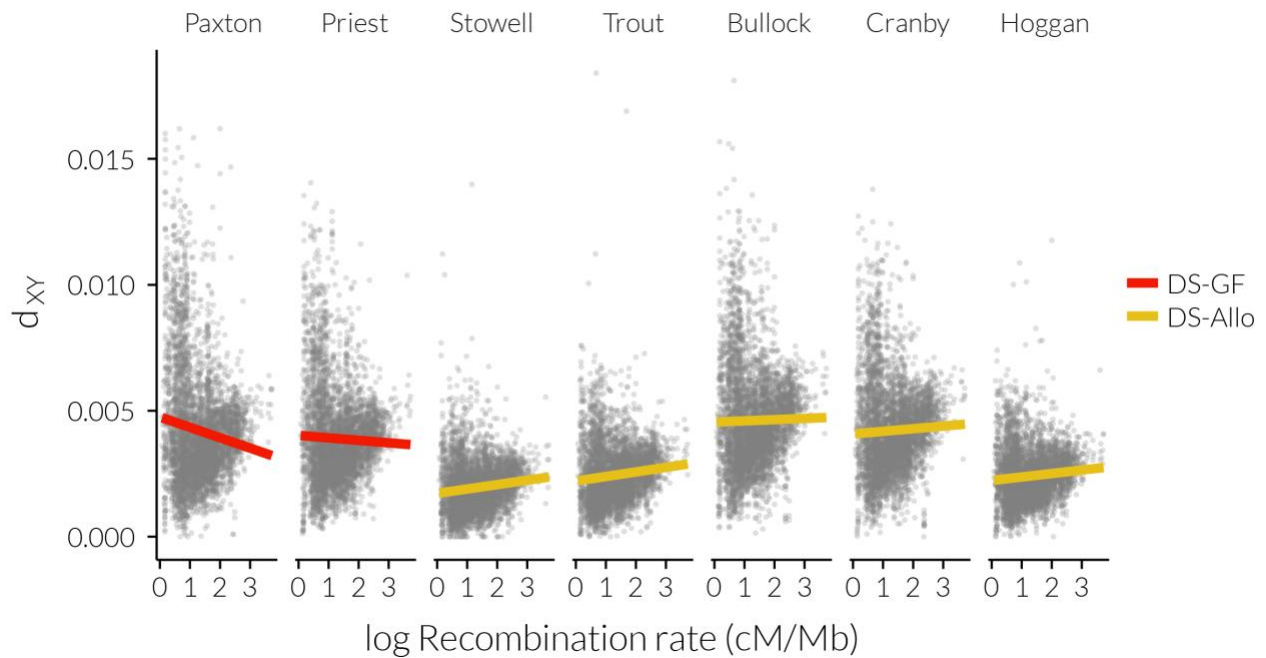


Figure 3 | The relationship between recombination rate and d_{xy} estimated from whole genome sequence from seven pairs of stickleback populations. Each panel depicts the relationship between recombination rate and d_{xy} in a single population, calculated by comparing the whole genome sequences of two individuals. Each point represents the value of d_{xy} in a single 1000 bp window. Points have been randomly down-sampled by a factor of 100 to aid in visualization. Colored lines represent lines of best fit. DS-GF (red) comparisons represent d_{xy} between two sympatric populations (a single benthic/limnetic pair), whereas DS-Allopatry (yellow) comparisons represent d_{xy} between two allopatric populations (solitary lake vs. marine). Values on the x axis were transformed via $\log(\text{value} + 1)$.

