Title

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

Permalink
https://escholarship.org/uc/item/88z1q14t

Journal
Molecular Ecology, 26(17)

ISSN
0962-1083

Authors

Samuk, Kieran
Owens, Gregory L
Delmore, Kira E
et al.

Publication Date
2017-09-01

DOI
10.1111/mec.14226

Peer reviewed

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

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

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

To approach this problem, we assembled a large population genomic dataset of threespine sticklebacks (*Gasterosteus aculeatus*) from across the northern hemisphere (Figure S1, Table S1). Threespine sticklebacks are a holarctic species of fish that have evolved into a variety of unique forms over the last 10,000 years (McKinnon & Rundle 2002). Notably, the various forms of stickleback have evolved repeatedly in the presence and absence of gene flow (McKinnon & Rundle 2002). This allows for statistical comparisons of the genomic distribution of adaptive alleles among groups of population pairs experiencing varying levels of divergent selection and gene flow. Here, we focused on comparing population pairs in which divergent selection occurs in the face of gene flow to population pairs experiencing selection alone, gene flow alone, or neither. Using this approach, we tested the theoretical prediction that when divergent selection and gene flow co-occur, adaptive alleles are more likely to fix in regions of low recombination and/or occur in tightly linked clusters throughout the genome.

Results

Population genomic dataset

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

Localizing candidates for adaptive divergence

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

Divergence in regions of low recombination

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

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

d_{XY} outliers also showed a tendency (albeit non-significant) to occur most often in regions of low recombination (Figure S2; permutation test: two-sided $p = 0.475$). That said, our estimates of d_{XY} from GBS/RAD dataset had considerable levels of noise, likely due to low marker density in the 75kb windows. We thus repeated the d_{XY} analysis, but restricted the analysis to whole genome datasets (see Methods). Using this reduced dataset and 75 kb windows, we found that the relationship between d_{XY} (both outlier status and mean d_{XY}) and recombination was negative in DS-GF comparison and positive in DS-Allo comparisons (Figure 3). This difference in slopes between regimes was highly significant (likelihood ratio test: $\chi^2_2 = 28.85$, $p = 5.41\times10^{-5}$). Thus, DS-GF comparisons exhibited unusually high levels of both relative and absolute divergence in regions of low recombination.

Ruling out potential sources of bias

Discretization of geographic distance

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

Differences in genome-wide F_{ST}

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

Differences in heterozygosity vs. recombination among regimes

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

Clustering of outlier SNPs

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

DS-GF population pairs showed more clustering of F_{ST} outlier SNPs than population pairs in other gene flow/selection regimes (Figure S4). Specifically, DS-GF outlier SNPs were on average approximately one standard deviation closer together in map distance than expected on the basis of overall SNP density (Figure S4, permutation test: two-sided $p < 0.0001$). Coefficients of variation for the distance between F_{ST} outlier SNPs showed similar results (Figure S4, permutation test: two-sided $p < 0.0001$), again indicating the highest levels of clustering in DS-GF comparisons.

Discussion

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

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

Note that the analyses presented here were not designed to detect changes in genome structure or the modification of recombination rate among populations. We assume that recombination rates are highly conserved between threespine stickleback populations. This is likely a reasonable assumption given that (a) recombination maps are highly similar among threespine stickleback populations from Europe and the United States (Roesti *et al.* 2013; Glazer *et al.* 2015) and (b) homologous chromosomes in the distantly-related ninespine stickleback show very similar patterns of recombination (Rastas *et al.* 2016). While modification of recombination can be important in some systems, our results pertain to the (likely far more common) scenario in which many loci with potentially varying linkage relationships underlie adaptation and DS-GF favors genetic architectures in which adaptive alleles are tightly linked over other architectures (Yeaman & Whitlock 2011). Future studies could extend our framework to study how gene flow shapes the evolution of recombination rate and genome structure.

