Locus-specific introgression in young hybrid swarms

Citation for published version:

Digital Object Identifier (DOI):
10.1111/mec.15862

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Molecular Ecology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
INTRODUCTION

The rate of hybridization between closely related species that have recently come into secondary contact is increasing, due to human-assisted migration and environmental change (Grabenstein & Taylor, 2018; Parmesan & Yohe, 2003). While such hybridization is not necessarily negative (Hamilton & Miller, 2016), in many cases hybridization can cause problems for native species. If F1s are
inviable or sterile then hybridization is a loss of reproductive effort (Allendorf et al., 2001). Alternatively, the presence of viable, fertile hybrid offspring can lead to populations with large numbers of hybrids, and in the most extreme cases, whole populations comprised only of hybrid individuals (Allendorf et al., 2001). Biodiversity can be lost through hybridization, either if all remaining members of a species are hybrids (extinction via hybridization; Allendorf et al., 2001; Allendorf & Luikart, 2009; Rhymer & Simberloff, 1996; Todesco et al., 2016), or if particular endemic alleles are replaced by novel alleles introduced by backcrossing and driven to fixation via selection (as described by Petit, 2004).

Hybrid zones, whether naturally occurring or due to human interference, can be used as “natural laboratories” for research into selection and the genetics of reproductive isolation between species (Hewitt, 1988). The rate of introgression of alleles between species is expected to be heterogenous across the genome, reflecting variation in selection (Baack & Rieseberg, 2007). Backcrossing coupled with recombination will separate haplotypes that are commonly found together and create novel haplotypes on which selection can act on alleles in unique genetic backgrounds (Arnold et al., 1999). Alleles that move quickly across the species barrier are assumed to be under positive selection in their new genetic background, while alleles that do not introgress between species are candidates for contributing to reproductive isolation (Baack & Rieseberg, 2007). Drift will also be acting on these alleles, particularly if hybridization is rare or one of the parental populations is small. In these cases, we expect substantial variation in the degree of introgression across loci, as a result of the sampling error introduced by reproduction and recombination (Baird et al., 2003). If non-native alleles are increasing in frequency, whether due to selection or drift, we should apply the precautionary principle until we can determine the effects (positive or negative) of these alleles. Identifying those endemic alleles that are most likely to be replaced by novel alleles gives a target for policy makers to reflect upon and consider protecting.

Geographic cline analyses have been used to determine the extent of hybridization between two species at a contact zone (Barton & Gale, 1993; Barton & Hewitt, 1985). Traditionally, the width of these geographic gradients of allele frequencies can be used to infer selection on each allele as it introgresses from one species to another across a landscape (Mallet et al., 1990). Genomic clines, which replace geographic gradients with hybrid indices, have been used in the same way, and have the advantage that they can be applied even when hybrids have a mosaic distribution, or exist as a hybrid swarm (Gompert & Buerkle, 2011, 2012; Lexer et al., 2007). Genomic clines use a multinomial regression that predicts the probability of a particular genotype (θ) given a hybrid index (h), where:

$$\theta = h + (2h - h^2) \times (a + (\beta(2h) - 1))$$

Here, $a$ is analogous to the location of the cline centre and can be interpreted as the direction of introgression, i.e., a positive $a$ means excess ancestry from species A to species B and negative $a$ means excess ancestry from species B to A. $\beta$ is analogous to the width of the cline and can be interpreted as the strength of the barrier to gene flow (Janoušek et al., 2015). Positive $\beta$ is interpreted as a narrow cline, where introgression is impeded, and negative $\beta$ is a wide cline, where introgression is faster than expected based on the genomic expectation (Gompert & Buerkle, 2009).

$a$ and $\beta$ are not explicitly expected to covary with each other (although they are not fully independent), nor are $a$ and $\beta$ necessarily expected to covary with divergence estimates between the parental species in the system such as $F_{ST}$ (Charlesworth, 1998). However, those loci that are both highly diverged between species (i.e., high $F_{ST}$) and slow moving (large positive $\beta$) are good candidates for loci involved in reproductive isolation (Gompert & Buerkle, 2009; Lexer et al., 2007), particularly if they are not expected to be highly diverged because of other genomic constraints (i.e., recombination cold spots; Burri et al., 2015; Cruickshank & Hahn, 2014). Studies of naturally occurring hybridization regularly find many markers, spread across the genome, with significant $a$ and $\beta$ estimates, and typically find more loci that are significant for $\alpha$ than $\beta$ (but see Pulido-Santacruz et al., 2018 who found no divergent $\alpha$ or $\beta$ SNPs between either woodcreeper (Willisornis) or antbird (Xiphorhynchus) species pairs). For example, Janoušek et al., (2015) found that as many as 70% of SNPs diverged from genome-wide expectations in a house mouse (Mus musculus musculus and M. m. domesticus) hybrid zone, Parchman et al., (2013) using 59,100 SNPs found more than 1000 significant for $\alpha$ and more than 400 significant for $\beta$ between manakin species (Manacus candei and M. vinellinus), and Sung et al., (2018) reported ~30% of 45,384 SNPs with significantly diverged $\alpha$ and ~1% of SNPs with significantly diverged $\beta$ between iris species (Iris hexagona and I. fulva). The vast number of reported genome-wide SNPs with excess $\alpha$ and $\beta$ from many systems are unlikely to all be related to selection, especially given that selection must be extremely strong to be detected at the genome-wide level in artificial selection studies (e.g., Castro et al., 2019). Simulations of admixed populations that varied in population size found that, particularly with a population size of only 100, both $a$ and $\beta$ estimates could be quite variable, and when loci under selection were simulated, particularly when there was weak selection and low levels of admixture, there were high false discovery rates (Gompert & Buerkle, 2011). Before genomic regions can be considered candidates to be responding to selection, careful consideration of expectations due to nonselective forces must be undertaken (Gompert & Buerkle, 2011).

