Title: Increased genetic marker density reveals high levels of admixture between red deer and introduced sika in Kintyre, Scotland

Authors: S. Eryn McFarlane*[1,2], Darren C. Hunter[1,3], Helen V. Senn[1,4], Stephanie L. Smith[1,5], Rebecca Holland[1], Jisca Huisman[1], Josephine M. Pemberton[1]

Author affiliations:

1. Institute of Evolutionary Biology, School of Biological Science, University of Edinburgh, Edinburgh, EH9 3FL, UK
2. Department of Biology, Lund University, Sölvegatan 37, 223 62 Lund, Sweden
3. Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, S10 2TN UK
4. WildGenes Laboratory, Royal Zoological Society of Scotland, Edinburgh, UK
5. The Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian, Edinburgh, EH25 9RG, UK

***Author for correspondence:** eryn.mcfarlane@ed.ac.uk

Abstract:

Hybridization is a natural process at species range boundaries, but increasing numbers of species are experiencing it due to human intervention. In such cases of anthropogenic hybridization, subsequent introgression can threaten the survival of native species. To date many such systems have been studied with too few genetic markers to assess the level of threat resulting from advanced backcrossing. Here we use 44,999 single nucleotide polymorphisms and the ADMIXTURE program to study two areas of Scotland where a panel of 22 diagnostic microsatellites previously identified introgression between native red deer (*Cervus elaphus*) and introduced Japanese sika (*Cervus nippon*). In Kintyre we reclassify 26% deer from the pure species categories to the hybrid category whereas in the NW Highlands we only reclassify 2%. As expected, the reclassified individuals are mostly advanced backcrosses. We also investigate the ability of marker panels selected on different posterior allele frequency criteria to find hybrids assigned by the full marker set, and show that in our data, ancestry informative markers (i.e. those that are highly differentiated between the species, but not fixed) are better than diagnostic markers (those markers that are fixed between the species)

because they are more evenly distributed in the genome. Diagnostic loci are concentrated on the X chromosome to the detriment of autosomal coverage.

Introduction:

Hybridization is of rising concern to conservation biologists (Rhymer & Simberloff 1996; Allendorf et al. 2001; Grabenstein & Taylor 2018). While hybridization regularly occurs naturally (Mallet 2005), anthropogenic hybridization resulting from human actions is expected to increase in frequency and intensity in the future, as introductions and range shifts due to climate change increase the occurrence of secondary contact among previously allopatric species (Crispo et al. 2011). A recent review of anthropogenic hybridization found that most cases of admixture result in the breakdown of reproductive isolation between previously distinct species (Grabenstein & Taylor 2018). Biologists have flagged anthropogenic hybridization as a conservation concern, and the rapidity with which hybrid swarms are formed (Grabenstein & Taylor 2018) suggests that this is a problem that is both common and urgent.

Backcrossing of hybrid individuals with parental species, leads to many backcrossed individuals that are difficult to differentiate from parental species individuals (McFarlane & Pemberton 2019). When backcrossed individuals mate with parental species individuals (instead of other hybrid individuals), bimodal hybrid zones can be formed, where there are many individuals with a small proportion of introgression from another species. This leads to a large proportion of individuals in the population of hybrid ancestry that are difficult to distinguish either genetically or phenotypically from parental species. In the case of genetic analysis, this is because with each generation of backcrossing a doubling of markers is needed to detect introgression with the same degree of confidence (Boecklen & Howard 1997). However, many current studies of anthropogenic hybridization use an average of 20 markers to detect introgression (Todesco et al. 2016), which is too few to reliably detect second-generation backcrosses in most systems (Boecklen & Howard 1997; Vähä & Primmer 2006). The limited detection power of the marker panels that are typically used in studies of anthropogenic hybridization are insufficient to detect on-going introgression and it seems likely that many thousands of markers will be necessary to accurately identify the majority of hybrid individuals (McFarlane & Pemberton 2019).

The use of genomic markers (i.e. thousands of SNPs as compared to tens of microsatellite markers) increases the detection probability of admixed individuals, particularly in systems where backcrossing has occurred for multiple generations. Theoretically with each generation

of backcrossing, the proportion of individuals that are homozygous at diagnostic markers for alleles from one parental species increases, with the power to detect backcrosses sharply decreasing with each additional generation (Boecklen & Howard 1997). For these reasons, it has been suggested that researchers consider not only divergence between hybridizing species (as suggested by (Vähä & Primmer 2006)), but also the number of elapsed generations since hybridization has begun and the recombination rate of the taxa studied when determining how many markers to genotype (McFarlane & Pemberton 2019). Empirically, use of high-density markers has indeed led to an increase in detection of backcrossed individuals. For example in a wolf-dog hybrid zone, 1 – 5% of wolves were found to be hybrid individuals when 16 – 18 microsatellite markers were used (Randi 2008), whereas 62% of sampled individuals were determined to be hybrid when 170 000 SNP markers were used (Pilot et al. 2018). Similarly, all individuals in a westslope cutthroat trout – rainbow trout hybrid zone that were previously thought to be parental species were discovered to be hybrid individuals when re-analyzed with more markers (Boyer et al. 2008; Hohenlohe et al. 2013). Use of higher density marker panels has the potential to change the qualitative and quantitative understanding of hybrid systems, including providing information on the direction of backcrossing, the number of generations since backcrossing began and the distribution (e.g. hybrid swarm, unimodal or bimodal) of the hybrid system (McFarlane & Pemberton 2019).

Even with the use of genomic markers, not all markers will contribute equally to hybrid detection. Diagnostic markers (DM), those that have completely fixed allele differences between populations, have more power than markers that are polymorphic in either population (Pritchard et al. 2000; Vähä & Primmer 2006). Ancestry informative markers (AIM), which have substantial allele frequency differences between parental populations are also more powerful than markers which are polymorphic within one or both parental populations (Shriver et al. 2003). Thus, if a situation requires the use of limited markers fewer markers (e.g. to lower the cost of genotyping individuals, or because of the need for rapid testing (Senn et al. 2018)), it would be best to focus on diagnostic or ancestry informative markers, with the caveat that more markers will always lead to improved detection of hybrid individuals (McFarlane & Pemberton 2019). However, it should be noted that both diagnostic and ancestry informative markers can have a biased distribution across the genome. If there is heterogeneity in introgression across the genome, then a limited number of markers may miss hot-spots of introgression. For example, genes that are associated with post zygotic

reproductive isolation are expected to be concentrated on sex chromosomes, due to the smaller effective population size and limited recombination of sex chromosomes (Qvarnström & Bailey 2009). Empirically, a disproportionate number of genes associated with reproductive isolation have been traced to the sex chromosomes (Coyne & Orr 2004), including in *Mus. musculus – M. domesticus* hybrids (Payseur et al. 2004), *Xiphophorus sp.* hybrids (Schartl 1988), and *Ficedula* hybrids (Sæther et al. 2007). Markers in strong LD with such genes are likely to be diagnostic (DM). AIM may also be disproportionately distributed across the genome, but perhaps not with the same intensity as markers that have fixed between species.

