**Crop-to-wild introgression in the European wild apple *Malus sylvestris* in
Northern Britain**

**MARKUS RUHSAM^{1*}, WILL JESSOP¹, AMANDINE CORNILLE², JAMES RENNY³ and RICK
WORRELL³**

¹*Royal Botanic Garden Edinburgh, Edinburgh, Midlothian, EH3 5LR, UK*

²*GQE– Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190, Gif-
sur-Yvette, France*

³*Forestry Consultancy, Aberfeldy, Upper Park, PH15 2EH, UK*

*Corresponding author. Tel: +44 131 248 2859; Email: m.ruhsam@rbge.ac.uk

Malus sylvestris (L.) Mill, the European wild apple, is a small tree native to Europe which reaches its north-western limit in Northern Britain. It has been identified as one of the main contributors to the domesticated apple *M. domestica* (Suckow) Borkh. There are concerns that wild populations of this rare tree are threatened by hybridisation with *M. domestica* throughout Europe. We genotyped 332 trees collected in the wild from Scotland and Northern England using 14 microsatellite markers to investigate levels of introgression in wild *M. sylvestris* populations. Our results showed that 70 % of the samples could be considered pure *M. sylvestris*, 27% showed varying degrees of introgression from *M. domestica* and 3% were pure *M. domestica*. Hybrids were more frequent in areas with more intensive land use and less semi-natural woodland. About 80% of hybrids were identified as backcrosses to *M. sylvestris*. Areas in the Southern Highlands had the highest frequency of pure *M. sylvestris* trees (87%). Morphological characters such as leaf size and hairiness as well as fruit size generally used by botanists for identification of *M. sylvestris* in the field proved to be unreliable as only 68% of the trees could be identified accurately. Given the use of wild apple trees it seems likely that many of the trees collected from agricultural areas are planted and not the product of natural *in situ* hybridisation and seed dispersal events thus compounding natural hybridisation rates. Areas with high frequencies of pure *M. sylvestris* trees should be prioritised for conservation and possible measures are discussed to safeguard *M. sylvestris* as a species in Northern Britain.

Introduction

European wild apple (*Malus sylvestris* (L.) Mill., crab apple), a small tree in the Rosaceae family, is native across Europe and reaches its north-western distribution limit in Scotland and Norway (<http://www.euforgen.org>, accessed 9 July 2018). It is a light demanding species which grows naturally in gaps within or at the edge of broadleaf forests or on low competition sites such as flood plains (Schnitzler *et al.*, 2014). However, it is more often associated with anthropogenic habitats such as hedges, wood pastures, semi-natural grasslands, and meadows (Buttenschön and Buttenschön, 1999). Wild apple is a rare tree species in Europe which is red-listed in Belgium (Coart *et al.*, 2003), Finland and Germany (Kik *et al.*, 2011), and part of conservation efforts in Denmark (Graudal *et al.*, 1995). This species mainly occurs as single trees scattered throughout the landscape at low densities and for example it is estimated that only about 5500 trees are left in Germany (BLE, 2013). *Malus sylvestris*, like all other apple species, is outbreeding due to a self-incompatibility system (Broothaerts, 2003) and is mainly pollinated by bees (Palmer-Jones and Clinch, 1966). Birds and mammals such as hares and wild pigs have been implicated in seed dispersal, however, there is evidence that domestic cattle are likely to be the main distributor of viable wild apple seeds (Buttenschön and Buttenschön, 1999).

Recent studies have highlighted the important recent contribution of *M. sylvestris* to the cultivated apple genome, *M. domestica* (Suckow) Borkh., during the domestication history from the Central Asian progenitor *M. sieversii* (Ldb.) M. Roem. (Cornille *et al.*, 2012, 2014). It has long been recognised that wild apple and domesticated apple hybridise and that it can be very difficult to reliably identify pure wild apple from hybrids using morphological characters such as the fruit width and the hairiness of leaves (Wagner, 1996; Stephan *et al.*, 2003). Not only has this led to discussions whether genuine wild *M. sylvestris* still exists today (Henning, 1947; Remmy and Gruber, 1993; Wagner, 1996) but has also raised concerns about the genetic integrity of wild apple. This has recently sparked a number of studies which investigated the extent of hybridisation between the two apple species in Europe using molecular markers (Coart *et al.*, 2003, 2006; Larsen *et al.*, 2006; Schnitzler *et al.*, 2014; Wagner *et al.*, 2014; Cornille *et al.*, 2015; Feurtey *et al.*, 2017). Although results varied between studies the overall picture emerged that hybrids may occur at significant frequencies in wild apple populations (7% to 36%, Cornille *et al.*, 2014).

61 *Malus sylvestris* is an understudied species in Northern Britain where it occurs at the north-western
62 limit of its distribution. Yet, there has been a long standing history of domestic apple cultivation in
63 Scotland which can be traced back to the 12th century (Hayes, 2008). Apples were also an important
64 commercial crop during the 19th and 20th century, a time when many orchards were established,
65 especially in the Tay and Clyde valleys (Ironsides Farrar, 2004). However, the extent of crop to wild
66 gene flow in *M. sylvestris* has never been assessed in the United Kingdom and although the general
67 distribution of crab apple is well documented in the online database of British plants
68 (<https://database.bsbi.org/maps/>, accessed 9 July 2018), many of these records might be
69 questionable due the difficulty in distinguishing pure wild apple from hybrids morphologically.

70 This raises the question of how many crab apple trees in the Scottish landscape are really pure *M.*
71 *sylvestris* given the ample opportunity *M. domestica* and *M. sylvestris* have had to hybridise.

72 The aims of this study are therefore 1) to assess the frequency of *M. sylvestris* and hybrid trees in the
73 wild 2) to assess the reliability of morphological characters used in identifying pure and hybrid
74 samples in the field and 3) to discuss implications for the conservation of this species in Northern
75 Britain.