The costs of low recombination

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

Heterogenous genomic divergence

Our findings also suggest that the patterns of heterogenous genomic divergence observed in many speciation studies (Marko & Hart 2011; Feder *et al.* 2012) may be partly a product of the interaction between gene flow and selection. Explaining this phenomenon has become a major question in speciation genetics, and many recent studies have shown that patterns of heterogenous divergence in the genome are correlated with recombination rate (Roesti *et al.* 2013; Renaut *et al.* 2013; Burri *et al.* 2015). The association between diversity, divergence and recombination is widely thought to be the result of linked selection, i.e. background selection and hitchhiking (Charlesworth 2012). Our results suggest that there is a general negative association between recombination rate and both diversity and divergence (probably generated by background selection) and this relationship can be further shaped by the effects of selection (hitchhiking) and gene flow (decay of divergence in regions of high recombination and/or favoring linkage between adaptive alleles).

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

The effect of divergent selection

Divergent selection alone is predicted to generate a correlation between recombination rate and genomic divergence across the genome (Barton 2010). This effect is particularly apparent in reduced representation datasets, such as the RAD and GBS datasets we analyzed here (Lowry *et al.* 2016). Our data support this prediction: all “divergent selection” comparisons (DS-GF and DS-Allo) show increased divergence in regions of low recombination (e.g. Figure 2B, red and yellow lines). However, the divergence-recombination correlation is significantly more negative in DS-GF populations, which we interpret as a unique joint effect of gene flow and divergent selection. Note that this pattern held when the analysis was restricted to whole-genome data (Figure 3), suggesting that low marker density is not the sole source of the low-recombination bias (although undoubtedly a contributor). Interestingly, gene flow alone (e.g. parallel selection + gene flow, blue lines in Figures 2 and 4) appears to not be sufficient to generate a bias for divergence in regions of low recombination.

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

Caveats

The main strength of the approach we applied here was that it allowed for replication within each gene-flow/selection regime, which is necessary for examining statistical differences between regimes in their recombination bias. However, the number of comparisons involved (1000+) also created serious computational bottlenecks, which precluded using more sophisticated methods for detecting natural selection and gene flow (Aeschbacher *et al.* 2016). Further, we do not have detailed knowledge of the demographic history and historical rates of introgression between any of the populations studied here. Both of these factors are known to affect patterns of divergence, and can potentially alter the relationship between divergence and recombination (Tine *et al.* 2014). It is possible that the more extreme recombination vs. divergence bias we observed in DS-GF populations was a result of an unusual demographic or introgression history that was somehow confounded with the contemporary "DS-GF" classification. For example, these comparisons may be enriched for populations that have experienced a period of allopatry, followed by the resumption of gene flow (secondary contact). However, this would still imply that divergent selection and gene flow interact to generate a low-recombination bias, as loci not involved in divergent selection should still flow freely between populations. Thus, while the mechanistic details behind the patterns we describe here are still unclear, we hope our study stimulates further studies of the relationship between gene flow, selection and recombination in shaping patterns of divergence.

Acknowledgements

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

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

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

Methods

Github Repository

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

Data Sources

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

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

Variant identification and processing

We identified variants using a standard, reference-based bioinformatics pipeline (see Github code repository for details). After demultiplexing, we used Trimmomatic v0.32 (Bolger *et al.* 2014) to filter low quality sequences and adapter contamination. We then aligned reads to the stickleback reference genome (BROAD S1, (Jones *et al.* 2012) using BWA v0.7.10 (Li & Durbin 2010), followed by realignment with STAMPY v1.0.23 (Lunter & Goodson 2011). We then followed the GATK v3.3.0 (Cachat *et al.* 2010) best practices workflow except that we skipped the MarkDuplicates step when reads were derived from reduced representation libraries (RAD and GBS). We realigned reads around indels using RealignTargetCreator, and IndelRealigner, identified variants in individuals using the HaplotypeCaller, and each dataset using GenotypeGVCFs. The results were sent to a VCF file containing all variant and invariant sites and converted to tabular format. All datasets were combined for processing.