Hybridization between native red deer (*Cervus elaphus*) and Japanese sika (*C. nippon*) in Scotland is a scenario in which the use of high-density markers to explore the extent of backcrossing seems very likely to provide new insights. Sika were introduced from Japan to Ireland in 1860 and to Scotland from 1870 (Ratcliffe 1987). Based on earlier work using microsatellite markers, there appear to be two different scenarios of admixture between red deer and sika in Scotland. Sika were introduced to the Kintyre Peninsula in 1893, and phenotypically hybrid individuals were first reported in the region in the 1974 (Ratcliffe 1987). Given a generation time of approximately 5 years in red deer (Coulson et al. 1998), we would expect at least 6 -7 generations of backcrossing in this system, as our dataset includes deer sampled as early as 2006. However, we should not discount the possibility of hybridization prior to introduction to Scotland, as hybridization was recorded in the Irish source population in the 1880's (Powerscourt 1884). Previous research in Scotland used 22 diagnostic microsatellites to determine that there is a range of admixture proportions in Kintyre, where backcrosses are common (Senn et al. 2010a; Smith et al. 2018). The use of a diagnostic mitochondrial marker demonstrated the existence of backcrosses with very low levels of nuclear introgression, beyond the detection of the microsatellite markers (Smith et al. 2018). Hybrid individuals have phenotypes that are closely correlated with their admixture score (Senn et al. 2010b). While intermediate hybrids may be detectable for phenotypes, advanced backcrosses are easily confounded with the parental species, and are difficult for stalkers to identify in the field (Senn & Pemberton 2009; Smith et al. 2014). Sika were introduced to the north west highland region between 1889 and 1900 (Ratcliffe 1987). While hybrid individuals have been documented in this region, in contrast to Kintyre, there does not seem to be a hybrid swarm (Smith et al. 2018).

Anthropogenic hybridization between red deer and Japanese sika provides an opportunity to use SNP markers to search for cryptic backcrossed individuals that were previously identified as parental individuals. Using substantially more markers than were previously available, this study has four aims: (1), to determine how many individuals are hybrids when scored using greatly increased marker density; (2) to determine the distribution of hybridization and introgression in each study area, specifically whether the hybrid zone is uni or bi directional and whether hybrids mate with other hybrids or with parental species individuals; (3) to study the performance of different classes of marker (diagnostic, ancestry informative and random) in determining hybrid status and understand the difference arising from marker number and genomic location.

Methods:

Study area and sampling:

Samples were collected in the Kintyre peninsula, and from the northwest highland region of Scotland (Figure 1). Samples were collected from 15 forestry sites by Forestry Commission Scotland rangers as part of routine culling operations. The Kintyre samples included 336 individuals shot in 2006-2007 and 187 shot in 2010-2011. The NW highland samples included 108 individuals sampled in 2010- 2011. All samples were shot as encountered, i.e. without bias in terms of phenotype (Smith et al. 2018). Sample collection consisted of ear tissue and has previously been described in Senn and Pemberton (2009) and (Smith et al. 2018). Samples were either preserved in 95% alcohol or frozen.

Microsatellite genotyping, mtDNA haplotyping

All individuals included in this study were previously genotyped at 22 diagnostic microsatellites (Senn et al. 2010a; Smith et al. 2018). Most individuals were also typed at the mitochondrial DNA control region, which includes a 39-bp sequence with a variable number of tandem repeats, of which red deer have a single repeat and Japanese sika have three repeats (Cook et al. 1999; Smith et al. 2018).

DNA extraction for SNP analysis

We used the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacture's instructions to re-extract DNA for SNP analysis, except that to obtain DNA at a sufficiently high concentration, in the last step we eluted twice in 25 ul TE. DNA concentration was assayed

using the Qubit™ dsDNA BR Assay Kit (Invitrogen™) and any samples below 50ng/ul were vacuum-concentrated, re-extracted or omitted from SNP analysis.

SNP Genotyping

SNP genotyping was carried out using the cervine Illumina iSelect HD Custom BeadChip using an iScan instrument following manufacturer's instructions (Huisman et al. 2016). To develop this chip five European red deer, and two wapiti (*Cervus canadensis* or *Cervus elaphus canadensis*) were whole genome sequenced (Brauning et al. 2015), in addition to two pools of 12 pure species red deer from the Isle of Rum, Scotland and 1 pool of 12 pure species sika (based on microsatellite data) from Kintyre, at 10x coverage. The majority of SNPs on the chip (53K attempted loci) were selected to be polymorphic within red deer, specifically within the Rum red deer. However, during the development of the chip 2250 SNPs were included that were expected to be diagnostic between red deer and sika, and 2250 SNPs were included that were expected to be diagnostic between red deer and wapiti. Within each marker set, SNPs were selected to be evenly spaced throughout the genome according to map positions in the bovine genome, with which the deer genome has high homology (Johnston et al 2017)(Slate et al. 2002). To check for consistency between batches and calculate the repeatability of each SNP, we used the same positive control twice on each 96-well plate (Huisman et al. 2016). Genotypes were scored using GenomeStudio (Illumina), using the clusters from (Huisman et al. 2016). SNPs that didn't successfully cluster according to Huisman (et al. 2016) were clustered manually. Quality control was carried out in PLINK (Purcell et al. 2007). We excluded individual samples with a call rate of less than 0.90. We deleted loci with minor allele frequency less than 0.001, and/or call rate of less than 0.90. We did not exclude SNPs based on the Hardy-Weinberg equilibrium, as any diagnostic markers fixed between the two species are not expected to be in HWE in an admixed population.

Estimating global F_{ST}

To estimate global F_{st} between red deer and sika in our two study areas, we considered only individuals that had been assessed as parental species from the phenotype by the ranger concerned. We used the R package 'assigner' (Gosselin et al. 2016) to estimate Weir and Cockerham's F_{st} (Weir & Cockerham 1984), with 95% confidence intervals (CI) based on bootstrapping.

Estimating admixture

For the microsatellite data, we used the STRUCTURE Q scores from Smith et al. 2018 to compare species assignment with the Q scores estimated using SNP markers (described below). However, unlike Smith et al. (2018), we used credibility intervals rather than a threshold cut off to assign individuals to species categories (see description below).