The red deer (Cervus elaphus) is an emblematic animal native to Scotland. It was named as one of “Scotland’s big 5” in a campaign to increase engagement with wildlife ran by the Scottish Government between 2013 and 2015 (NatureScot, 2016; Scottish Wildlife Trust, 2013), and is known for its large body size, large antlers and bright red summer coat. Red deer are abundant through much of Scotland and they are popular for hunting (deer stalking) and with tourists and unpopular for their ecological impacts, particularly on young trees. The red deer is currently a species of least concern, but the greatest
threat to them in Europe is introgression by Japanese sika (C. nippon; IUCN, 2020). Physically smaller sika were introduced to Scotland in the late 19th century, and have since hybridized with the red deer (Ratcliffe, 1987). In some parts of the Kintyre peninsula, Argyll, more than 40% of sampled phenotypic red deer and sika individuals are hybrids according to 50,000 SNP markers, with the majority being the result of multiple generations of backcrossing (McFarlane, Hunter, et al., 2020; McFarlane et al., 2020). Hybrid deer tend towards an intermediate phenotype and thus are smaller, have smaller antlers, and are more likely to have the spots typical of sika than parental species red deer (Bartos et al., 1981). Initial hybridization may be constrained by the substantial size difference between species, but it is clear that at least some F1s and many backcrosses are fertile (Harrington, 1979; McFarlane, Hunter, et al., 2020; McFarlane & Pemberton, 2019; McFarlane, Senn, et al., 2020). While there is a trend from red deer in the north to sika in the south of Kintyre, the distribution of hybrids does not follow a cline, being instead concentrated in specific areas (Senn et al., 2010), and we have thus recently redefined this system as a “bivariate hybrid zone” (McFarlane, Hunter, et al., 2020; McFarlane & Pemberton, 2019; McFarlane, Senn, et al., 2020). Additionally, in a study using 20 microsatellite markers, there was no evidence that the number of hybrid individuals was changing over a period of 15 years (Senn et al., 2010). Taken together, the red deer and sika system in Scotland is an excellent model for understanding how hybridization between a native and introduced species can play out genetically.

In this study, we sought evidence among red-sika hybrids that specific genome regions have introgressed more or less than expected under neutrality, in ways that might be interpreted as being due to selection. We estimated genetic divergence between red deer and sika in a native and introduced species red deer (Bartos et al., 1981). Initial hybridization may be constrained by the substantial size difference between species, but it is clear that at least some F1s and many backcrosses are fertile (Harrington, 1979; McFarlane, Hunter, et al., 2020; McFarlane & Pemberton, 2019; McFarlane, Senn, et al., 2020). Additionally, in a study using 20 microsatellite markers, there was no evidence that the number of hybrid individuals was changing over a period of 15 years (Senn et al., 2010). Taken together, the red deer and sika system in Scotland is an excellent model for understanding how hybridization between a native and introduced species can play out genetically.

In this study, we sought evidence among red-sika hybrids that specific genome regions have introgressed more or less than expected under neutrality, in ways that might be interpreted as being due to selection. We used 45K SNP genotypes in 222 Kintyre hybrid deer to estimate genomic clines and show that, as in the other studies cited above, many loci exceed background expectation in terms of direction of introgression $a$ and cline width $b$. We paid particular attention to the X chromosome, as we have previously found relatively more diagnostic and ancestry informative markers on the X than on the autosomes (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020). We then conduct population genetic simulations to investigate admixture scenarios that shed light on the likely roles of drift and selection in generating these results.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample collection

Five hundred thirteen deer samples were collected from 15 forestry sites in the Kintyre region of Scotland between 2006 and 2011. These samples were collected by the Forestry Commission Scotland (now Forestry and Land Scotland) as part of normal deer control measures. Deer were shot as encountered, without regard to the phenotype of the animal (Smith et al., 2018). Sample collection consisted of ear tissue and has been previously described elsewhere (Senn & Pemberton, 2009; Smith et al., 2018). Samples were either preserved in 95% ethanol or frozen for long-term storage.