Calculation of divergence metrics

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

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

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

Classification of Populations

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

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

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

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

Addition of Genomic Variables

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

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

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

Tendency for adaptive divergence in regions of low recombination

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

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

Permutation tests

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

Clustering vs. geographic distance and overall divergence

To ensure our results were not influenced by our discrete geographic categorization scheme, we examined how the tendency for F_{ST} outliers to occur in regions of low-recombination varied with pairwise geographic distance. To do this, we regressed the low recombination outlier tendency (regression coefficients from above) on geographic distance between populations using the R function “lm”. The linear model was of the form recombination bias = distance + ecology + distance * ecology (interaction). We then assessed significance of the model terms using a permutation test similar to the one previously described (see code supplement)

The results of (Burri *et al.* 2015) and (Roesti *et al.* 2013) suggest that the tendency for F_{ST} outliers to occur in regions of low recombination may be highest at intermediate levels of overall genetic divergence

(F_{ST} = 0.3-0.5). Overall F_{ST} thus represents a potential source of bias, as our use of geographic distance as a proxy for gene flow is naturally confounded with overall F_{ST} – with isolation by distance, more distant populations will have higher divergence, all else being equal. To test if this may have influenced our results, we examined the correlation between low-recombination clustering tendency and overall F_{ST}. To obtain overall F_{ST} estimates between each pair of populations, we divided the sum of the numerator terms by the sum of the denominator terms of all locus-specific F_{ST} values for each pair (Weir & Cockerham 1984). This yielded a single average F_{ST} value for each pair of populations. We then employed the same approach as the analysis of distance, with a linear model the form recombination bias = F_{ST} + ecology + F_{ST} * ecology (interaction). We assess the significance of this difference again via permutation test (see code supplement).

Increased clustering of outlier SNPs

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

We first asked: do map distances between nearest-neighbour outlier loci differ significantly from the expected map distances of identical numbers of nearest-neighbour SNPs? This approach was designed to account for disparities in SNP density that might occur due to differences in sequencing outcomes between our various datasets. To do this, we first partitioned each SNP data set by chromosome. Then, for each chromosome we identified the number of outlier loci using the previously described method. We then drew 10,000 samples of random SNPs from each chromosome equal to the number of outliers on that chromosome, and calculated the mean map distance between each SNP and its nearest neighbor in the random sample. We then compared the empirical mean nearest neighbor map distance of outliers to this null distribution for each chromosome within each individual comparison dataset. We then used permutation tests to compare (a) the proportion of chromosomes that were significantly over-clustered and (b) the difference between the average NND between outliers and the average NND expected between SNPs, in units of standard deviations, between the four selection and gene flow regimes.

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

Whole genome data collection

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

Whole genome d_{XY} calculation and analysis

We used the GATK best practices workflow described above to call variants on the eight populations above (not including the reference). We emitted VCF files containing all variant and invariant sites for each population. We then computed d_{XY} in 75,000 base pair windows using the method described previously (see "Calculation of Divergence Metrics" above; code available in repository). For the two pairs of DS-GF populations (Paxton and Priest), we computed d_{XY} between sympatric populations within each lake. For the remaining DS-Allopatry populations, we computed d_{XY} between each population and a marine population (Bear Paw Lake, i.e. the reference genome). We allowed for missing sites, and for windows with no variable sites. Prior to analysis, we inspected relationships between the number of genotyped sites in each window and d_{XY}. We found that the variance in d_{XY} was highly inflated in windows containing fewer than 7500 genotyped sites (variant and invariant). We thus excluded all windows with less than 7500 sites (out of 75,000) from the analysis. As before, we classified windows with dxy values exceeding the 95th percentile as "outlier windows".

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

References

Aeschbacher S, Selby JP, Willis JH, Coop G (2016) *Population-genomic inference of the strength and timing of selection against gene flow*. Cold Spring Harbor Labs Journals.

Barton NH (2010) Genetic linkage and natural selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 2559–2569.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, **25**, 1656–1665.