For the SNP data, we used ADMIXTURE (Alexander & Lange 2011) to estimate the admixture (Q) score for each individual. ADMIXTURE uses maximum likelihood to estimate individual ancestry proportions (Alexander et al. 2009). We ran ADMIXTURE unsupervised separately on the deer from Kintyre and those from the NW highlands, as these are two separate populations and we did not want to inadvertently introduce population structure while examining hybridization (Gompert & Buerkle 2016). We assumed two ancestral populations. We used the bootstrapping function in ADMIXTURE to estimate standard errors, from which we calculated 95% confidence intervals as 1.96*SE. We also used fastSTRUCTURE (Raj et al. 2014) to estimate Q scores which were highly correlated with the Q scores from ADMIXTURE (Pearson's correlation coefficient R= 0.9999, t=6720.1, df=631, p<2.2e-16), but used ADMIXTURE Q scores in all the analyses reported below in order to report error around point estimates (McFarlane & Pemberton 2019). It should be noted that neither analysis accounts for linkage between markers.

To categorize animals into pure red deer, pure sika deer or hybrid, we determined whether the 95% confidence intervals for the Q scores overlapped 0.99999 for pure red deer, or 0.00001 for pure sika deer. These are the highest and lowest possible Q scores produced by ADMIXTURE, and, with many markers, confidence intervals were extremely narrow. All other individuals were classified as 'hybrids'. These stringent criteria for assigning an individual to a parental species means that we are more likely to wrongly assign parental species individuals as hybrids than wrongly assign hybrid individuals as parental (i.e. low accuracy vs. low efficiency as defined by (Vähä & Primmer 2006).

Effect of varying SNP selection

We wanted to compare the efficiency of using our entire marker panel against using either only diagnostic markers, those fixed for different alleles in each parental population, ancestry informative markers, those with distinct allele frequency differences in the two parental

populations (Shriver et al. 2003), or markers indiscriminately sampled from the SNPchip that were neither diagnostic nor ancestry informative. Diagnostic markers (DM) were defined as having allele frequencies of more than or equal to 0.99999 in either direction. Ancestry informative markers (AIM) were defined as having extreme allele frequencies in each population, estimated as (p1-(1-p1)+(1-p2)-p2)/2 >0.95, where p1 is the allele frequency in one population, and p2 is the allele frequency in the other population.

To compare the power of each marker type, without confounding by the number of markers used, we first determined the number of DM that were available, which was 629, and then randomly sampled this number of markers from each of the AIM (excluding DM) and the entire marker panel (excluding DM and AIM). We also included a marker set that was ten times as many randomly sampled SNPs (6290) as well as the entire SNP panel and the 22 microsatellite markers used in Smith et al. 2018. Note that the 629 AIM and the 629 random and 6290 random panels were a new random draw of markers each time the loop was run. We refer to these datasets as 45K, 629 DM, 629 AIM, 629 random, 6290 random later in the text. We then reran the ADMIXTURE analysis with each of the five marker sets for the Kintyre sample and re-estimated the number of individuals in each class (red, sika, hybrid). We focused on the deer from Kintyre because the sample size was larger there because of prior evidence of a large number of hybrid individuals in this region (Senn et al. 2010a), a finding confirmed below. To understand the uncertainty around the number of individuals in each class generated by the different marker sets, we ran each analysis 100 times, and varied the 'set seed' in ADMIXTURE (-s time) to allow for slight differences between replicates. We present the mean estimated number of red deer, hybrid and sika called, as well as the standard errors around each estimate from the replicated analyses.

For each category of marker, we compared the number of deer per iteration that were identified to species class differently when compared to the 45K marker set. We refer to these deer as 'mismatches'. We used a linear model with the number of mismatches as the response variable explained by the marker category to ask which marker category gave results most similar to those from the 45K panel. We report the mean number of mismatches ± standard error, as well as an ANOVA of the linear model. Since we found large differences between marker groups, we explored the location of markers as a possible explanation of the observed differences.

All data and code are available at XX.

Results:

After quality control, there were 513 individual deer from ten sampling areas on the Kintyre peninsula, 108 deer from three sites in the NW Highlands, and 44 999 SNPs (Figure 1). After excluding individuals that were assessed to be hybrids by rangers (N = 30 on Kintyre and N= 27 in the Highlands), and using all markers, F_{st} between the two species was 0.5321 (95% CI = 0.529 – 0.534) in Kintyre and 0.4751 (95% CI = 0.472 – 0.478) in the NW Highlands. Note that the inclusion of putatively diagnostic markers is likely to have inflated F_{st} compared with the true value.

Kintyre

The use of SNP genotyping led to a different classification of many animals when compared to microsatellite genotyping. Using the definitions outlined above, we found 159 pure red deer (i.e. those with a CI that overlapped 0.99999), with an average Q score of 0.9986+/-0.003, 132 pure sika deer (CI overlapped 0.00001), with an average Q score of 0.00169+/- 0.003 and 222 hybrid deer with an average Q score of 0.6144+/- 0.393 (Figure 2a). Hybrid individuals had Q scores that spanned from 0.008 to 0.992, suggesting many generations of backcrossing have elapsed. In comparison to previous work that used 22 microsatellites (e.g. Smith et al. 2018), we classified many more individuals as hybrids (N= 222, vs 119 with microsatellites), and the majority of newly classified hybrids had a Q score of >0.85 (red deer like hybrids) or <0.15 (sika like hybrids). This is in spite of a high correlation of Q score estimates between microsatellites and SNPs (R^2= 0.997, p<2.2e-16). This pattern of many individuals with low levels of introgression is consistent with a bimodal hybrid zone, as has been demonstrated in previous studies (Goodman et al. 1999; Senn & Pemberton 2009). We found three potential F1 individuals, for which 95% CIs overlapped Q=0.5, but on further examination of the heterozygosity of diagnostic markers we determined that these individuals were advanced hybrids with intermediate Q scores, since these individuals were not heterozygous for a high proportion of diagnostic markers. Finally, of the hybrid individuals that we identified, 167 had a red deer mitochondrial marker, and 45 had a sika mitochondrial marker (note that mtDNA was not available for 14 individuals). This is in contrast to 103 hybrids with red mtDNA and 13

individuals with sika mtDNA markers found using microsatellite markers (89.1% red mtDNA, 10.9.% sika mtDNA with microsatellite data vs. 78.7% red mtDNA and 22.2% with SNP data).

NW Highlands

In the NW Highlands, we found 45 red deer (mean Q=0.9999+/-0001), 59 sika (mean Q=0.00001+/-0.0000), and 4 hybrids (Q=0.2333+/-0.4223). This is similar to the 45 red deer, 61 sika and 2 hybrids that were previously reported (Figure 2b). The Q scores for three of these hybrid individuals ranged from 0.0199±0.006 to 0.0251±0.007, i.e. they were very sika like hybrids, while the 4th individual had a Q score of 0.8667±0.133 (red like hybrid). Two of these hybrid individuals had red deer mtDNA; the other two had sika mtDNA. Given the lack of hybrids in this region, we did not conduct further analysis on this dataset.