Sampling and plant material

Our fieldwork comprised both locating trees and attempting to develop a representative sample as the project progressed. Sampling was carried out based on field records in the BSBI online atlas for *Malus sylvestris* s. l. (<https://database.bsbi.org/maps/>, accessed 9 March 2017) and the Forestry Commission Native Woodland Survey of Scotland (<https://scotland.forestry.gov.uk/supporting/strategy-policy-guidance/native-woodland-survey-of-scotland-nwss>, accessed 9 March 2017). Additionally, samples were located via professional contacts and by searching likely habitats encountered during field work. However, time and budget constraints meant that it was impossible to search the whole of Northern Britain comprehensively although most areas were covered (Figure 1). Usually all trees encountered were sampled, but in a few woods with multiple trees, samples were selected to cover the range of morphological characteristics present. The collected trees were allocated to one of the following seven geographical regions (Figure 1): Lake District (LD, n=24), Dumfries (DF, n=89), Central Belt (CB, n=66), Southern Highlands (SH, n=122), West Coast (WC, n=9), Aberdeenshire (AB, n=9) and Northern Highlands (NH, n=13, including one sample from Shetland). The delimitation of these regions is broadly based on geographical features and the intensity of land usage. The regions of the Lake District (LD), Dumfries (DF), Southern and Northern Highlands (SH and NH, respectively) are mainly upland, characterised by less intense agricultural usage, relatively high incidence of semi-natural woodland and lower population densities. In contrast, the areas where samples were collected in the Central Belt (CB), Aberdeenshire (AB) and the West Coast (WC) are mainly low lying with little remaining semi-natural woodland and more intense agricultural land use and/or higher population densities.

A total of 366 silica dried leaf samples were assayed including 332 trees collected throughout Scotland and adjacent parts of Northern England 34 trees served as reference samples for six *Malus* species including 24 samples from the living apple collection of the Royal Botanic Garden Edinburgh (*Malus domestica* (n=11), *M. baccata* (L.) Borkh. (n=5), *M. hupehensis* (Pamp.) Rehd. (n=3), *M. orientalis* Uglitzk. (n=2) and *M. sieversii* (n=3)) as well as 10 samples from France which were used in Cornille et al. (2015) and were shown to be genetically pure *M. domestica* (n=6) and *M. sylvestris* (n=4).

Microsatellite markers and polymerase chain reaction (PCR) amplification

DNA was extracted using the DNeasy Plant kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. All samples were assayed for 15 microsatellite loci using the four multiplex reactions MP01, MP02, MP03 and MP04 from Cornille *et al.* (2012; Table S9 and Supplementary Table S1 this publication). The suitability of these markers for population genetic analyses, i. e. their neutral and unlinked nature as well as the absence of null alleles has been demonstrated in previous studies (Cornille *et al.*, 2012, 2015; Cornille *et al.*, 2013a; Cornille *et al.*, 2013b). Each multiplex PCR was carried out in volumes of 10 µl using the following protocol: 1x PCR buffer (Bioline, London, UK), 0.2 mM dNTP (Bioline, London, UK), 2.5 mM MgCl₂, 0.2 µM of each forward and reverse primer in the multiplex, 1x polysorbate-20 (Tween-20) (TBT-Par), 0.05 units BioTaq (Bioline, London, UK) and 1 µL of unquantified DNA. The mixture was then cycled through the profile: 5 min at 95°C, 35 cycles of 30 sec at 94°C, 1.5 min at 55°C and 1 min at 72°C, ending with 20 min at 72°C to complete extension. All PCR reactions were performed on a Biorad Tetrad 2 thermal cycler and run on an ABI 3730 automated sequencer at the Edinburgh Genomics facility. LIZ-500 labelled internal size standard (Applied Biosystems, Foster City, California, USA) was added to each sample to size fragments. The data generated were analysed with the software GENEMAPPER, v 4.1 (Applied Biosystems, Foster City, California, USA) scoring electropherograms manually.

Morphology

We used the morphological characters described in field guides to delimit *M. sylvestris* from *M. domestica*: the lack of hairs on the petiole and underside of the leaves, the fruit size (< 30 mm) and leaf size (Stace, 2010). However, it has been shown that there is large morphological variability within species and that hybrid samples display a continuum of intermediate characters which makes identification based on a few characters challenging (Wagner, 1996). Nevertheless, this is the reality of field work and apple trees recorded in online databases will have been identified based on this limited number of morphological characters, some of which might not have been available at the time of identification, such as fruit size. To test the reliability of such a procedure we identified trees based on the available characters (leaf hairiness, leaf and fruit size) at the time of leaf collection for genetic analysis. Hairiness was assessed on a scale of 0-4 (0= no hairs or very few hairs on petiole only; 4 = felted hairs on petiole and all of lower leaf surface) using a schema similar to Tollefsrud *et al.* (2014).

Trees were recorded either as 'crab apple', 'hybrid', 'cultivated apple', 'possibly crab apple' or 'possibly hybrid'.

Statistical analyses

STRUCTURE v.2.3.4 (Pritchard *et al.*, 2000) was used to investigate the population structure and levels of admixture in the 332 field collected and 34 reference *Malus* samples. This software uses a Bayesian algorithm to infer the number of distinct genetic groups K by minimizing deviations from Hardy–Weinberg and linkage equilibrium within each cluster. The analyses were carried out for $K=1$ to 10 using 1,000,000 MCMC iterations after a burnin of 30,000 steps and were repeated 20 times for each K , with the 'admixture' and 'correlated allele frequencies' options selected. If sampling is unbalanced, i.e. genetic clusters have widely different sample sizes, then STRUCTURE has been shown to yield poor estimates of individual ancestry and K if the default settings are used (Wang, 2017). We therefore followed Wang's (2017) recommendation and used the alternative option allowing a separate α which is a measure of the relative admixture level between populations ('Separate α for each Population' ticked). This option assumes distinct α values for the assumed K populations, and allows individuals to have their ancestry originating from the assumed K populations at K different prior probabilities. In contrast, the default setting uses a single α value for all populations and specifies that each individual has its ancestry originating from each of the assumed K populations at an equal prior probability of $1/K$. We used StrAuto (Chhatre and Emerson, 2017), a python script, which automatically distributes STRUCTURE jobs among multiple processors reducing analysis time considerably. To identify the most likely number of distinct genetic groups K in the dataset the ΔK statistic (Evanno *et al.*, 2005), which is based on the rate of change in the log probability of data between successive K values, as well as $\text{Pr}[X|K]$ (the probability of obtaining the genotype data X given K) of Pritchard *et al.* (2000) were calculated using the online version of the software CLUMPAK (Kopelman *et al.*, 2015, <http://clumpak.tau.ac.il>). STRUCTURE output files were processed using the R package 'pophelper' (Francis, 2017).