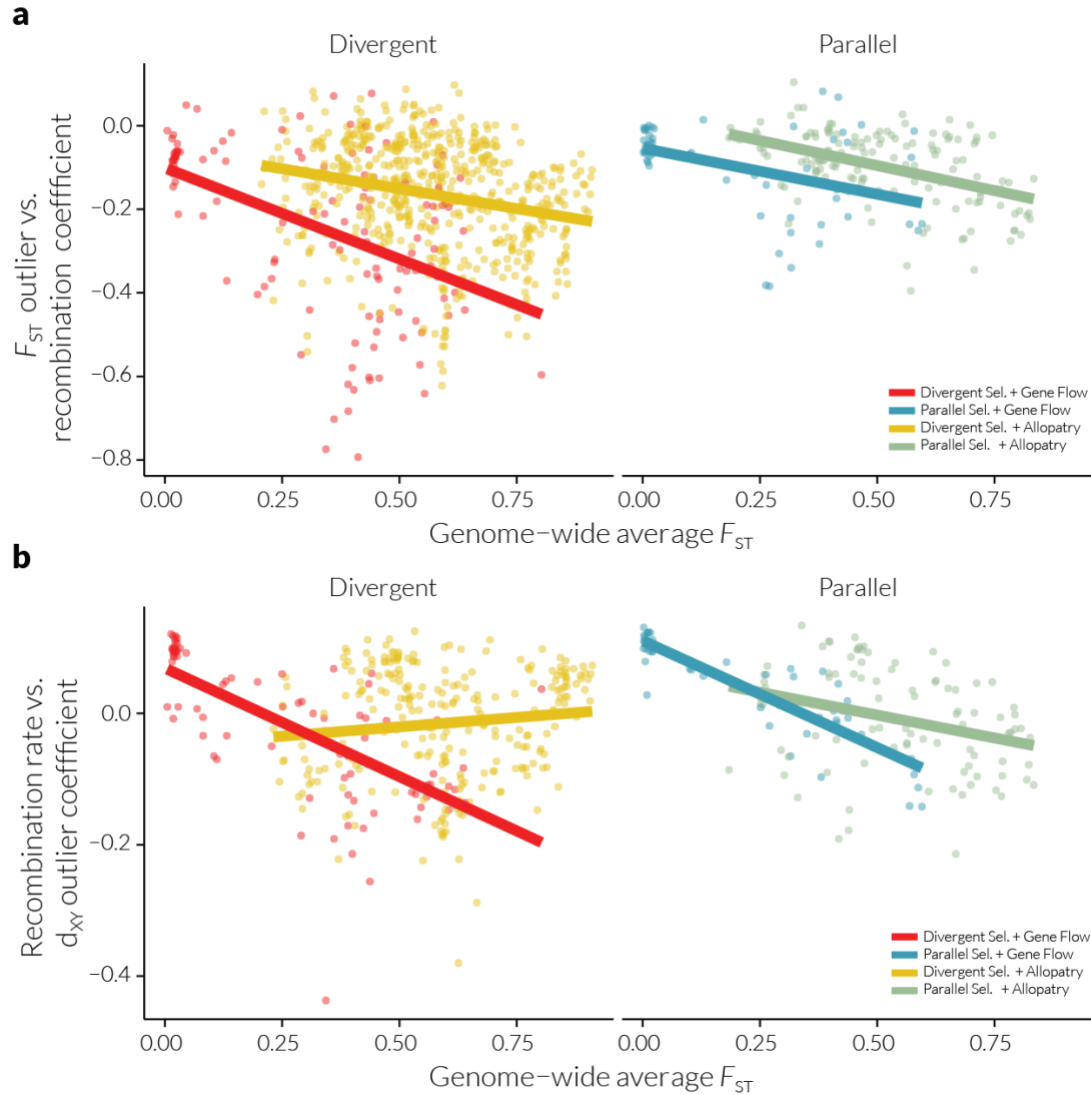


Figure 4 | The relationship between the tendency for divergence outliers to occur in regions of low recombination (y-axis) and overall genetic divergence (x-axis) when measured for (a) the F_{ST} outliers and (b) d_{XY} outliers. Y-axis values are regression coefficients derived by performing logistic regressions of outlier probability vs. recombination rate for 75 kb genomic windows in each comparison. X-axis values are averages of F_{ST} at all loci across the genome for each comparison. Each point represents a single comparison of two populations. Colors indicate different gene flow + selection regimes, with divergent and parallel selection separated for clarity in each of (a) and (b).