Diagnostic and Ancestry Informative Marker Panels in Kintyre

We found substantial differences in the number of individuals assigned to each species by each of five marker panels (all 45K, 629 DM, 629 AIM, 629 random sample, 6290 random sample). Generally, we found that the 629 diagnostic markers and the sample of 629 random markers estimated fewer hybrids in the population than the other marker groups (Figure 3). We found that the number of mismatches (deer identified to a different species by a marker type than by the full 45K panel) was affected by marker type such that the 629 AIM had the fewest mismatched individuals (23.91±0.30), followed by the 629 DM (71.89±0.32), the 6290 random sample (82.27±9.0) and 629 random subsample of markers (95.55±5.8). The marker type used explained 20.4% of the variance in mismatches ($F_{3,396}$=33.8, p=2.2e-16).

Inspection of the genomic location of the DM and AIM marker groups revealed a large contrast in genomic location. In the DM group, 269 of the 629 markers (42.8%) were on the X chromosome, while the autosomes had uneven coverage and chromosomes 12, 16, 22, 23 and 25 had no markers at all. In the AIM group, 186 of the 3205 markers (5.8%) were on the X and all chromosomes were marked, though of course this may not have been true of the individual draws in the 629 AIM analyses. Among all markers passing QC, 2360 of 44,999 were on the X. Among those markers originally selected to be diagnostic for sika and actually diagnostic for sika, a remarkable 152/282 or 53.9% were on the X.

Discussion:

We found many more hybrid individuals on Kintyre using 45K SNPs than had previously been reported using 22 microsatellites. The majority of the newly discovered hybrids were those with either a very high or very low Q score, suggesting that we mainly have higher efficiency (i.e. we are more likely to assign true hybrids as hybrid individuals; (Vähä & Primmer 2006) than previous studies of this system. However, it should be noted that we estimated a very similar F_{st} as that reported in previous studies (0.53 compared to 0.58, using microsatellite markers selected to be diagnostic (Senn & Pemberton 2009)), suggesting that while increasing the number of markers allowed us to detect more hybrid individuals, more markers did not greatly affect the estimate of F_{st}.

We reclassified many more individuals as hybrids in Kintyre than in the NW highlands. On Kintyre, we detected 103 hybrids among the 394 individuals that had previously been determined to be parental species. In contrast, we only reclassified 2 individuals (of 106) from the NW Highlands as hybrids using higher density markers. This suggests that reproductive isolation is still strong in the NW highlands, but has broken down to some extent in Kintyre, where a bimodal hybrid zone has formed. This does not necessarily mean that parental type sika and parental type red deer are regularly mating with each other, especially as we identified no F1 individuals in our data set. However, it is the case that there is the opportunity for substantial genetic exchange between red deer and sika in Kintyre. Further, since Kintyre is a peninsula attached to the Scottish mainland, there is the opportunity for it to become a source population from which admixed individuals can disperse. Previous work using microsatellite markers found a significantly higher number of advanced backcrossed individuals when moving away from the point of introduction (Senn et al. 2010a). While the extent of backcrossing did not increase as time progressed (Senn et al. 2010a), if there had been a discrete number of F1s generated at some point, then increased generations of backcrossing would make contemporary, late generation backcrosses very difficult to detect either phenotypically or genetically. In any case, there is an expectation that there are backcrossed individuals in the north and south regions of Kintyre, and an opportunity for sika gene flow into red deer in the rest of Scotland, and thus a conservation concern. Future work could then use higher density markers to re-examine the cline dynamics (both spatial and temporal) along Kintyre to determine if there are temporal changes that were previously

undetectable and if the tension zone of hybridization has moved over time (Barton & Hewitt 1985).

Even with 45K markers, we are unlikely to detect all of the backcrosses that could be present in the Kintyre population. The most extreme Q score that we found in a hybrid (i.e. an individual with credible intervals that excluded 0 or 1) was 0.0078±0.003, which would be the approximate average Q score for a sixth generation backcross. Given that hybridization was already apparent 6 – 7 generations ago (Ratcliffe 1987), there could be more extreme hybrid individuals that we could not detect. Further, more than 1000 diagnostic markers would be needed to detect backcrosses past the 9th generation with high power (McFarlane & Pemberton 2019). With 629 diagnostic markers, we would expect to have high power to detect 8th generation backcrosses (expect to detect 91.5% of 8th generation backcrosses), but this assumes unlinked markers. However, we are certainly getting additional power to detect backcrosses from the other, non-diagnostic markers, so it is difficult to be certain where our detection ability tapers off. Individuals that are the product of consistent backcrossing since an initial F1 hybridization event could be homozygous for parental alleles from one species for all informative markers that we have used here, making them impossible to detect (Boecklen & Howard 1997).

Whether or not all backcrosses are detectable, it is a debatable matter whether there is conservation management value in detecting such advanced backcrosses. If an individual is less than 0.08%, should this individual be considered a hybrid when a management decision is made? This is a difficult question to answer (Allendorf et al. 2001), and we believe it should be answered based on policy and biological considerations, rather than determined based on the statistical power a given set of markers allows. A gene based theory of conservation has been proposed (Petit 2004), in which individuals of conservation concern are determined to be hybrids based on the alleles carried for specific genes of interest for the native parental population (i.e. rare alleles, or alleles associated with distinct phenotypes). However, such 'gene targeted conservation' is probably out of reach for this and many other study systems of conservation concern (Kardos & Shafer 2018). More practically, a mixed strategy of preserving those individuals that are genetically parental species (i.e. narrow Q score CIs that overlap with 0 or 1) *and* score highly on emblematic red deer phenotypes (e.g. summer coat without spots, long pointed ears, large antlers) could reduce the threat to red deer from sika introgression. A

similar strategy is currently in use to preserve both the genetics and phenotypes of Scottish wildcats, where individuals with both high phenotypic and genetic scores are preferentially protected and used in a captive breeding program (Senn et al. 2018).

A surprise in our analysis was that the diagnostic marker panel of 629 markers performed worse, in terms of finding hybrids, compared with equivalent-sized panels of ancestry informative markers (Figure 3). Diagnostic markers are the gold standard in the field (Pritchard et al. 2000), so this observation requires explanation. 71.89±0.32 individuals were mis-categorized by the 629 DM panel when compared to the full marker panel, and these animals had an average Q score of 0.976±0.02 for red like hybrids, and 0.018±0.01 for sika like hybrids, suggesting that the DM panel missed highly backcrossed individuals. The 629 AIM panel, despite comprising slightly less diagnostic markers probably outperformed the 629 DM panel because of a better distribution of the markers across the genome. Any introgression that occurred on a chromosome without diagnostic markers (12, 16, 22, 23 and 25), or in unmarked regions within chromosomes, would not be detected by the DM markers, and the individual would not be assigned as a hybrid. Markers not in the DM or AIM categories, here called 'random markers' generated more mismatches with the full marker set than the 629 AIM panels, and the smaller sample size of 629 such markers was worse in this regard than the larger panel of 6920 such markers. This is as expected if the markers are relatively uninformative but power is gained by having more of them.