To assess the frequency of pure and hybrid apple trees in our dataset we considered two approaches. The first approach used the estimated individual membership coefficients from STRUCTURE (Pritchard *et al.*, 2000). These are the proportions P for each of the identified (K) genetic clusters that have contributed to an individual's genomic makeup (Pritchard *et al.*, 2000).

There is no objective threshold for P which can be used to define pure and hybrid samples but a value ranging between 0.05 to 0.2 has been generally applied in the literature (Larsen *et al.*, 2006; Vähä and Primmer, 2006; Starr *et al.*, 2013; Feurtey *et al.*, 2017). A higher threshold can result in the misclassification of hybrids as pure and a lower threshold might have the opposite effect and identify a sample as hybrid when it is actually pure. In previous *Malus* studies values of $P \leq 0.1$ (Cornille *et al.*, 2013b; Cornille *et al.*, 2015), $P \leq 0.15$ (Larsen *et al.*, 2006) and $P \leq 0.2$ (Coart *et al.*, 2006; Wagner *et al.*, 2014; Feurtey *et al.*, 2017) have been used to classify pure individuals. We therefore calculated the percentage of pure and hybrid samples using all of the above thresholds, resulting in three scenarios. In a conservative scenario only samples with a membership coefficient for the *M. domestica* gene pool $P_{\text{dom}} \leq 0.1$ were considered to be pure *M. sylvestris* trees. Samples with $0.10 < P_{\text{dom}} < 0.45$ were classed as backcrosses to *M. sylvestris*, with $0.45 < P_{\text{dom}} < 0.55$ as F1, with $0.55 < P_{\text{dom}} < 0.90$ as backcrosses to *M. domestica* and with $P_{\text{dom}} \geq 0.90$ as pure *M. domestica*. In the two other less stringent scenarios samples with $P_{\text{dom}} \leq 0.15$ and $P_{\text{dom}} \leq 0.2$ were considered to be pure *M. sylvestris* and samples with $P_{\text{dom}} \geq 0.85$ and $P_{\text{dom}} \geq 0.80$ as pure *M. domestica*. The threshold for the other classes stayed the same apart from the backcross boundaries being adjusted to reflect the altered threshold for purity. The second approach employed the software NEWHYBRIDS v1.1 (Anderson and Thompson, 2002), which uses a Bayesian algorithm to estimate the probability of individuals belonging to any one of six distinct classes (pure parental, F1, F2 and backcross categories). We allocated individuals to a specific class if they had a posterior probability $\geq 55\%$ of belonging to this class. However, if the probability was split more or less equally between two classes, e. g. 45% backcross to *M. sylvestris* and 55% F1 then the individual status was left undecided. The analysis was carried out with the default parameters, Jeffreys priors, a burn-in phase of 100,000 steps and 500,000 sweeps. No prior information about pure parental samples was included.

To investigate the population structure of pure *M. sylvestris* samples in Scotland a second STRUCTURE analysis with the same parameters specified above was carried out on a subset of the data. The dataset only included pure *M. sylvestris* samples i.e. samples which had a membership coefficient for the *M. domestica* gene pool $P_{\text{dom}} \leq 0.1$ in the analysis using the full dataset.

Standard genetic diversity estimates and population differentiation were calculated for all pure *M. sylvestris* populations ($P_{\text{dom}} \leq 0.1$) with more than ten pure individuals using the R package 'diveRsity'

192 (Keenan *et al.*, 2013) and GenAEx v6.5 (Peakall and Smouse, 2012). To test for isolation-by-distance
193 a Mantel test was carried out between the geographic and genetic distance matrix of pure samples in
194 GenAEx v6.5 using 9999 permutations (Peakall and Smouse, 2012).

195 The full dataset was also subjected to a multivariate analysis (Principal Coordinate Analysis - PCO) in
196 GenAEx v6.5 (Peakall and Smouse, 2012) using the pairwise genetic distances and the
197 'standardised covariance' option.

198 To assess the reliability of morphology in identifying apple trees correctly, field and genetic
199 identification using the $P_{\text{dom}} \leq 0.1$ threshold were compared for each sample.

Results

Genetic structure

Microsatellite locus CH04e05 in MP02 was difficult to score and was excluded from further analyses which are therefore based on 14 loci only. The STRUCTURE analysis using the full dataset (n=366) indicated that, depending on the evaluation statistic used, the most likely number of genetic groups was either K=2 (Evanno *et al.*, 2005) or K=3 (Pritchard *et al.*, 2000) (Supplementary Figure S1). For K=2, a *M. sylvestris* group and a group containing cultivated samples and the other four reference *Malus* species were differentiated. For K=3, a *M. sylvestris* and a *M. domestica*/*M. sieversii* group were retrieved, as well as a group containing the other three *Malus* species (*M. hupehensis*, *M. baccata* and *M. orientalis*). We considered K=3 to be the most likely number of groups in our dataset for the following reasons.

Firstly, it has been shown that overall Pritchard *et al.*'s (2000) method is more accurate than Evanno *et al.*'s (2005) especially if sampling is unbalanced, i.e. the inferred genetic groups have unequal sample sizes (Wang, 2017). This is the case in our dataset where the number of reference samples is much smaller than the number of samples collected in the field. Secondly, K=3 differentiated between a *M. domestica* and *M. baccata* - *M. hupehensis* - *M. orientalis* group which is biologically a more meaningful grouping than an undifferentiated group containing five different *Malus* species. This is also more consistent with Cornille *et al.*'s (2012) analysis of a much larger dataset which yielded distinct groups for each of their reference *Malus* species (*M. baccata*, *M. sieversii*, *M. orientalis*, *M. domestica*). The reason why the different reference *Malus* species did not fall into their own respective clusters in our dataset is probably due to the very small number of samples in each reference group (n=2 to 5 per species), and the existence of considerable levels of admixture and ancestral polymorphism which often clusters *M. baccata* and *M. hupehensis* as well as *M. domestica* and *M. sieversii* together (Cornille *et al.*, 2012; Duan *et al.*, 2017). Thirdly, the K=3 analysis resulted in the same genetic clusters and individual membership coefficients in all 20 repeat runs whereas results for K>3 produced up to four alternative ('minor') modes suggesting conflicting signals in these clustering solutions (Supplementary Figure S2).