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

Cachat JM, Stewart A, Utterback E *et al.* (2010) Deconstructing Adult Zebrafish Behavior with Swim Trace Visualizations. In: *Zebrafish Ecology and Behaviour*, pp. 191–201. Humana Press, Totowa, NJ.

Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J (2016) Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F 2 Recombinant Crosses as an Example. *Genome biology and evolution*, **8**, 78–93.

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

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

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

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

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

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

Ecology, **74**, 699.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106 Suppl 1**, 9955–9962.

Schluter D, Rambaut A (1996) Ecological Speciation in Postglacial Fishes [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **351**, 807–814.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Taylor EB, McPhail JD (2000) Historical contingency and ecological determinism interact to prime speciation in sticklebacks, Gasterosteus. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 2375–2384.

Team RC (2015) *R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2012)*. URL: http:// www. R-project. org.

Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.

Vavrek MJ (2011) fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, **14**.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358.

Wielgoss S, Barrick JE, Tenaillon O *et al.* (2011) Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli (BJ Andrews, Ed,). *G3: Genes|Genomes|Genetics*, **1**, 183–186.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution*, **65**, 1897–1911.

Yoshida K, Makino T, Yamaguchi K *et al.* (2014) Sex Chromosome Turnover Contributes to Genomic Divergence between Incipient Stickleback Species (J Zhang, Ed,). *PLoS Genetics*, **10**, e1004223.