The question arises as to why so many of the diagnostic markers were on the X chromosome. Although the selection criteria during the construction of the cervine SNP array included even spacing across the chromosomes (against the bovine map), both within the selected red deer polymorphic and the sika diagnostic loci, we suspect that the categorization into diagnostic versus ancestry informative markers has led to a concentration of X-linked markers into the diagnostic category. Since introgression is impeded on the sex chromosomes because of limited recombination, there may be long runs of fixed markers on the sex chromosomes in any hybridizing system (Barton 1979; Baack & Rieseberg 2007). While thinning markers for linkage will alleviate this problem to some extent, it will also limit the number of informative markers in an analysis, and thus might not be the best solution when more markers will lead to higher power. Many researchers in hybrid systems use anonymous and unmapped markers (e.g. from RADseq SNP discovery) and select diagnostic markers for analysis. Given the

demonstrable downside to using diagnostic markers in our system (concentration on the X at the cost of autosome coverage, leading to fewer identified hybrids) we suggest that future analyses of anthropogenic hybrid zones should include both DM and AIM, and probably all possible markers, in their analyses. We anticipate that this will lead to increased discovery of hybrid individuals, as it has done in the red deer – sika bimodal hybrid zone on Kintyre.

Acknowledgements: We thank the Forestry Commission Scotland rangers, particularly Kevin McKillop, for providing samples, the Wellcome Trust Clinical Research Facility Genetics Core, Edinburgh for performing the genotyping, and Paul Fisher and Rudi Brauning for SNP array development. This work was funded by a European Research Council Advanced Grant to JMP and a Vetenskapsrådet (Swedish Research Council) International Postdoc Fellowship to SEM.

Works Cited:

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics **12**:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research **19**:1655-1664.

Allendorf FW, Leary RF, Spruell P, Wenburg JK. 2001. The problems with hybrids: setting conservation guidelines. Trends in ecology & evolution **16**:613-622.

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. Current opinion in genetics & development **17**:513-518.

Barton N. 1979. The dynamics of hybrid zones. Heredity **43**:341.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. Ann. Rev. Ecol. Syst. **16**:113-148.

Boecklen WJ, Howard DJ. 1997. Genetic analysis of hybrid zones: numbers of markers and power of resolution. Ecology **78**:2611-2616.

Boyer MC, Muhlfeld CC, Allendorf FW. 2008. Rainbow trout (Oncorhynchus mykiss) invasion and the spread of hybridization with native westslope cutthroat trout (Oncorhynchus clarkii lewisi). Canadian Journal of Fisheries and Aquatic Sciences **65**:658-669.

Brauning R, Fisher PJ, McCulloch AF, Smithies RJ, Ward JF, Bixley MJ, Lawley CT, Rowe SJ, McEwan JC. 2015. Utilization of high throughput genome sequencing technology for large scale single nucleotide polymorphism discovery in red deer and Canadian elk. bioRxiv:027318.

Cook CE, Wang Y, Sensabaugh G. 1999. A Mitochondrial Control Region and CytochromebPhylogeny of Sika Deer (Cervus nippon) and Report of Tandem Repeats in the Control Region. Molecular phylogenetics and evolution **12**:47-56.

Coulson T, Pemberton J, Albon S, Beaumont M, Marshall T, Slate J, Guinness F, Clutton-Brock T. 1998. Microsatellites reveal heterosis in red deer. Proceedings of the Royal Society of London B: Biological Sciences **265**:489-495.

Coyne JA, Orr HA 2004. Speciation. SinauerAssociates Inc., Sunderland, Massachusetts, USA.

Crispo E, Moore JS, Lee‐Yaw JA, Gray SM, Haller BC. 2011. Broken barriers: Human‐induced changes to gene flow and introgression in animals. BioEssays **33**:508-518.

Gompert Z, Buerkle CA. 2016. What, if anything, are hybrids: enduring truths and challenges associated with population structure and gene flow. Evolutionary applications **9**:909-923.

Goodman SJ, Barton NH, Swanson G, Abernethy K, Pemberton JM. 1999. Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus Cervus) in Argyll, Scotland. Genetics **152**:355-371.

Gosselin T, Anderson EC, Bradbury I. 2016. assigner: Assignment Analysis with GBS/RAD Data using R, Available from https://github.com/thierrygosselin/assigner.

Grabenstein KC, Taylor SA. 2018. Breaking Barriers: Causes, Consequences, and Experimental Utility of Human-Mediated Hybridization. Trends in Ecology & Evolution.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps‐Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired‐end RAD sequencing. Molecular ecology **22**:3002-3013.

Huisman J, Kruuk LE, Ellis PA, Clutton-Brock T, Pemberton JM. 2016. Inbreeding depression across the lifespan in a wild mammal population. Proceedings of the National Academy of Sciences **113**:3585-3590.

Kahle D, Wickham H. 2013. ggmap: Spatial Visualization with ggplot2. R Journal **5**.

Kardos M, Shafer AB. 2018. The Peril of Gene-Targeted Conservation. Trends in Ecology & Evolution.
Mallet J. 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution **20**:229-237.
McFarlane SE, Pemberton JM. 2019. Detecting the true extent of introgression during anthropogenic hybridization. Trends in Ecology & Evolution.
Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution **58**:2064-2078.
Petit RJ. 2004. Biological invasions at the gene level. Diversity and Distributions **10**:159-165.
Pilot M, Greco C, vonHoldt BM, Randi E, Jędrzejewski W, Sidorovich VE, Konopiński MK, Ostrander EA, Wayne RK. 2018. Widespread, long‐term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. Evolutionary Applications **11**:662-680.
Powerscourt V. 1884. On the acclimatization of the Japanese deer at Powerscourt. Pages 207-209. Proceedings of the Zoological Society of London. Wiley Online Library.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945-959.
Purcell S, et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics **81**.
Qvarnström A, Bailey RI. 2009. Speciation through evolution of sex-linked genes. Heredity **102**:4.
Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics **197**:573-589.
Randi E. 2008. Detecting hybridization between wild species and their domesticated relatives. Molecular ecology **17**:285-293.
Ratcliffe P. 1987. Distribution and current status of sika deer, Cervus nippon, in Great Britain. Mammal Review **17**:39-58.
Rhymer JM, Simberloff D. 1996. Extinction by hybridization and introgression. Annual Review of Ecology and Systematics:83-109.
Sæther SA, Sætre G-P, Borge T, Wiley C, Svedin N, Andersson G, Veen T, Haavie J, Servedio MR, Bureš S. 2007. Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. science **318**:95-97.
Schartl M. 1988. A sex chromosomal restriction-fragment-length marker linked to melanoma-determining Tu loci in Xiphophorus. Genetics **119**:679-685.
Senn HV, Barton NH, Goodman SJ, Swanson G, Abernethy K, Pemberton JM. 2010a. Investigating temporal changes in hybridization and introgression in a predominantly bimodal hybridizing population of invasive sika (Cervus nippon) and native red deer (C. elaphus) on the Kintyre Peninsula, Scotland. Molecular Ecology **19**:910-924.
Senn HV, Ghazali M, Kaden J, Barclay D, Harrower B, Campbell RD, Macdonald DW, Kitchener AC. 2018. Distinguishing the victim from the threat: SNP‐based methods reveal the extent of introgressive hybridization between wildcats and domestic cats in Scotland and inform future in situ and ex situ management options for species restoration. Evolutionary applications.
Senn HV, Pemberton JM. 2009. Variable extent of hybridization between invasive sika (Cervus nippon) and native red deer (C. elaphus) in a small geographical area. Molecular ecology **18**:862-876.