To assess levels of introgression in Northern British populations of *M. sylvestris* we combined the estimated membership coefficients coming from the *M. domestica* and *M. baccata* - *M. hupehensis* - *M. orientalis* group to obtain a figure for the total contribution of genetic clusters other than *M. sylvestris* in each individual. As individual contributions from the *M. baccata* - *M. hupehensis* - *M. orientalis* group were well below 1% for the vast majority of samples (Figure 2) we subsequently refer to this combined figure as *M. domestica* contribution.

The STRUCTURE analysis revealed considerable levels of introgression in the trees collected throughout Scotland and Northern England (Figure 2, Supplementary Table S2). We have used three thresholds for the permissible proportion of *M. domestica* P_{dom} ($P_{\text{dom}} \leq 0.1, 0.15$ and 0.2) in a pure *M. sylvestris* sample. This resulted in somewhat differing proportions for pure and hybrid classes (Table 1) and generated uncertainty for a few individuals regarding their 'true' class. However, it did not affect the overall picture and the interpretation is the same regardless of the threshold used. Both the STRUCTURE and NEWHYBRIDS results indicated that although 70 to 80% of the samples collected in the field were likely to be pure *M. sylvestris*, between 16 to 27% of the samples showed varying degrees of introgression from *M. domestica* (Figure 2, Table 1). Samples of hybrid origin were mainly identified as backcrosses to *M. sylvestris* (12%-21%) whereas F1 (~3%), F2 (0.3%) and backcrosses to *M. domestica* (1%-2%) were rare. There was generally good agreement between the two different methods used to classify hybrid and pure samples (Table 1). A small percentage of samples (4%) could not be confidently allocated to a single category by NEWHYBRIDS as they shared a more or less equal probability to be either a backcross to *M. sylvestris* or a pure *M. sylvestris* (3.7%) or an F1 (0.3%) or F2 (0.3%) individual (Table 1).

The PCO analysis revealed a relatively well defined cluster of reference individuals consisting of *M. domestica* and reference samples for the other four *Malus* species (*M. baccata*, *M. hupehensis*, *M. orientalis* and *M. sieversii*) which was well separated from a pure *M. sylvestris* group (Figure 3). The space between these two groups was filled by hybrids which intermingled to a large extent with the pure *M. sylvestris* group and to a much lesser extent with the *M. domestica* cloud. Relatively little of the total variation in the dataset was explained by axis 1 (6.5%) and axis 2 (3.6%).

Geographic variation of introgression rates

The proportion of introgressed trees varied considerably between geographical regions. Using the more conservative threshold of $P_{dom} \leq 0.1$ for pure *M. sylvestris* samples, the West Coast, Aberdeenshire and the Central Belt were identified as regions with more than 50% of the trees showing introgression. In contrast, regions with $\leq 25\%$ introgressed trees were the Lake District, Dumfries and the Southern Highlands (Figure 4A). The average contribution of the *M. domestica* genepool to the identified hybrids over all regions was 27% and reached its highest value with 34% on the West Coast (Figure 4B). Interestingly, three trees from the West Coast and two from the Trossachs-Loch Lomond area (SH) had contributions of >0.9 from the *M. baccata* - *M. hupehensis* - *M. orientalis* genetic cluster (Figure 2, Supplementary Table S2).

To assess the population structure within pure *M. sylvestris* populations ($n=227$) we used a pruned data set including only samples which were estimated to have a membership coefficient for the *M. domestica* gene pool $P_{dom} \leq 0.1$ in the first analysis. The analysis did not reveal any population structure within Scotland and Northern England as all samples had more or less equal proportions from K genetic groups for each K from 2 to 6 (Supplementary Figure S3).

Genetic diversity estimates of pure M. sylvestris populations

Four out of the seven geographic regions had more than ten pure ($P_{dom} \leq 0.1$) *M. sylvestris* trees and were used for genetic diversity estimates. Genetic diversity (H_e) and allelic richness (A_r) were generally high in all populations with no evidence for inbreeding (F_{is} not significantly different from zero) (Table 2). Overall population differentiation was significant but very low ($G_{st}=0.018 \pm 0.003$; $D_{est}=0.066 \pm 0.013$). Equally, pairwise population differentiation was very low but significant with the highest differentiation displayed between the Lake District and the Central Belt ($G_{st}=0.012$; Supplementary Table S3). The Mantel test revealed a very weak isolation-by-distance pattern ($R_{xy}=0.05$, $P=0.004$) with hardly any of the variation explained by a linear relationship ($R^2=0.003$). This suggests that there is no pattern of isolation-by-distance in Scottish crab apple, i. e. geographic proximity of trees does not correlate with closer genetic relatedness.

Reliability of identification based on morphological characters taken in the field

A total of 239 trees were used to test the reliability of morphological characters recorded in the field to assess whether trees were either pure or of hybrid origin. Field identifications were compared to the

283 genetic identifications using a threshold of $P_{\text{dom}} \leq 0.1$ for pure *M. sylvestris* samples and $P_{\text{dom}} > 0.1$ for
284 hybrid samples. 58% (n=139) of the samples were correctly identified in the field using morphology
285 (Table 3) whereas 25% (n=59) were identified incorrectly. A total of 17% (n=41) of the samples could
286 not be identified with confidence in the field and were tentatively allocated to either 'possibly crab
287 apple' or 'possibly hybrid'. This 'informed' guess turned out to be correct in 58% (n=24) of the cases
288 and incorrect in 42% (n=17). Over all 68% (n=163) of trees were correctly identified (Table 3).

Discussion

A number of studies have investigated levels of hybridisation between the European wild crab apple (*Malus sylvestris*) and the domesticated apple (*M. domestica*) on different geographical scales ranging from regional (Schnitzler *et al.*, 2014; Wagner *et al.*, 2014) to country (Coart *et al.*, 2003, 2006; Larsen *et al.*, 2006; Wagner *et al.*, 2014; Cornille *et al.*, 2015; Feurtey *et al.*, 2017) and European wide (Cornille *et al.*, 2012, 2013b, 2015). However, no study has ever focused on *M. sylvestris* in the United Kingdom where reliable distribution data are rare and nothing is known about the extent of introgression in this enigmatic tree. One aim of this study was to investigate the frequency of hybrids in wild Northern British *Malus sylvestris* populations which occur in the UK at the north-western edge of their natural distribution range.