### 2.2 | DNA extraction and SNP genotyping

SNPs were genotyped on the Cervine Illumina iSelect HD Custom BeadChip using an iScan instrument following manufacturer’s instructions as in Huisman et al., (2016). When this SNP chip was developed, SNPs were selected to be spaced evenly throughout the genome based on the bovine genome with which the deer genome has high homology (Johnston et al., 2017). In the present study we have used the bovine map as this allows use of all of the SNPs, including those that are not polymorphic in red deer, and thus were difficult to map. The majority of the 53K SNPs (45K after quality control; see below) included were selected to be polymorphic in red deer, 4500 SNPs were selected to be diagnostic between either red deer and sika or red deer and wapiti (Cervus canadensis) (Brauning et al., 2015). While one pool of 12 sika from Kintyre were whole genome sequenced during the development of this SNP chip, the main focus was on polymorphic SNPs in red deer on Rum, a well-studied, isolated island population of red deer in the inner Hebrides (Brauning et al., 2015). We used the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer’s instructions to extract DNA for SNP analysis, with the exception that we eluted twice in 25 μl buffer TE to obtain DNA at a sufficiently high concentration. Concentration was assayed using the Qubit dsDNA BR Assay Kit (Invitrogen). Any samples below 50 ng/μl were vacuum-concentrated, re-extracted or omitted from SNP analysis. We used a positive control twice on each 96 well plate to check for consistency between batches (Huisman et al., 2016). We scored genotypes using GenomeStudio using the clusters from Huisman et al., (2016), and clustered SNPs manually if they could not be resolved in these clusters (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020).

All quality control was carried out in PLINK (Purcell et al., 2007). We excluded individual samples with a call rate of <0.90, and deleted loci with a minor allele frequency of <0.001 and/or a call rate of <0.90 (as in McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020). We did not exclude SNPs based on Hardy Weinberg Equilibrium (HWE) as highly differentiated markers between red and sika are not expected to be in HWE.

In McFarlane, Hunter, et al., (2020) and McFarlane, Senn, et al., (2020), we used ADMIXTURE (Alexander et al., 2009) to assign a Q score to each individual. Using the credible intervals (CI), we assessed individuals as pure sika, if the CI overlapped 0, pure red deer if the CI overlapped 1 or hybrid if the CIs overlapped neither 0 or 1 (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020). Of the 513 genotyped deer from Kintyre, 222 were assigned as hybrids, 159 as red deer and 132 as sika. We use these species assignments in the analyses in the present paper.

### 2.3 | Diversity

We estimated genetic divergence between red deer and sika in Kintyre using the hierfstat package in R (Goudet, 2005). We compared only individuals that previous analysis identified as pure
species red deer or sika (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020) and we estimated $F_{ST}$ at each individual locus following Nei (1987). We used a linear model in R (R Core Team, 2013) with $F_{ST}$ as the response variable, and the X chromosome as a reference to ask how the $F_{ST}$ of SNPs on the autosomes differed from those SNPs on the X chromosome.

2.4 Bayesian genomic clines

The genomic clines method is designed to detect loci with alleles that have introgressed at rates that deviate from genome-wide expectations, as those alleles that move faster than expected might be under selection in the novel parental genomic background and those loci that move more slowly than expected might be related to post zygotic reproductive isolation ( Lexer et al., 2007). We used the program bgc ( Gompert & Buerkle, 2012) to estimate Bayesian genomic clines across the hybrid individuals in our population. bgc compares the genotype of each locus in each individual to that individual's hybrid index to estimate values of $\alpha$, which is comparable to a geographic cline centre and $\beta$, comparable to a geographic cline slope ( Gompert & Buerkle, 2012).

Red deer and sika were each assigned to parental populations, and all admixed individuals were put into a “hybrid population”. This is in contrast to some previous analyses in which individuals are separated based on whether they are from a population in which admixture occurs ( Royer et al., 2016; Taylor et al., 2014; Trier et al., 2014). We calculated allele frequencies for the two parental populations using PLINK ( Purcell et al., 2007), while hybrid genotypes were considered individually. We ran bgc on both all 44,997 SNPs and on thinned SNPs ($n = 6684$, thinned according to recombination in PLINK, $r < 0.2$), and found the same frequency of SNPs significant for $\alpha$ and $\beta$ across the genome, so below we report analyses run on all SNPs. We ran bgc five independent times, for 50,000 iterations each time, with a burnin of 25,000 and a thinning interval of 200, and assessed convergence by eye. To be as conservative as possible when determining which loci significantly deviated from the genome-wide expectation, we used the widest possible confidence intervals for each locus from the five chains ( Janousek et al., 2015). Loci with credible intervals that did not overlap with 0 are referred to as “excess” loci. Additionally, we assumed a normal distribution for each $\alpha$ and $\beta$ with the same mean and standard deviation as the empirical data. We then asked which SNPs had $\alpha$ or $\beta$ estimates in the 2.5% upper and lower tails of this distribution. Those loci outside of the 95% distribution are referred to as “outlier loci”.

2.5 SLiM simulations to characterize expectations under drift

We wanted to determine the impact of population size and history on the potential role of drift in hybridized populations. Theoretically, there is an expectation that rare, recent hybridization should result in extremely variable rates of introgression across the genome ( Baird et al., 2003). We used SLiM ( Haller & Messer, 2017) to build some simple models that varied the rate of admixture, the length of time admixture has been occurring and the abundance ratio of each parental type population (1:1 or 3:1). We simulated 1000 individuals with a single chromosome of 1e7 markers, split into two populations of either 500 of each species or 250 and 750 of the two species, and allowed these parental populations to evolve for 3000 generations with a standard rate of neutral mutation (0.01), typically resulting in an $F_{ST}$ between 0.40 and 0.60. Note that we did not simulate any markers to be under selection. We then allowed migration and hybridization between the two populations at a given rate (0.002, 0.02, or 0.2) for a given number of generations (10, 100 or 1000). While 1000 generations is substantially longer than the red deer and sika have been in secondary contact, we designed these simulations to be useful to a wide variety of systems, including those with much more ancient hybridization. We then took the SNP data for all individuals alive at 10, 100 or 1000 generations and put them through our PLINK-ADMIXTURE-bgc pipeline (as above). One deviation from the above pipeline is that due to computational constraints bgc was only run for 2500 iterations, with a burnin of 200 iterations and a sampling interval of two. We ran bgc five times for each simulation, and, as with the empirical analyses, categorized loci based on the widest possible CIs. As bgc analyses may not have converged over such a small numbers of iterations, this could lead to wider CIs than if convergence had occurred in all chains, making this analysis conservative with respect to finding excess loci. We ran each simulation 50 times to determine what proportion of markers deviated significantly from the genome-wide expectation. We did not identify outlier loci for $\alpha$ and $\beta$, as this is less commonly done in the literature, and is difficult to standardize across studies.