Senn HV, Swanson GM, Goodman SJ, Barton NH, Pemberton JM. 2010b. Phenotypic correlates of hybridisation between red and sika deer (genus Cervus). Journal of Animal Ecology **79**:414-425.

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Human genetics **112**:387-399.

Slate J, Van Stijn TC, Anderson RM, McEwan KM, Maqbool NJ, Mathias HC, Bixley MJ, Stevens DR, Molenaar AJ, Beever JE. 2002. A deer (subfamily Cervinae) genetic linkage map and the evolution of ruminant genomes. Genetics **160**:1587-1597.

Smith SL, Carden RF, Coad B, Birkitt T, Pemberton JM. 2014. A survey of the hybridisation status of Cervus deer species on the island of Ireland. Conservation Genetics **15**:823-835.

Smith SL, Senn HV, Pérez‐Espona S, Wyman MT, Heap E, Pemberton JM. 2018. Introgression of exotic Cervus (nippon and canadensis) into red deer (Cervus elaphus) populations in Scotland and the English Lake District. Ecology and Evolution.

Todesco M, Pascual MA, Owens GL, Ostevik KL, Moyers BT, Hübner S, Heredia SM, Hahn MA, Caseys C, Bock DG. 2016. Hybridization and extinction. Evolutionary applications **9**:892-908.

Vähä J-P, Primmer CR. 2006. Efficiency of model‐based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Molecular Ecology **15**:63-72.

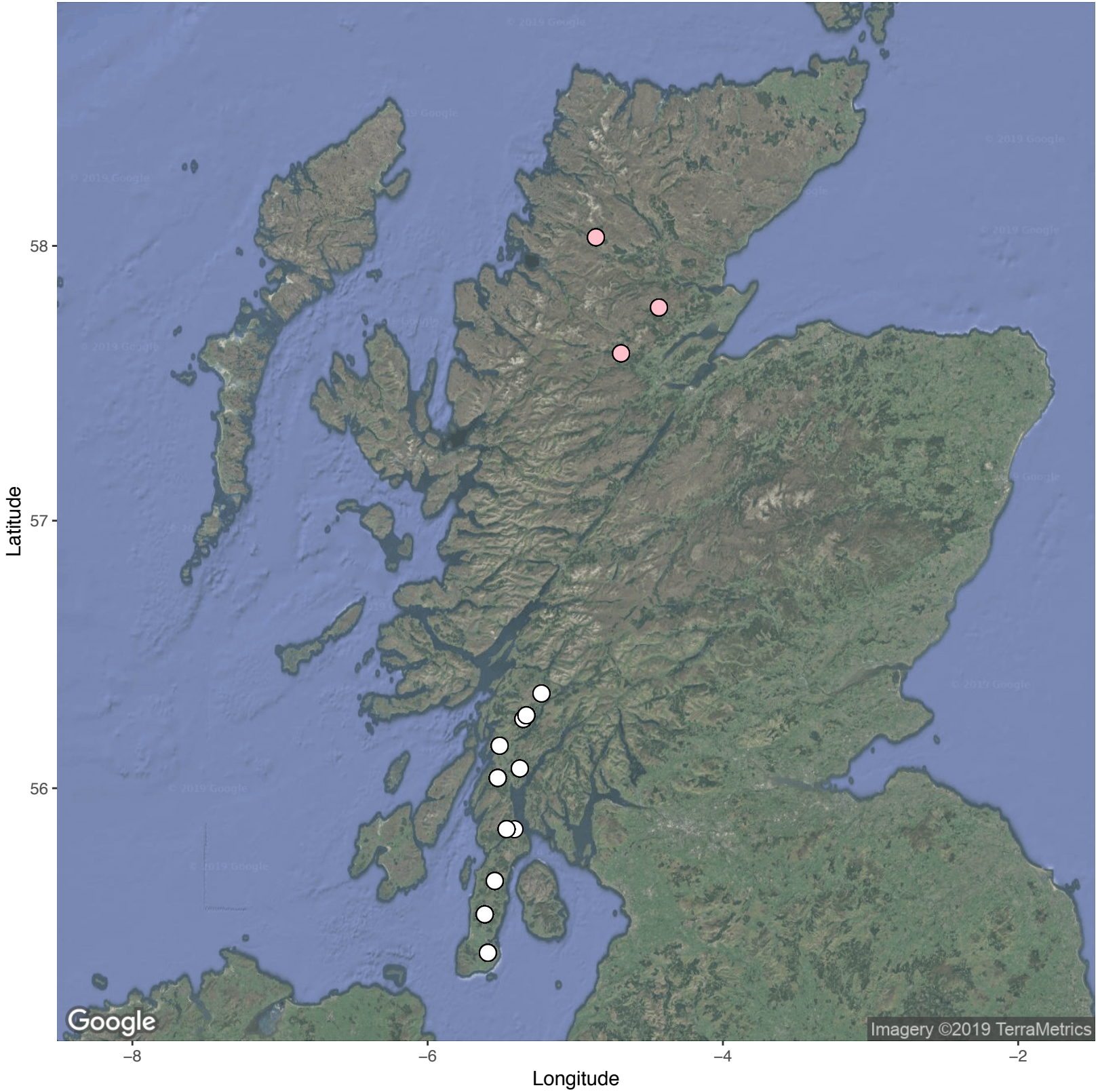
Weir BS, Cockerham CC. 1984. Estimating F‐statistics for the analysis of population structure. Evolution **38**:1358-1370.