Extent of hybridisation between M. sylvestris and M. domestica

The results from both the NEWHYBRIDS and STRUCTURE ($P_{\text{dom}} \leq 0.1$) analyses indicate that around 70% of the trees collected in the field are pure *M. sylvestris* and about 26% are the product of hybridisation. Previous studies on hybridisation between *M. sylvestris* and *M. domestica* reported hybrid proportions of 36.7% [292/796, Europe wide] (Cornille *et al.*, 2013b), 23.1% [436/1189, Europe wide] (Cornille *et al.*, 2015), 13.9% [58/417, Germany] (Wagner *et al.*, 2014), 13.5% [8/59, Luxembourg] (Wagner *et al.*, 2014), 11.2% [20/178, Denmark; 71/634, majority Belgium] (Larsen *et al.*, 2006; Coart *et al.*, 2006), and about 4% [9/255(202), France] (Schnitzler *et al.*, 2014). Although caution is necessary when comparing hybridisation rates from different studies due to differences in marker numbers, samples and applied thresholds, Scottish levels of hybridisation appear to be comparable to the ones found in other European countries. However, there were clear differences in the extent of hybridisation between regions within Scotland. Strongholds for pure *M. sylvestris* trees are clearly the Southern Highlands where only 13% of the samples turned out to be of hybrid origin. This was closely followed by Dumfriesshire and the English Lake District with 25% hybrids. Most of these areas are characterised by rugged terrain, a relatively high proportion of woodland of natural origin and are likely to have relatively low numbers of cultivated apple trees and orchards. In contrast, the Central Belt is the most densely populated area in Scotland which is characterised by low lying fertile agricultural land and a higher incidence of woodland and trees of planted origin. An orchard survey of the Forth Valley (Hayes, 2010) and Clyde Valley (Ironside Farrar, 2004), which are part of

this region, listed 67 and 63 orchards respectively, many of which have grown or grow cultivated apples. It is therefore not surprising that ~ 50% of the assayed samples are of hybrid origin in the Central Belt. Our results indicate similar hybrid frequencies in Aberdeenshire, along the West Coast and in parts of the Northern Highlands, however, these are based on much smaller sample sizes. This suggests that hybrid trees are more common in regions where apple cultivation is more prevalent (and has been for many centuries), and where trees (including crab apples) are typically more likely to have been planted rather than regenerated naturally.

Interestingly, five trees (two from the Southern Highlands and three from the West Coast) had more than 90% of their genetic make up derived from the *M. baccata*-*M. hupehensis*-*M. orientalis* genetic group. These are likely to be either *M. baccata* trees which are sometimes planted as ornamental trees (Stace, 2010, p. 200) and have been recorded from the West Coast (<https://database.bsbi.org/maps/?taxonid=2cd4p9h.asm>, accessed 9 March 2017) or *M. hupehensis* trees which have been planted by, amongst others, the Forestry Commission, as a component of a mix of trees used for landscaping purposes (Peter Quelch, pers. comm.). Although such ornamental trees will contribute to the pollen pool which *M. sylvestris* trees are exposed to, cross-pollination with subsequent hybrid establishment is likely to be a very rare event as all of the assayed hybrid samples had much less than 10% contribution from this genetic group (Figure 2).

Hybrid classes detected in samples from Northern Britain

Our data indicate that the most successful hybridisation events are backcrosses to *M. sylvestris* (crosses between a pure *M. sylvestris* and an F1 or later generation hybrid) as ~20% of all samples and about 80% of all hybrids fall into this category. In comparison, only 3% of the samples (or about 15% of all hybrids) were classified to be F1 (i. e. the mating between a pure *M. sylvestris* and a pure *M. domestica* tree) regardless of analysis. The generation of F1 with the subsequent successful establishment of the hybrid seed seems to be a relatively rare event. Backcrosses to *M. domestica* and F2 seem to be even rarer, as only 1-2% and 0.3%, respectively, were classified as such.

A few studies have investigated pollen dispersal distances of apple trees, and show that the maximum distance can vary considerably from as short as 300 m (Larsen and Kjær, 2009) to as far as 4 km (Feurtey *et al.*, 2017) and 11 km (Reim *et al.*, 2015). Feurtey *et al.* (2017) reported that

although about 5% of the pollination events occurred between trees more than 1 km apart, 75% occurred at distances of less than 100 m and 25% even below 15 m. Interestingly, pollination distances decreased with the level of *M. domestica* ancestry of the father tree so that trees genetically closer to *M. sylvestris* had higher pollen dispersal distances than more cultivated apple trees (Feurtey *et al.*, 2017). Feurtey *et al.* (2017) also showed that the number of hybridization events increased with decreasing tree density, suggesting that when fewer crab apple trees are close together pollinators have to fly larger distances which presumably increases the chance of encountering domesticated apple trees.

Our results indicate that a large proportion of hybridisation events in the wild were facilitated by *M. sylvestris* pollen as 80% of all detected hybrids are likely to be backcrosses to *M. sylvestris*. This is somewhat unexpected regarding the rare occurrence of F1. One explanation for this could be that cultivated apple trees are rare and therefore do not contribute significantly to the pollen pool. However, this is unlikely as 44% of the samples collected in the Central Belt, where a substantial number of orchards exist, were classified as backcrosses to *M. sylvestris*. An alternative explanation is that a large proportion of the samples might not necessarily be the product of a natural hybridisation event which occurred *in situ* but instead are planted specimens of unknown origin. In the densely populated Central Belt, a landscape which has also been heavily modified by agriculture and human activities, very little natural habitat is left. Indeed, the vast majority of samples was collected along country roads, field borders and hedges. Only a few samples came from semi-natural looking habitats, which also showed clear signs of human influence. It is therefore very likely that the samples collected in these areas were planted and are derived from nursery stock of unknown origin. A recent study investigating the genetic integrity of crab apple trees sold by nurseries reported that all samples sold as wild crab apples from commercial seed companies had substantial levels of *M. domestica* ancestry (Feurtey *et al.* 2017). It is therefore possible that the high frequency of backcrosses to *M. sylvestris* in certain areas of Scotland is not necessarily a sign of natural hybridisation events in these areas but may be partly a reflection of the purity of the crab apple stock sold by nurseries.