3 RESULTS

3.1 Diversity

$F_{ST}$ varied widely among markers (Figure 1a) and across the genome (Figure S1). While each chromosome had SNPs with $F_{ST}$ estimates that ranged from 0 to 1 (mean autosomal $F_{ST} = 0.499 \pm 0.33$), the X chromosome had a higher $F_{ST}$ on average than all other chromosomes with the exception of Chromosome 25 (Figure 1b, Table S1).

3.2 bgc

We found substantial variation between loci in the location and rate of genomic clines between red deer and sika. Positive $\alpha$ can be interpreted as introgression from red deer to sika, while negative $\alpha$ is introgression from sika to red deer. While most of the 44,997 SNPs that we examined were not significant, there were many SNPs that were excess or outliers compared to the genome-wide expectation based on hybrid indices. As noted above, a SNP was considered...
significantly excess if the 95% confidence interval did not overlap zero, and considered an outlier if the point estimate was not within the 95% distribution for the overall genome. A total of 691 (324 negative and 367 positive) SNPs were in excess for $\alpha$ estimates, but not for $\beta$ estimates, 3483 (255 negative and 3228 positive) SNPs had $\beta$ estimates that were in excess but not $\alpha$ estimates and 4437 other SNPs (60 negative $\alpha$ and $\beta$, 0 negative $\alpha$ and positive $\beta$, 3034 positive $\alpha$ and negative $\beta$, 1343 positive $\alpha$ and $\beta$) were in excess for both $\alpha$ and $\beta$ (Table 1). 1168 SNPs were $\alpha$ outliers but not $\beta$ outliers (1 negative, 1167 positive), 678 SNPs (568 negative, 110 positive) were outliers for $\beta$ but not $\alpha$ and 2450 were outliers for both $\alpha$ and $\beta$ (0 negative $\alpha$ and $\beta$, 0 negative $\alpha$ and positive $\beta$, 2438 positive $\alpha$ and negative $\beta$, 12 positive $\alpha$ and $\beta$). We found substantially more excess loci with positive $\alpha$ estimates than negative $\alpha$ estimates (4744 vs. 384) and substantially more positive $\alpha$ outliers than negative outliers (3617 vs. 1). We found more positive than negative $\beta$ excess SNPs (4571 vs. 3349), but substantially fewer positive than negative $\beta$ outlier SNPs (122 vs. 3006). Excess SNPs (for either $\alpha$ or $\beta$) are spread across the entire genome, and occur on every chromosome (Figure 2a,b), as are outlier SNPs.

When we examined only those diagnostic and ancestry informative markers we have previously identified ($n = 3793$; McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020), we found 226 (five negative and 221 positive) that were significantly $\alpha$ excess but not $\beta$ excess, 87 (14 negative and 73 positive) that were significantly $\beta$ excess but not $\alpha$ excess, and 2315 (2 negative $\alpha$ and $\beta$, 0 negative $\alpha$ and positive $\beta$, 2285 positive $\alpha$ and negative $\beta$, 28 positive $\alpha$ and $\beta$) that were both $\alpha$ and $\beta$ excess. Of the AIMs, we found 346 (0 negative and 346 positive) that were $\alpha$ but not $\beta$ outliers, 313 (309 negative and four positive) that were $\beta$ but not $\alpha$ outliers and 1870 SNPs (0 negative $\alpha$ and $\beta$, 0 negative $\alpha$ and positive $\beta$, 1870 positive $\alpha$ and negative $\beta$, 0 positive $\alpha$ and $\beta$) that were significant outliers for $\alpha$ and $\beta$. We found substantially more excess loci with positive $\alpha$ estimates than negative $\alpha$ estimates (4744 vs. 384) and substantially more positive $\alpha$ outliers than negative outliers (3617 vs. 1). We found more positive than negative $\beta$ excess SNPs (4571 vs. 3349), but substantially fewer positive than negative $\beta$ outlier SNPs (122 vs. 3006). Excess SNPs (for either $\alpha$ or $\beta$) are spread across the entire genome, and occur on every chromosome (Figure 2a,b), as are outlier SNPs.

![Figure 1](image1.png)

**Figure 1** (a) Frequency of SNPs within 0.05 $F_{ST}$ bins, estimated using pure sika and red deer (see text). (b) Boxplot showing $F_{ST}$ between red deer and sika on each (bovine) chromosome. Each box shows the median, 25th and 75th percentile for each chromosome and each whisker extends to the fifth and 95th percentiles of $F_{ST}$ for each chromosome.