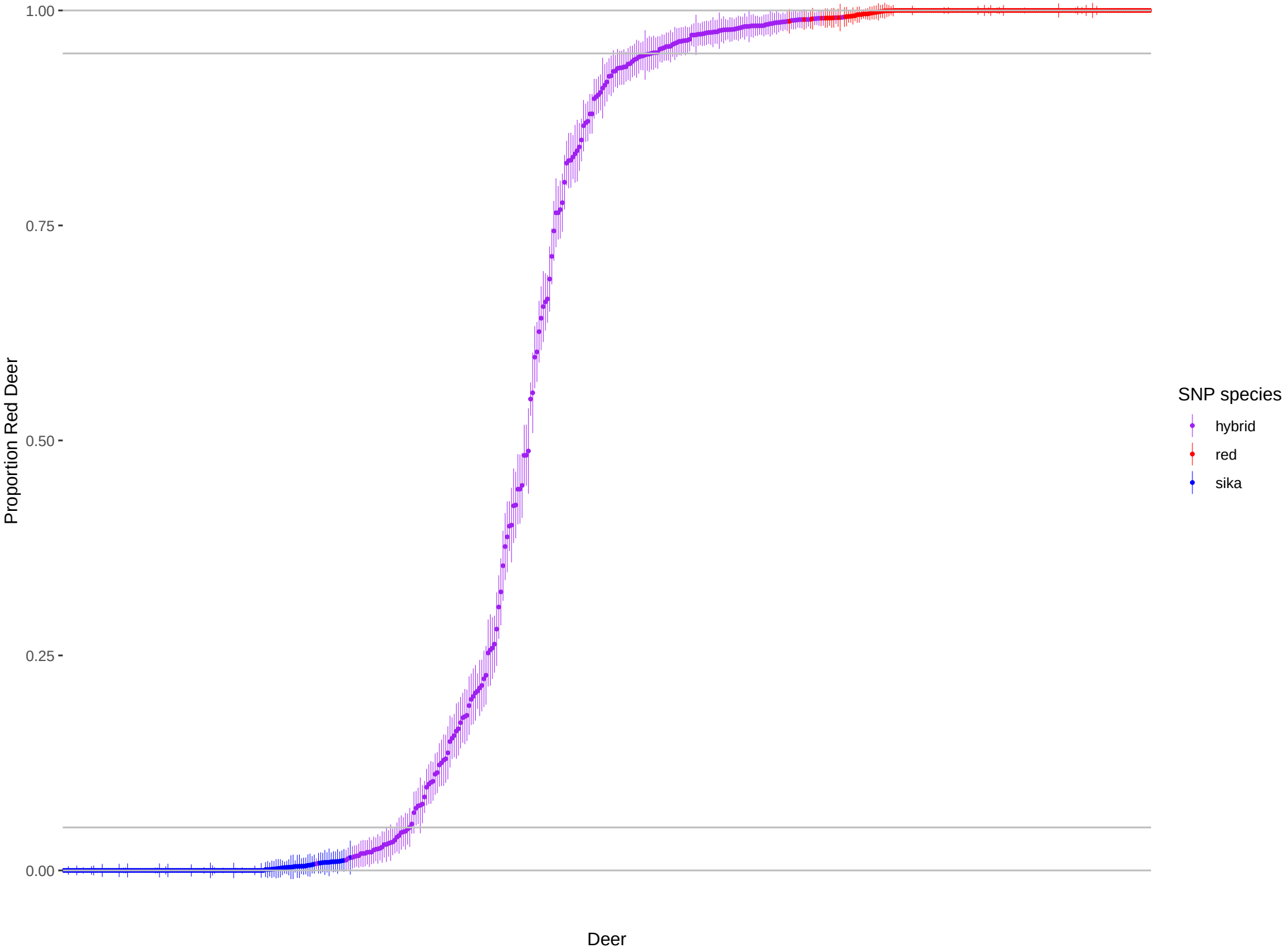
Figures:

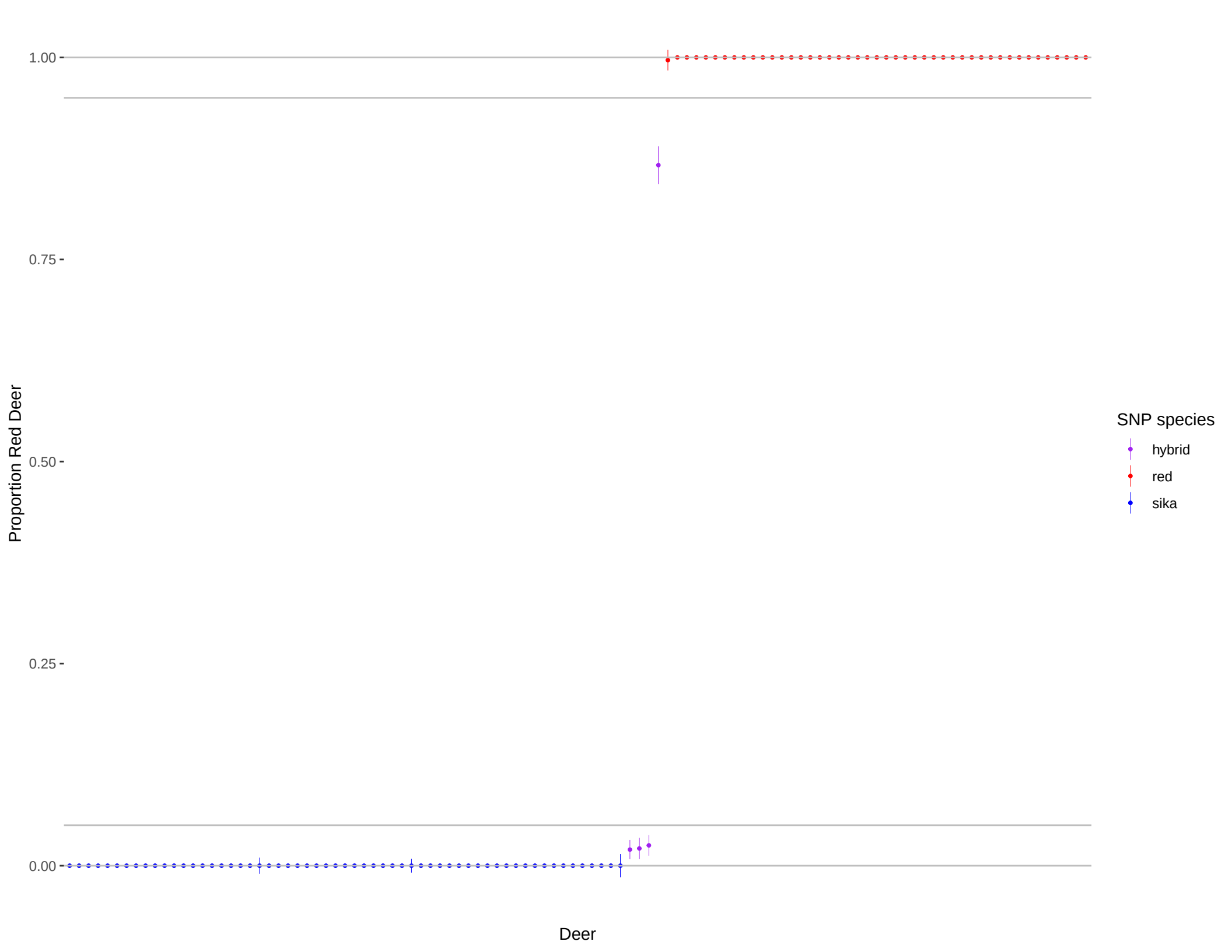
Figure 1: A map of Scotland, showing the approximate sampling locations for deer from Kintyre (white) and the NW Highlands (pink) that were genotyped on a 50K cervine SNP array. The map is from Google Maps, accessed using ggmap (Kahle & Wickham 2013).