It is noteworthy that five out of the 17 (~30%) reference samples included for *M. domestica* had contributions of $P_{syl} \geq 10\%$ from the *M. sylvestris* genome ranging from $13\% < P_{syl} < 32\%$. If collected in the wild, these samples would have been regarded as backcrosses to *M. domestica*. However,

these five individuals belong to well known cultivars ('Red Falstaff', 'Early Julyan', 'Alderman', 'pomme de fer' and 'Corte pendu') which might indicate that the samples identified as backcrosses to *M. domestica* or F1 are in fact *M. domestica* cultivars. Cornille *et al.* (2012) have highlighted that many cultivars including commercial varieties such as Granny Smith can have large admixture proportions from the *M. sylvestris* genepool. For example, the English apple cultivar 'Cox Orange Pippin' has a 55% *M. sylvestris* contribution (Cornille *et al.*, 2012, Table S2) and would therefore be classified as F1 on the basis of admixture proportions. The Cox is the UK's most popular dessert apple and accounts for about 42% of the UK's apple production (<https://quadram.ac.uk/spotlight/apple-facts/>, accessed 10 May 2018). This opens up the possibility of an alternative interpretation where some hybrid samples collected from the wild are not necessarily the product of a recent hybridisation event but could actually be 'escaped' cultivars which have grown from the seed of an apple core carelessly thrown out of a car or discarded at the wayside. Interestingly, about 88% (7/8) of the samples which had a higher than 50% posterior probability to be an F1 in the NEWHYBRIDS analysis were collected either along roads (n=5), on the bank of an abandoned railway line (n=1) or next to a foot path (n=1).

No population genetic structure in pure M. sylvestris populations

The STRUCTURE analysis of a pruned data set where all trees with a domesticated membership coefficient of $P_{Dom} > 10\%$ had been removed did not reveal any population structure. This is also apparent in the very low genetic differentiation between all populations as well as in pairwise population comparisons and the lack of an isolation-by-distance pattern. This suggests that crab apple pollen is effectively dispersed over long distances in Northern Britain which is supported by pollination studies reporting dispersal distances of up to 11 km (Reim *et al.*, 2015) but might also partly be due to human mediated dispersal. The lack of a significant population structure is also consistent with previous studies. Although Cornille *et al.* (2015) detected five different genetic clusters in wild apple on a European scale, population structure within each of these populations or on a smaller regional scale is generally weak or non-existent (Cornille *et al.*, 2013a; Schnitzler *et al.*, 2014; Wagner *et al.*, 2014).

Is morphology a useful indicator to distinguish between pure and hybrid samples?

Only 68% of our field identifications turned out to be correct. This highlights the difficulty in reliably distinguishing between pure wild apple and hybrid samples solely based on morphology. Morphological characters which have been used in the literature to delimit pure *M. sylvestris* are the lack of hairs on the petiole and underside of the leaves, the fruit size (< 30 mm) and leaf size (Wagner, 1996; Stephan *et al.*, 2003). However, it has been shown that there is large morphological variability within species and that hybrid samples display a continuum of intermediate characters (Wagner, 1996). As noted by Cornille *et al.* (2012) even cultivars can have large *M. sylvestris* contributions which will have a compounding effect on the morphology and these discriminating characters. This is clearly echoed by our results as some trees which were identified as crab apple in the field turned out to be domestic apple genetically. In some cases a sample was identified as a domestic apple when it was a crab apple genetically. This suggests that identifications based on morphology need to be treated with caution.

However, there is also reason to be cautious about viewing genetic data as the silver bullet for all apple identification problems. Some of the cultivars assayed by Cornille *et al.* (2012) had very high contributions from *M. sylvestris* - in the case of the Russian cultivar 'borovitsky', which is better known as 'Duchess of Oldenburg' in Britain (Smith, 1971), to the point that it would have been classified as pure *M. sylvestris* genetically but more likely as a hybrid or domesticated apple morphologically due to the larger fruit size ($P_{syl}=0.933$, Table S2 in Cornille *et al.* (2012)).

This clearly highlights that there is no foolproof way to obtain accurate identification of trees either in the field or in the lab alone. It therefore seems that a combination of field assessment and genetic data will provide the most reliable basis for the identification of an unknown sample.

Implications for conservation

There is evidence for a long standing history of apple cultivation in Scotland which can be traced back to the 12th century (Hayes, 2008). Many orchards were founded throughout the centuries with a peak period of fruit growing in the early 20th century (Ironsides Farrar, 2004). This clearly shows that there has been ample opportunity for hybridisation between wild apple and domesticated apple. Wild apple trees have also been planted as alternative pollinators for eating and cooking apples and the fruits have been used for animal fodder as well as cider and jelly making (Cox and Beaton, 2002, p. 101).

These planted trees might have been raised from local stock in the past but nowadays are more likely to be of unknown origin with a high chance of being introgressed by *M. domestica* (Feurtey *et al.*, 2017). It is also worth mentioning that the term 'crab apple' does not necessarily refer to the European wild apple *M. sylvestris* but is generally used for any *Malus* species producing small fruits to distinguish it from the cultivated apple. Nurseries offer a large number of crab apple varieties many of which have nothing to do with the European wild apple *M. sylvestris* and often are of obscure parentage (Fiala and Daniels, 1994). In some instances their origin is well documented as in the case of the popular crab apple *Malus* 'Evereste' which is a hybrid between *Malus domestica* 'Rome Beauty' x *Malus floribunda* clone 821 (Decourtye, 1977). This array of crab apple varieties is confusing for customers and might result in the unintentional planting of exotic instead of native wild apple trees. It is possible that the five samples identified as belonging to the '*M. baccata* group' in the STRUCTURE analysis could be an example for this.