**Table 1**

<table>
<thead>
<tr>
<th>$\alpha$ category</th>
<th>$\beta$ category</th>
<th>Introggression interpretation</th>
<th>45K SNPs Excess CI ≠0</th>
<th>95% outlier</th>
<th>AIM Excess CI ≠0</th>
<th>95% outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Fast into red deer</td>
<td>60 (0.001)</td>
<td>0 (0.000)</td>
<td>2 (0.001)</td>
<td>0 (0.000)</td>
</tr>
<tr>
<td>Negative</td>
<td>Not significant</td>
<td>Into red deer</td>
<td>324 (0.007)</td>
<td>1 (0.000)</td>
<td>5 (0.001)</td>
<td>0 (0.000)</td>
</tr>
<tr>
<td>Negative</td>
<td>Positive</td>
<td>Slow into red deer</td>
<td>0 (0.000)</td>
<td>0 (0.000)</td>
<td>0 (0.000)</td>
<td>0 (0.000)</td>
</tr>
<tr>
<td>Not significant</td>
<td>Negative</td>
<td>Fast in both directions</td>
<td>255 (0.006)</td>
<td>568 (0.013)</td>
<td>14 (0.004)</td>
<td>309 (0.081)</td>
</tr>
<tr>
<td>Not significant</td>
<td>Not significant</td>
<td>No movement</td>
<td>36,386 (0.809)</td>
<td>40,701 (0.905)</td>
<td>1165 (0.307)</td>
<td>1309 (0.341)</td>
</tr>
<tr>
<td>Not significant</td>
<td>Positive</td>
<td>Slow in both directions</td>
<td>3228 (0.072)</td>
<td>110 (0.002)</td>
<td>73 (0.019)</td>
<td>4 (0.001)</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Fast into sika</td>
<td>3034 (0.067)</td>
<td>2438 (0.054)</td>
<td>2285 (0.602)</td>
<td>1870 (0.487)</td>
</tr>
<tr>
<td>Positive</td>
<td>Not significant</td>
<td>Into sika</td>
<td>367 (0.008)</td>
<td>1167 (0.026)</td>
<td>221 (0.058)</td>
<td>346 (0.090)</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Slow into sika</td>
<td>1343 (0.030)</td>
<td>12 (0.000)</td>
<td>28 (0.007)</td>
<td>0 (0.000)</td>
</tr>
</tbody>
</table>
\( \beta \) (Table 1). As was the case when we used all the SNPs, we found many more excess loci with positive \( \alpha \) than negative \( \alpha \) (2534 vs. 7) and many more positive than negative \( \alpha \) outlier AIM SNPs (2234 vs. 0), suggesting more extreme introgression from red deer into sika than from sika into red deer. We found fewer positive than negative excess \( \beta \) AIM SNPs (101 vs. 2301), and fewer positive than negative outlier \( \beta \) AIM SNPs (4 vs. 2179). Similarly to when we examined all SNPs, excess and outlier \( \alpha \) and \( \beta \) SNPs were found across the genome. In contrast to when we examined all SNPs, there was a substantially higher proportion of AIM SNPs that were different than the genome wide expectation (69.3% DM&AM significant excess vs. 19.1% from all SNPs and 65.5% AIM significant outlier vs. 9.5% from all SNPs).

3.3 | SLiM simulations

Across the scenarios that we simulated, we found that the majority of simulated loci were not significant for either \( \alpha \) or \( \beta \) estimates. However, we did find that in cases where there had only been 10 generations of admixture, and a low level of hybridization, most loci had either a positive or negative \( \beta \) estimate, suggesting faster or slower than expected movement through the cline (Figure 3, panels “sle”, “slo” and “sme”). While the proportion of loci with significant \( \beta \) decreased with increasing number of generations and increased admixture, there are loci with significant \( \beta \) found in every other simulated scenario, with sometimes as many as 40% of loci introgressing at extreme rates when compared to the average rate of introgression.
across the entire genome. Additionally, in scenarios where hybridization has been progressing for longer (Figure 3, m and l rows), as many as 15% of loci have negative alpha estimates. This appears to be more extreme with increased rates of hybridization.

4 | DISCUSSION

Using 44,997 SNPs, we found extremely variable $F_{ST}$ between red deer and sika across all chromosomes, although the X chromosome had a substantially higher $F_{ST}$ than the autosomes. We also found 5128 $\alpha$ excess SNPs, of which 3618 were outliers and 3618 $\beta$ excess SNPs of which 3128 were outliers (Table 1). When we compared these excess and outlier SNPs to our list of AIMs, we found a high proportion of AIM loci were excess and/or outliers (Table 1). This suggests that some caution should be used when interpreting the results of genomic clines of diagnostic or ancestry informative markers, as there could be a relationship between informativeness and extreme clines of these markers.

The higher $F_{ST}$ on the X chromosome is not unexpected. Fixed differences (i.e., diagnostic markers) between the two species are concentrated on the X chromosome (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020). More generally, limited recombination on the sex chromosomes and a smaller effective population size ($N_e$) result in a prediction of limited introgression on the X (Baack & Rieseberg, 2007; Barton, 1979). Genes that are associated with postzygotic isolation have been recorded on the X in other systems (Qvarnström & Bailey, 2009). As we discuss below, those markers with extremely diverged $F_{ST}$ and slow rates of introgression, particularly those on the X chromosome, are good candidates for further examination for a role in reproductive isolation.