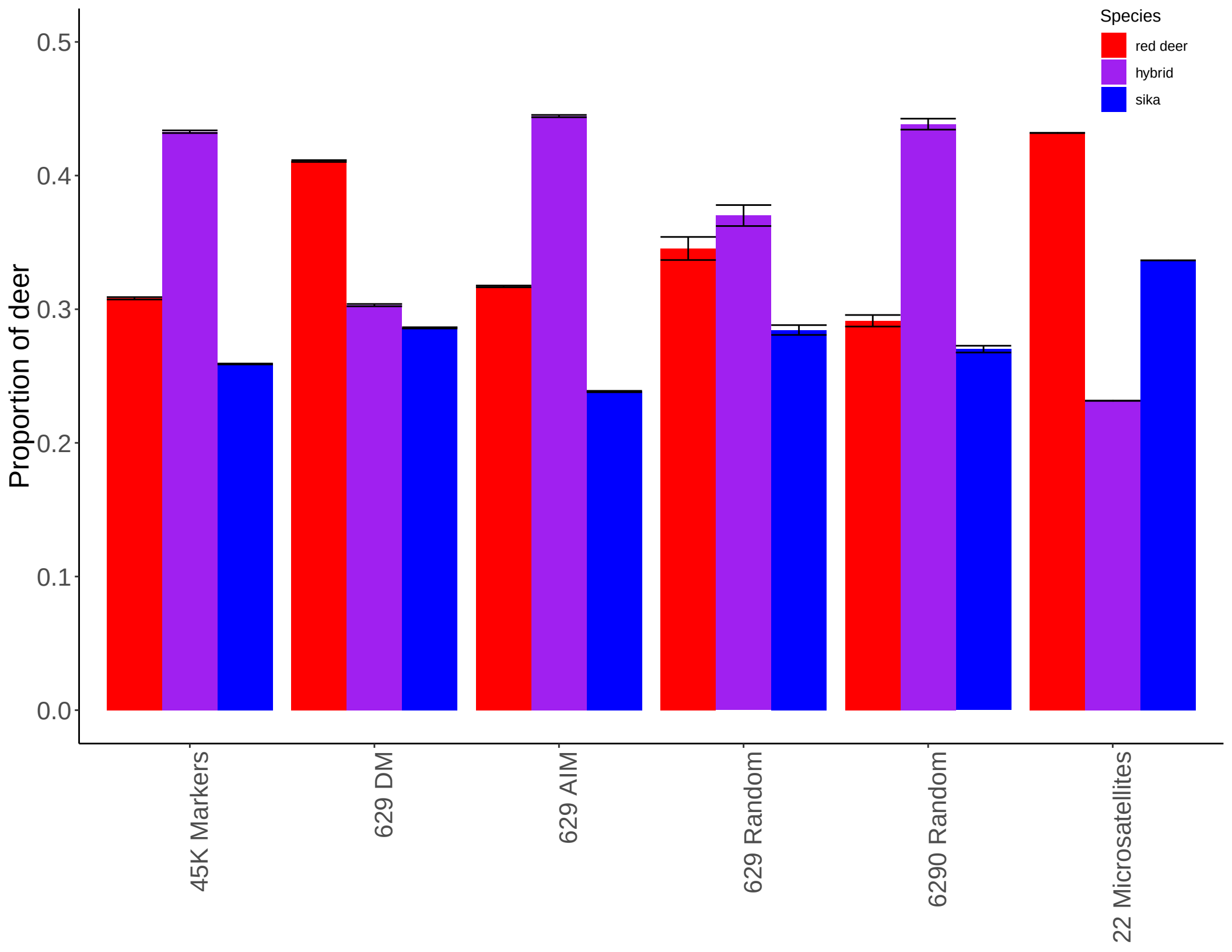
Figure 2: Estimates of admixture proportion (Q scores) and 95% confidence intervals for individual deer from (A) Kintyre and (B) the NW Highlands using 44,999 SNPs. Individuals are arranged by Q score following ADMIXTURE analysis. Individuals were assessed as members of the parental species if the 95% CI overlapped either 0.99999 (red deer; shown in red) or 0.00001 (sika, shown in blue), otherwise individuals were assessed as hybrid (purple). Horizontal grey lines indicate 0, 0.05, 0.95 and 1, where 0.05 and 0.95 were the thresholds used to assign species in Smith et al. 2018.

Figure 3: The number of deer assigned to each species category (red deer, hybrid or sika) according to all 44,999 (45K) markers, a panel of 629 diagnostic makers (629 DM), panels of 629 ancestry informative markers (629 AIM), or sets of either 629 or 6290 random markers (629 and 6290 random). These marker panels are then compared to the microsatellite results from Smith et al. (2018). Markers were assessed as either AIM or DM using posterior allele frequencies in ADMIXTURE (see text). Each assessment illustrated was run 100 times to allow for variation in the chosen panels (except for the DM panel) which numbered 629 markers, and with different start points for each analysis. Markers are non-nested, such that AIM do not include DM, and the random subsets of markers do not include DM or AIM. Red deer are depicted in red, hybrids are purple and sika are blue.


58
57
56
Latitude
−8
−6
−4
−2
Longitude
Google
Imagery ©2019 TerraMetrics


1.00
0.75
0.50
0.25
0.00
Proportion Red Deer
Deer
SNP species
hybrid
red
sika


1.00
0.75
0.50
0.25
0.00
Proportion Red Deer
Deer
SNP species
hybrid
red
sika


Species
red deer
hybrid
sika
Proportion of deer
0.0
0.1
0.2
0.3
0.4
0.5
45K Markers
629 DM
629 AIM
629 Random
6290 Random
22 Microsatellites