All this highlights the difficulty of assessing how much *M. sylvestris* growing in Scotland can be considered 'wild', i. e. has not been planted. This is certainly true for densely populated lowland areas and other locations where agriculture has had a major impact on the landscape. In the Central Belt and Aberdeenshire for example many of the trees which turned out to be pure *M. sylvestris* grow in places with obvious human presence and interference such as hedges, next to roads, along fences of small semi-natural woodlands, pastures, and field edges. Given the value this species has had for humans it is hard to imagine that these individuals which sometimes seem to be many decades old are the remnants of a natural population. The origin of these trees is probably threefold. Firstly, in rare cases they could be the product of a natural dispersal event; or, more likely, they are planted trees which were either raised from seeds collected from local trees in the countryside (planted or natural) or from a commercial seed orchard.

In contrast, in upland areas such as the Loch Lomond National Park in the Southern Highlands, parts of Dumfries and Galloway and the Lake District, where 75-87% of the collected samples can be considered pure, many of the wild apple trees grow in seemingly natural or semi-natural habitat with little reason to suspect deliberate planting. The component of hybrid trees in these areas is probably increasing only slowly due to the high frequency of pure wild apple trees and the lower frequency of domesticated apples. These populations should therefore be of priority conservation concern in order

to safeguard *M. sylvestris* as a species in Northern Britain. They could be used as sources for the establishment of seed orchards with certified pure *M. sylvestris* trees. This would be a timely conservation measure considering the difficulty to obtain unadulterated material from commercial nurseries (Feurtey *et al.*, 2017). Orchards should have high tree densities and be sited at some distance from known *M. domestica* populations to minimise the potential of hybridisation with cultivated apple (Reim *et al.*, 2015). Stock from such a seed orchard, or from controlled wild collections in the uplands would serve as reliable sources of pure wild apple, which could either be used for re-introductions in areas where the occurrence of truly wild *M. sylvestris* is doubtful or to support regeneration more widely. Despite the prolific fruit production of many crab apple trees we have seen very little in terms of natural regeneration in the field. This is concerning as wild apple trees are apparently not particularly long lived trees (80-100 years, Stephan *et al.*, 2003) and a substantial number seem to be of fairly old age in the wild. Another approach to encourage regeneration is to introduce managed low intensity grazing by cows or horses. It has been shown that cattle and horses are effective dispersers of wild apple seeds, which readily germinate in dung patches and hoof marks. Successful regeneration could then be encouraged by managed grazing in subsequent years (Buttenschön and Buttenschön, 1985, 1999). Deer might also be contributing to the seed dispersal of wild apple trees.

We also encountered a few wild apple trees in forests with no fruit set and poor vitality which was probably due to the disadvantageous light conditions. As *M. sylvestris* has high light requirements, pruning or thinning of nearby trees might be an inexpensive and effective measure to improve the local light condition for a particular crab apple tree resulting in increased vitality and flowering (Stephan *et al.*, 2003).

Conclusion

Our data suggest that there is a significant amount of hybridisation between crab and domesticated apple in Northern Britain. Given the use of wild apple trees and their association with man-made habitats it seems likely that many of the hybrid (and pure) samples collected from lowland agricultural areas are actually planted trees and not the product of a natural hybridisation event with subsequent seed dispersal and establishment. The overall hybridisation rate reported in this study is therefore likely to be the result of natural processes as well as human activities, and is probably be an

488 overestimation. In areas with a high frequency of pure wild apple trees the component of hybrid trees
489 is probably increasing only slowly, and basic conservation measures should be effective in
490 safeguarding *M. sylvestris* as a species in the long term. To estimate levels of natural hybridisation
491 between wild and domesticated apple in the wild, genotyping of seedlings collected from pure *M.*
492 *sylvestris* trees in different habitats is required. This would allow the potential threat of introgression to
493 be evaluated more accurately.

494 **Funding**

495 This work was supported by the Patsy Wood Trust; the Scottish Forestry Trust and Sustainable
496 Forestry SCIO; the Falkland Stewardship Trust; and the Forestry Commission Scotland via the Native
497 Woodland Survey of Scotland. The Royal Botanic Garden Edinburgh is supported by the Scottish
498 Government's Rural and Environment Science and Analytical Services Division.

499 **Acknowledgements**

500 Special thanks go to Robin Buxton, Colin Hughes, Bob Stubbs and Andrew Fairbairn. We are grateful
501 to the many people who located trees for us or sent in leaves for sampling; especially Mary-Ann
502 Smythe, Richard Cunningham, Crystal Maw, Neil Mackenzie, Joyce Gilbert, Iain Brodie, Paul Harvey
503 and John Parrot. We also thank the editor and two anonymous reviewers for their comments which
504 improved an earlier version of this manuscript.

505 **Conflicts of interest**

506 None declared

507

References

- Anderson EC, Thompson EA (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229.
- BLE (2013). *BUNDESANSTALT FÜR LANDWIRTSCHAFT UND ERNÄHRUNG: Erfassung und Dokumentation genetischer Ressourcen seltener und gefährdeter Baumarten in Deutschland. Teillos 2: Wild-Apfel (Malus sylvestris) und Wild-Birne (Pyrus pyraaster).*
- Broothaerts W (2003). New findings in apple S-genotype analysis resolve previous confusion and request the re-numbering of some S-alleles. *Theor. Appl. Genet.* **106**, 703–714.
- Buttenschøn J, Buttenschøn RM (1985). Grazing experiments with cattle and sheep on nutrient poor, acidic grassland and heath, 4: Establishment of woody species. *Nat. Jutl.* **21**, 117–140.
- Buttenschøn RM, Buttenschøn J (1999). Population dynamics of *Malus sylvestris* stands in grazed and ungrazed , semi-natural grasslands and fragmented woodlands in Mols Bjerger , Denmark. *Ann. Bot. Fenn.* **35**, 233–246.
- Chhatre VE, Emerson KJ (2017). StrAuto: automation and parallelization of STRUCTURE analysis. *BMC Bioinformatics* **18**, 192.
- Coart E, VAN Glabeke S, DE Loose M, Larsen AS, Roldán-Ruiz I (2006). Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Mol. Ecol.* **15**, 2171–82.
- Coart E, Vekemans X, Smulders MJM, Wagner I, Van Huylenbroeck J, Van Bockstaele E, *et al.* (2003). Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Mol. Ecol.* **12**, 845–857.
- Cornille A, Feurtey A, Gélin U, Ropars J, Misvanderbrugge K, Gladieux P, *et al.* (2015). Anthropogenic and natural drivers of gene flow in a temperate wild fruit tree: A basis for conservation and breeding programs in apples. *Evol. Appl.* **8**, 373–384.
- Cornille A, Giraud T, Bellard C, Tellier A, Le Cam B, Smulders MJM, *et al.* (2013a). Postglacial recolonization history of the European crabapple (*Malus sylvestris* Mill.), a wild contributor to the domesticated apple. *Mol. Ecol.* **22**, 2249–2263.