Our empirical findings of many SNPs with significant excess $\alpha$ estimates is in strong contrast to the few excess $\alpha$ loci in the simulations with recent hybridization (10 or 100 generations), although we are cautious not to interpret these excess alleles as under selection. Specifically, we found 4474 SNPs with positive excess $\alpha$ (and 3617 outliers), and 384 SNPs with negative excess $\alpha$ (and 1 outlier), which indicates alleles that have moved from red deer to sika or sika to red deer more than expected based on the genomic expectation. Previous simulations using bgc have found substantial variation in $\alpha$ estimates when smaller sample sizes were simulated, even if the simulation was for only 25 generations with an admixture rate of 0.2 (Gompert & Buerkle, 2011). Our empirical data set contains only 222 hybrid individuals, which is a small population compared to most of our simulations. It should be noted that the hybrid population size in our simulations varied (between approximately 45 and

FIGURE 3. We used SLiM (Haller & Messer, 2017) to simulate admixing populations that had been in secondary contact for either a short (s, 10 generations, top row), medium (m, 100 generations, middle row), or long (l, 1000 generations, bottom row) length of time since admixture started. For each length of secondary contact, we also simulated rates of migration and interbreeding between populations, as either low (l, 0.002, left two columns), medium (m, 0.02, middle two columns), or high (h, 0.2, right two columns), and the abundance ratio of each pure population, as either even (e, 1:1) or odd (o, 1:3). Each simulation was run 50 times, no selection was simulated, and we categorized (into nine categories; legend) the direction and rate of introgression among simulated hybrid individuals using bgc. Overall, introgression at most loci did not deviate from genome-wide expectation, but especially in cases with a short time since admixture started and a low rate of admixture (top, left two panels), many loci introgressed faster than genome-wide expectation despite the total absence of any selection in the simulations.
approximately 800), as it was a function of the admixture rate, and the stochasticity built into these individual-based simulations. In any case, the 222 deer hybrids from Kintyre are substantially fewer than the 500 or 1000 hybrid individuals that were simulated in the best performing models by Gompert and Buerkle (2011). This is good reason to be cautious about interpreting excess or outlier $\alpha$ estimates as evidence for selection on these loci.

We found substantially more loci with positive than negative excess and outlier $\alpha$'s, indicating that more alleles have shifted from red deer into sika than from sika into red deer. There are three possible explanations for this, which are difficult to distinguish in our study system. First, there could be asymmetry in backcrossing, such that there is more backcrossing into sika than there is into red deer. This was previously indicated in an analysis of microsatellite data by Goodman et al. (1999) who estimated that the rate of backcrossing into sika was twice the rate of backcrossing into red deer ($H = 0.002$ vs. $H = 0.001$), although based on mitochondrial DNA, it is clear that backcrossing does proceed in both directions (Smith et al., 2018). Second, the pattern of increased positive vs. negative $\alpha$ estimates could be due to marker selection. The SNP chip we used was mainly designed to provide polymorphic loci for studies within red deer, and has just 2250 SNPs that were selected to be diagnostic between red deer and sika (Brauning et al., 2015), although ultimately only 629 SNPs are diagnostic in our study population (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020). Across the other loci on the SNP chip, the sika population is less diverse than the red deer probably due to a demographic history of bottlenecks and the fact the chip was primarily designed for use in red deer. These two features together make it difficult to document whether shared alleles are introgressing from sika into red deer, whereas it is easier to document the introgression of private alleles from a large, outbred, polymorphic population of red deer into sika. Further, it is difficult to quantify the relative contribution of each of these processes. The third possible mechanism explaining the seemingly higher proportion of red deer alleles introgressing into sika than in the other direction is that, as sika are an introduced species in the UK, it is possible that some alleles that are introgressing from red deer to sika are indeed the result of adaptive introgression, because they increase the fitness of hybrids. Adaptive introgression can involve a faster response to selection in a new environment than selection on a new mutation since the allele is already proven, albeit in a different background (Hedrick, 2013), and has been suggested to be a potentially positive conservation outcome of anthropogenic hybridization (Hamilton & Miller, 2016). Without fitness estimates, it is extremely difficult to demonstrate adaptive introgression in wild populations (Taylor & Larson, 2019), making it difficult to tease apart these three possibilities.