Data Accessibility

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

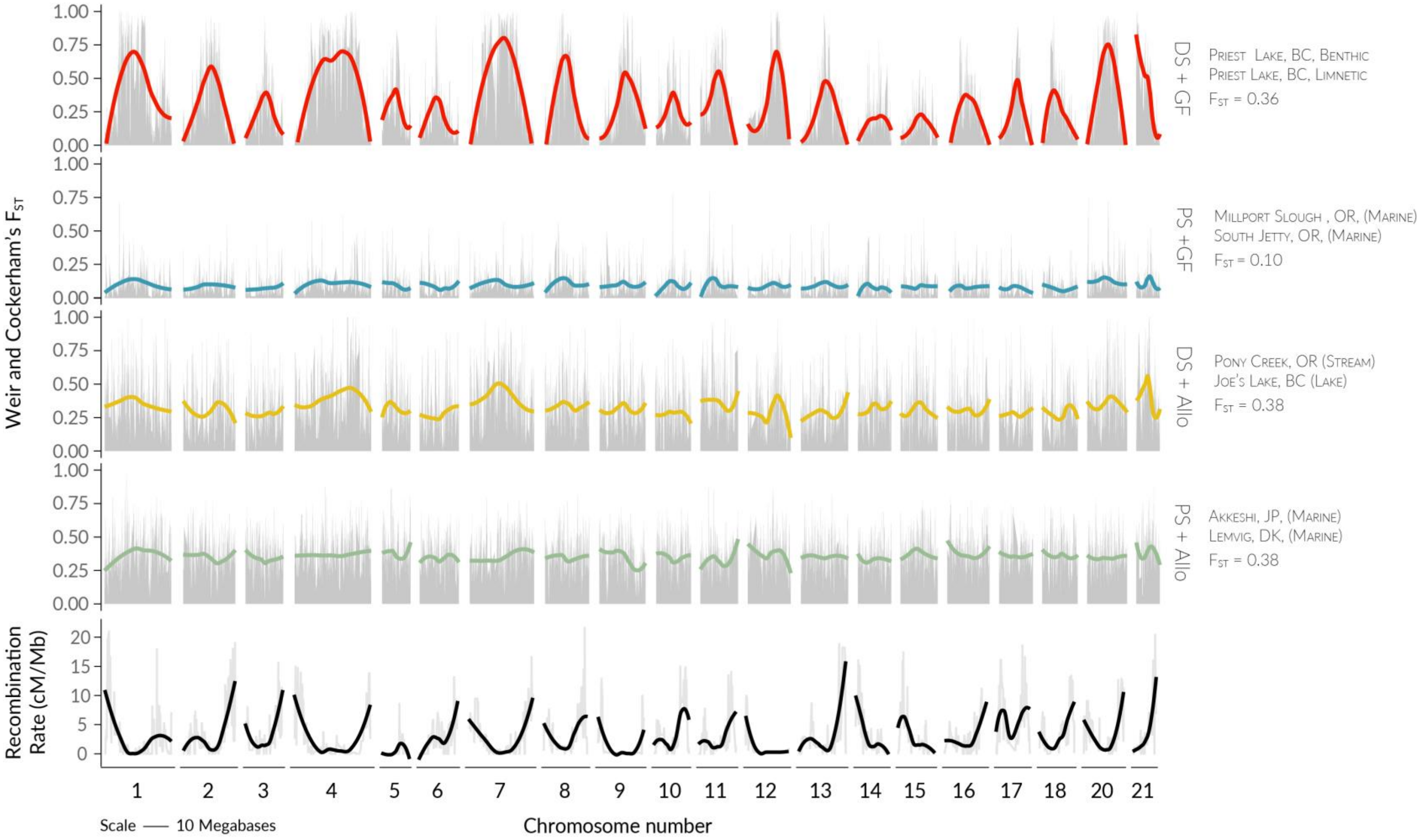

Figure 1 | Representative plots of genome-wide F_{ST} between single pairs of populations from four gene-flow and selection regimes. Each coloured line represents a loess smooth of F_{ST} vs. chromosomal position for a single chromosome (numbered along bottom). Raw F_{ST} (calculated in 75000 base-pair 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, dxy, dS, etc.) for each representative comparison are provided in supplemental figures S6-S9.