Empirically, we found 3349 (~6.7%) SNPs with a negative, excess $\beta$ estimate (3006 negative $\beta$ outliers), suggesting that these SNPs were introgressing faster than expected between red deer and sika. While red deer and sika have been hybridizing in Scotland for at least 6–7 generations, it is possible they may have hybridized prior to introduction to Scotland, as hybridization was reported in the Irish source population before animals were introduced to Kintyre (Powerscourt, 1884). Either way, this is a case of recent hybridization. The rate of backcrossing has previously been estimated using 11 microsatellite markers as 0.002 into sika and 0.001 into red deer (Goodman et al., 1999), which is consistent with our simulated “low” admixture parameter. The ratio of red deer to sika is variable across Kintyre (Smith et al., 2018). Thus, our empirical work is most consistent with the “sle” or the “slo” simulations, where we found that most SNPs were excess $\beta$, either positive or negative (Figure 3). Thus, we found substantially fewer significant negative $\beta$ SNPs than we may have expected from the simulations, highlighting that these simulations are just a toy example, rather than a highly accurate simulation of this natural system. For comparison, many studies of hybridization that have used bgc have not found significant $\beta$ estimates. For example, a recent study of ibis (Plegadis falcinellus, P. chihi, P. ridgwayi) hybridization using diagnostic markers found no significant negative $\beta$ SNPs, in spite of the ibis hybrid zone probably only being 60 or so years old (Oswald et al., 2019). In contrast, a study of recent sole ( Solea aegyptiaca x S. senegalensis) hybridization found 52% of all loci exhibited an extreme $\beta$ value, with 26% of all loci exhibiting a negative $\beta$ estimate (Souissi et al., 2018). For an example of research on an older hybrid zone, black-tailed deer (Odocoileus hemionus columbianus) and mule deer (O. h. hemionus) have been hybridizing for approximately 8000 years, and when genomic clines were estimated using 95 SNPs, four were found to have extreme $\beta$ estimates (two positive and two negative; Haines et al., 2019). Overall, comparison of genomic cline estimates across studies and taxa is difficult, particularly given the expectation for extreme $\beta$ values due to drift (Baird et al., 2003), the potential for extremely different results depending on the marker panel used (Table 1), the age of a hybrid zone, and rate of admixture between species (Figure 3). As such, a more comprehensive meta-analysis approach is probably needed to understand factors driving genomic cline variation across taxa.

Although we cannot be sure that any loci demonstrate selection in our study system we found a number of SNPs that exhibited extreme introgression as judged by $\alpha$ or $\beta$ estimates. For example, there are 298 SNPs with $F_{ST} = 1$ and a significantly negative $\beta$, suggesting that they are highly diverged between the two species, and are introgressing more quickly than would be expected in the hybrid population. This is what we would expect if there was adaptive introgression. We did not find any SNPs with $F_{ST} = 1$ and significantly positive $\beta$, as we might have expected to detect if there were loci with large effects on reproductive isolation. Additionally, there are no obvious regions where there is strong divergence and extreme alpha or beta among linked SNPs (Figures S2 and S3), suggesting little evidence for strong selection or “islands of differentiation” (Cruickshank & Hahn, 2014; Turner & Hahn, 2010; Wolf & Ellegren, 2017). Simulations of genomic clines that included epistatic interactions on reproductive isolation (i.e., Bateson-Dobzhansky-Muller interactions; Dobzhansky, 1937; Muller, 1940) are difficult to detect using bgc (Gompert & Buerkle, 2011), so we would not claim the lack of evidence in this case as evidence of the absence of genes involved in reproductive
isolation in this system. Substantially more work is needed to address this question.

There is an expectation that when there is recent, rare hybridization, the genomic outcome of introgression is extremely stochastic (Baird et al., 2003), and it has previously been noted how difficult it is to derive a null distribution for locus-specific introgression (Gompert & Buerkle, 2011). Drift can substantially increase or decrease the frequency of different blocks, in the complete absence of selection. This is consistent with what we saw in our SLiM simulations, where, when we simulated 10 generations of admixture with a rate of admixture of 0.002, we found in some cases that 50% of markers had wider clines and 50% of markers had narrower clines than predicted from the genome-wide expectation (Figure 3). As noted above, the hybrid population sizes also varied with admixture rate, particularly when hybridization was rare and had only been ongoing for 10 generations (scenarios sle and slo). This is consistent with untargeted sampling in wild populations, as, if hybridization is recent and rare, there will be proportionately fewer hybrids in the population. This confirms that extreme $\alpha$ estimates should not be taken as evidence of selection (Gompert & Buerkle, 2012), or of adaptive introgression (Taylor & Larson, 2019), as this introgression happens in the absence of selection. This is particularly true when hybridization is recent and rare, leading to relatively few hybrids in the population. Previous neutral simulations of 25 generations of admixture with an admixture rate of 0.2, comparable to our she and sho simulations but with a simulated population size of 100, found substantial variation in the estimated $\alpha$ or $\beta$ estimates, with $\alpha$ being more variable than $\beta$ (Gompert & Buerkle, 2011). These simulations found that $\alpha$ or $\beta$ were less variable when the population sizes simulated were 500 or 1000, although some outlier $\alpha$ or $\beta$ loci were still found in some simulations in these cases (Gompert & Buerkle, 2011). As this pattern was less extreme when hybridization had been progressing for many generations (i.e., 100 or 1000), this provides an additional rationale for researchers to quantify the length of time admixture has been occurring in their system prior to drawing conclusions (Loh et al., 2013; McFarlane & Pemberton, 2019). The strength of evidence for adaptive introgression from genomic clines is, therefore, weak in more recently admixed systems, including many examples of anthropogenic hybridization. To make the case that adaptive introgression is occurring, particularly in a recent case of anthropogenic hybridization, studies must incorporate independent fitness estimates to demonstrate selection.