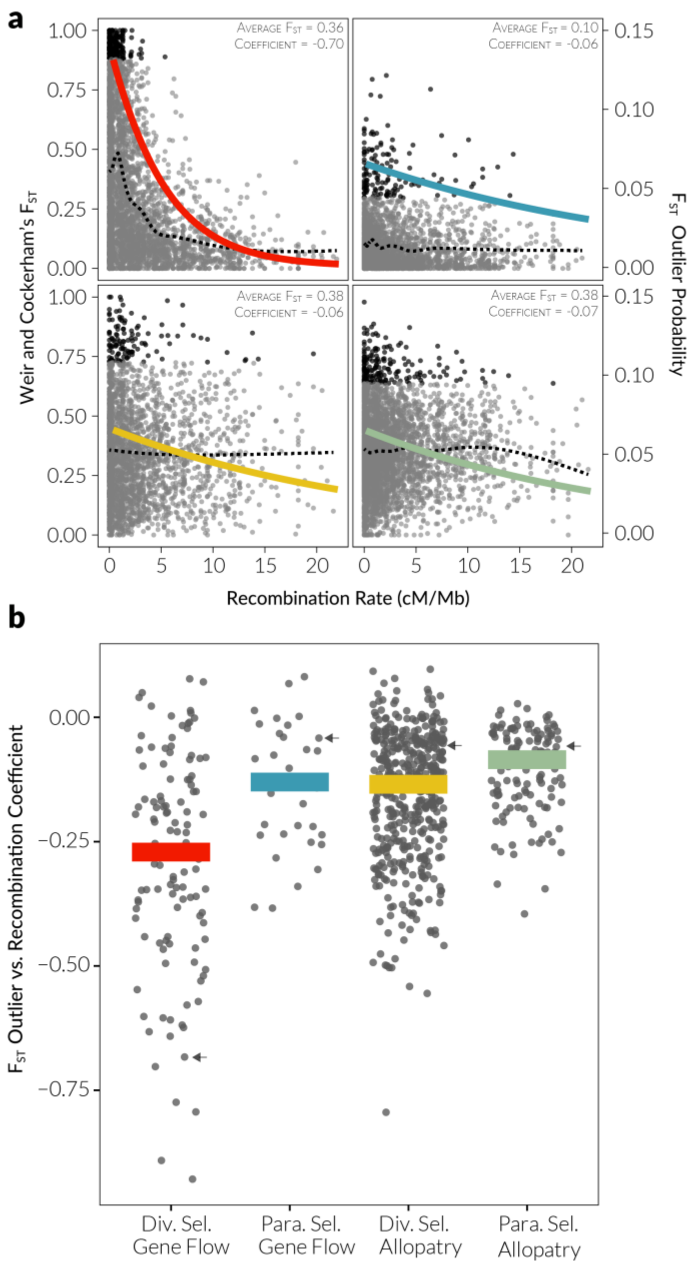

Figure 2 | Patterns of low recombination bias among the four gene flow and selection regimes. **(a)** Representative logistic regressions of outlier status against recombination rate. Each panel corresponds to a population shown in Figure 1. Regressions are corrected for variation in mutation rate and 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.

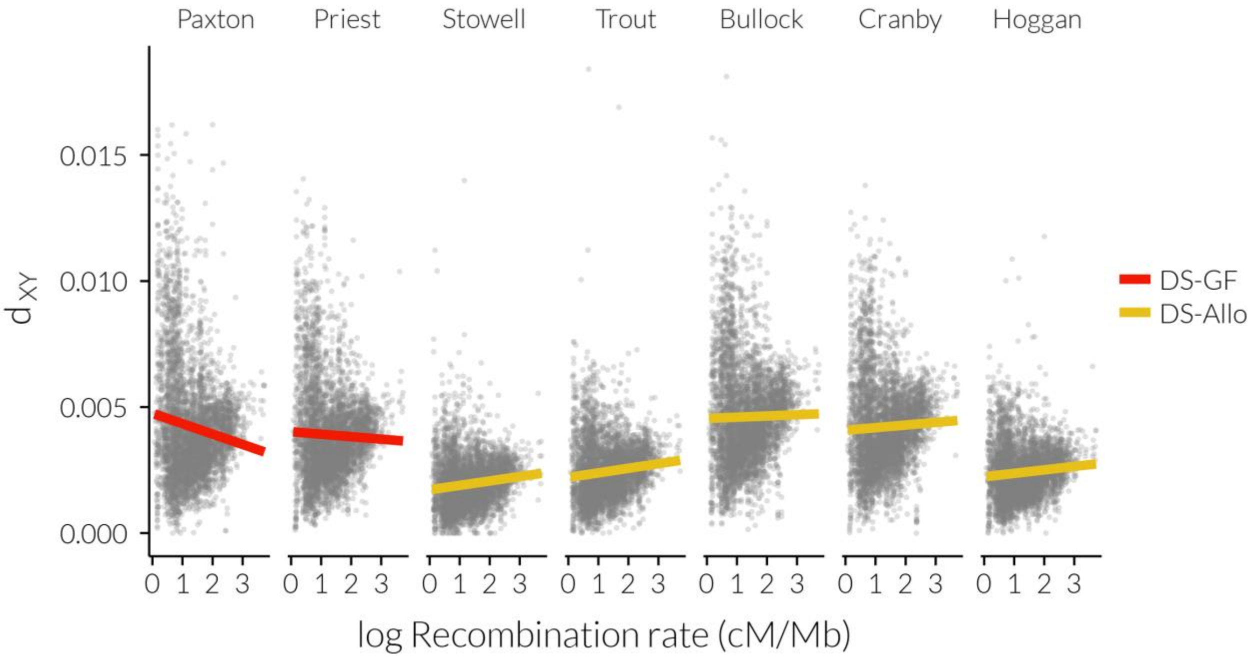

Figure 3 | The relationship between recombination rate and d_{XY} estimated from whole genome sequence from seven pairs of stickleback populations. Each panel depicts the relationship between recombination rate and d_{XY} in a single population, calculated by comparing the whole genome sequences of two 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 dxy between two allopatric populations (solitary lake vs. marine). Values on the x axis were transformed via log(value + 1).