To conserve a species in the presence of hybridization, we must first quantify both the number of individuals in the population that are hybrids, and the proportion of alleles that could be replaced by introduced alleles, i.e. in line with the gene-based theory of conservation (Petit, 2004). In our study area, we found approximately 43% of individuals are hybrids (McFarlane, Hunter, et al., 2020; McFarlane, Senn, et al., 2020) and in the present study, we have identified 60 SNPs with both an excessive negative $\alpha$ and excessive negative $\beta$ estimate, indicative of introgressive alleles moving from the introduced sika into the native red deer faster than expected. These SNPs are spread across 26 different chromosomes. Whether the pattern of these SNPs is the result of selection or drift, it is still the case that there are sika alleles that are spreading into red deer populations via hybridization faster than those at other loci. These are the genome regions that are of potential conservation concern for Scottish red deer as these alleles may most quickly replace their red deer alternates. Techniques such as admixture mapping could be used to try to link SNPs to phenotypes known to be under selection (Buerkle & Lexer, 2008), and then cross check these SNPs against those introgressing fastest. Such gene-targeted conservatism is unlikely to be successful (Kardos & Shafer, 2018), particularly since many of the traits of interest in red deer (e.g., redness, antler size and shape, body size) are likely to be polygenic (Santure & Garant, 2018). Specifically, body size has been found to be polygenic in a variety of taxa, including Soay sheep (Ovis aries; Bérénos et al., 2015), bighorn sheep (Ovis canadensis; Miller et al., 2018), and polar bears (Ursus maritimus; Malenfant et al., 2018). Antler shape has been found to be polygenic in Scottish red deer (Peters et al., in prep). Altogether, it seems unlikely that the 60 SNPs we have identified here would have large impacts on the phenotypic traits of interest that policy makers would seek to conserve in Scottish red deer.

Genomic clines can be used to identify loci showing extreme introgression. However, genomic clines cannot be used to identify definitively alleles under selection (Gompert & Buerkle, 2011, 2012), so different methods must be employed to distinguish between alleles undergoing adaptive introgression or involved in reproductive isolation and those loci that deviate from genomic expectations due to stochastic processes. One approach would be to study replicate hybrid zones, on the assumption that stochastic processes will act independently in each instance of secondary contact, but selection will not. Loci which have consistent excess $\beta$ estimates would be the best candidates for being under selection, either for or against introgression into a novel background. In house mice, it was found that 28/41 SNPs had different genomic clines between two replicates, as assessed using a likelihood ratio test that compared the clines, encompassing both $\alpha$ and $\beta$, suggesting that few if any of the extreme SNPs could be related to genetic incompatibilities or adaptive introgression (Teeter et al., 2010). While it should be noted that detecting signals of even very strong selection at the genome wide level is extremely difficult, requires substantial power and a strong signal (Castro et al., 2019), those SNPs with extreme $\beta$ across multiple replicate hybrid zones would be strong candidates for being involved in either adaptive introgression, or impeding gene flow between species. Future research on red deer x sika hybridization could capitalize on replicate hybrid populations across Europe (e.g., Ireland: Smith et al., 2014; Lithuania: Ražanskė et al., 2017; and Poland: Biedrzycka et al., 2012) where the many points of sika introduction have generated natural replications of this.

Genomic clines allow for the estimation of local introgression, allowing us to identify particular regions of the genome that could be involved in adaptive introgression or, conversely, reproductive isolation between species at secondary contact. The difficulty
in differentiating the patterns caused by selection versus those caused by drift, as we have simulated here, should not be taken as a deterrent to using such methods. Instead, we believe that there is an opportunity to make use of replicated hybrid zones, including those with varying degrees of hybridization (Mandeville et al., 2019) to compare genomic clines, as those loci with consistent rates of introgression in independent populations are more likely to be under selection. Further, future research could employ meta-analysis techniques across studies and species to quantify rates of introgression across the genome, while controlling for phylogeny, to make generalizations about selection across the genome at secondary contact.

ACKNOWLEDGEMENTS

We thank the Forestry and Land Scotland rangers, especially Fraser Robinson and Kevin McKillop for collecting samples, the Wellcome Trust Clinical Research Facility Genetics Core, Edinburgh for performing the genotyping and Paul Fisher and Rudi Brauning for SNP array development. We are also grateful to Nick Barton and Stuart Baird for discussions about the null expectation of genomic clines, as well as Alan Alexander, Zachary Gompert, Elizabeth Mandeville and Piotr Zielinski for assistance with and discussion of bgc, and for the comments from three constructive reviewers. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). This project was funded by a European Research Council Advanced Grant to JMP, a Vetenskapsrådet (Swedish Research Council) International Postdoc Fellowship to SEM and Natural Environment Research Council PhD Studentships to HVS and SLS.

AUTHOR CONTRIBUTIONS

S. Eryn McFarlane and Josephine M. Pemberton conceptualized the project. Helen V. Senn and Stephanie L. Smith collected the data. S. Eryn McFarlane analysed the data and wrote the manuscript and all authors contributed to discussions, revised and approved the final version.

DATA AVAILABILITY STATEMENT

All data and code are available at https://figshare.com/projects/Locus-specific_introgression_in_young_hybrid_swarms_drift_dominates_selection/76473

ORCID

S. Eryn McFarlane https://orcid.org/0000-0002-0706-458X

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** McFarlane SE, Senn HV, Smith SL, Pemberton JM. Locus-specific introgression in young hybrid swarms: Drift may dominate selection. *Mol Ecol.* 2021:00:1–12. [https://doi.org/10.1111/mec.15862](https://doi.org/10.1111/mec.15862)