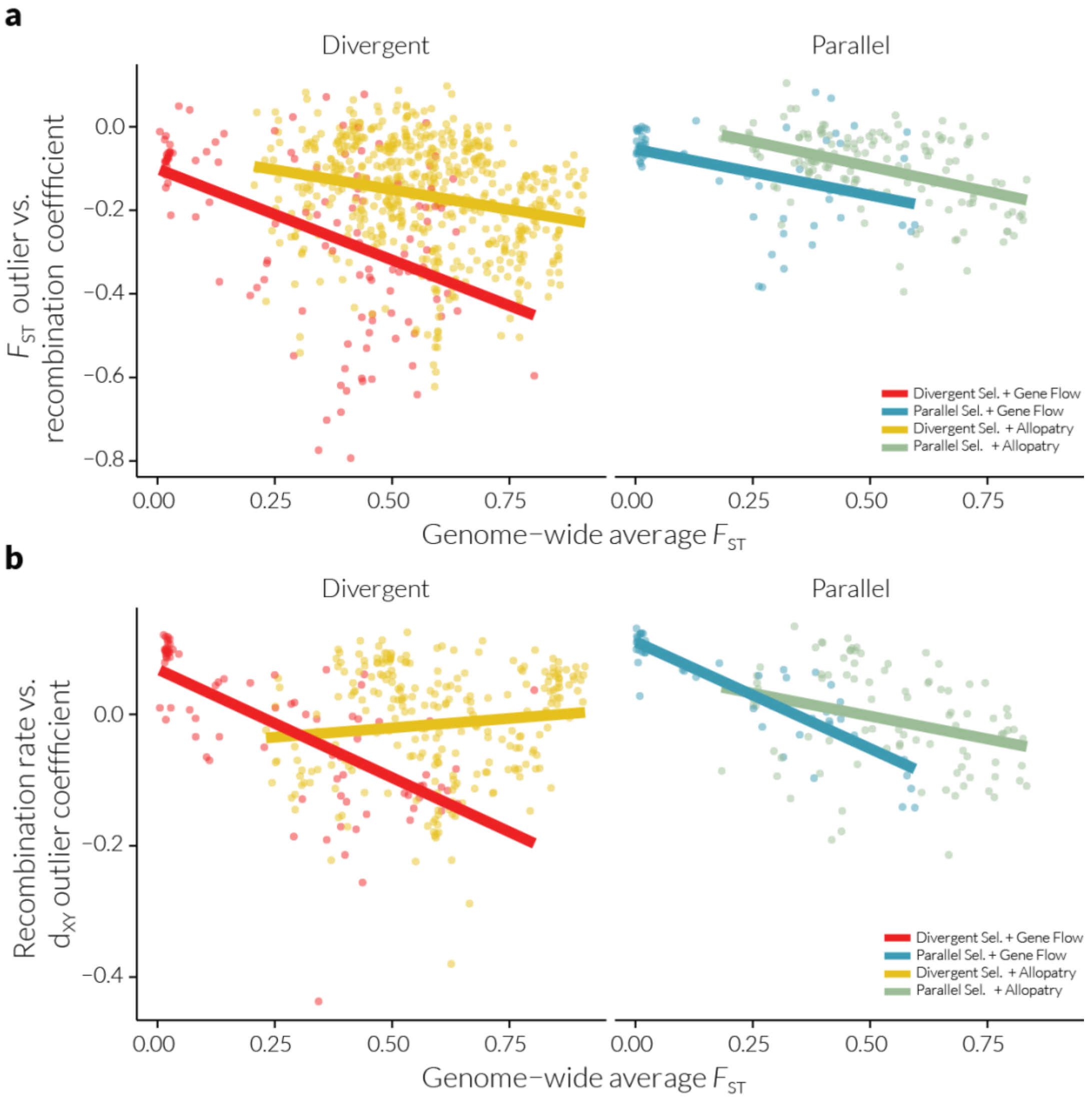

Figure 4 | The relationship between the tendency for divergence outliers to occur in regions of low recombination (y-axis) and overall genetic divergence (x-axis) when measured for (a) the F_{ST} outliers and (b) d_{XY} outliers. Y-axis values are regression coefficients derived by performing logistic 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).