538 Cornille A, Giraud T, Smulders MJM, Roldán-Ruiz I, Gladieux P, Juniper BE, *et al.* (2014). The
539 domestication and evolutionary ecology of apples. *Trends Genet.* **30**, 57–65.

540 Cornille A, Gladieux P, Giraud T (2013b). Crop-to-wild gene flow and spatial genetic structure in the
541 closest wild relatives of the cultivated apple. *Evol. Appl.* **6**, 737–748.

542 Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F, Le Cam B, *et al.* (2012). New
543 Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild
544 Apple to the Genome of Cultivated Varieties. *PLoS Genet.* **8**, e1002703.

545 Cox K, Beaton C (2002). *Fruit and Vegetables for Scotland*. Birlinn Ltd, 298 pp.

546 Decourtye L (1977). ‘Evereste’, un nouveau pommier d’ornement. *L’Hortic Fr.* **54**, 309–311.

547 Duan N, Bai Y, Sun H, Wang N, Ma Y, Li M, *et al.* (2017). Genome re-sequencing reveals the history
548 of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* **8**, 249.

549 Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the
550 software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620.

551 Feurtey A, Cornille A, Shykoff JA, Snirc A, Giraud T (2017). Crop-to-wild gene flow and its fitness
552 consequences for a wild fruit tree: towards a comprehensive conservation strategy of the wild
553 apple in Europe. *Evol. Appl.* **10**, 180–188.

554 Fiala JL, Daniels GS (1994). *Flowering Crabapples: The Genus Malus*. Timber Press, 273 pp.

555 Francis RM (2017). pophelper: an R package and web app to analyse and visualize population
556 structure. *Mol. Ecol. Resour.* **17**, 27–32.

557 Graudal L, Kjær ED, Canger S (1995). A systematic approach to the conservation of genetic
558 resources of trees and shrubs in Denmark. *For. Ecol. Manage.* **73**, 117–134.

559 Hayes CW (2008). Ancient Orchards on the Banks of the River Tay. *Landsc. Archaeol. Ecol.* **7**, 63–
560 75.

561 Hayes CW (2010). *Survey of Traditional Orchards in the Forth Valley 2009*. 1-42 pp.

562 Henning W (1947). Morphologisch-systematische und genetische Untersuchungen an Arten und
563 Artbastarden der Gattung Malus. *Der Züchter* **17**, 289–349.

564 Ironside Farrar (2004). *A Clyde Valley orchards survey*. Scottish Natural Heritage Commissioned
565 Report No. 023 (ROAME No. FO2LI21).

566 Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA (2013). diveRsity : An R package for the
567 estimation and exploration of population genetics parameters and their associated errors.

568 *Methods Ecol. Evol.* **4**, 782–788.

569 Kik C, Korpelainen H, Vögel R, Asdal Å, Eliáš P, Draper D, *et al.* (2011). *Malus sylvestris*. The IUCN
570 Red List of Threatened Species 2011. *IUCN*, e.T172170A,
571 <http://dx.doi.org/10.2305/IUCN.UK.2011>.

572 Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015). Clumpak: a program for
573 identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol.*
574 *Resour.* **15**, 1179–1191.

575 Larsen A, Asmussen CB, Coart E, Olrik DC, Kjær ED (2006). Hybridization and genetic variation in
576 Danish populations of European crab apple (*Malus sylvestris*). *Tree Genet. Genomes* **2**, 86–97.

577 Larsen AS, Kjær ED (2009). Pollen mediated gene flow in a native population of *Malus sylvestris* and
578 its implications for contemporary gene conservation management. *Conserv. Genet.* **10**, 1637–
579 1646.

580 Palmer-Jones T, Clinch PG (1966). Observations on the pollination of apple trees (*Malus sylvestris*
581 Mill.). *New Zeal. J. Agric. Res.* **9**, 191–196.

582 Peakall R, Smouse PE (2012). GenALEX 6.5: Genetic analysis in Excel. Population genetic software
583 for teaching and research-an update. *Bioinformatics* **28**, 2537–2539.

584 Pritchard JK, Stephens M, Donnelly P (2000). Inference of Population Structure Using Multilocus
585 Genotype Data. *Genetics* **155**, 945–959.

586 Reim S, Proft A, Heinz S, Lochschmidt F, Höfer M, Tröber U, *et al.* (2015). Pollen movement in a
587 *Malus sylvestris* population and conclusions for conservation measures. *Plant Genet. Resour.*
588 *Characterisation Util.* **15**, 12–20.

589 Remmy K, Gruber F (1993). Untersuchungen zur Verbreitung und Morphologie des Wild-Apfels
590 (*Malus sylvestris* (L.) Mill.). *Mitt. Deutsch. Dendrol. Ges.* 71–94.

591 Schnitzler A, Arnold C, Cornille A, Bachmann O, Schnitzler C (2014). Wild European apple (*Malus*
592 *sylvestris* (L.) Mill.) Population dynamics: Insight from genetics and ecology in the Rhine Valley.
593 Priorities for a future conservation programme. *PLoS One* **9**, 1–11.

594 Smith MWG (1971). *National Apple Register of the United Kingdom*. Ministry of Agriculture, Fisheries
595 and Food, 652 pp.

596 Stace C (2010). *New flora of the British Isles*. Cambridge University Press, 1266 pp.

597 Starr TN, Gadek KE, Yoder JB, Flatz R, Smith CI (2013). Asymmetric hybridization and gene flow

598 between Joshua trees (Agavaceae: *Yucca*) reflect differences in pollinator host specificity. *Mol.*
599 *Ecol.* **22**, 437–449.

600 Stephan R, Wagner I, Kleinschmit J (2003). *Technical Guide Lines for Genetic Conservation and Use*
601 *for Wild Apple and Pear (Malus sylvestris and Pyrus pyraaster).*

602 Tollefsrud MM, Sønstebo JH, Arvid P (2014). *Registrering og genetisk karakterisering av villeple i*
603 *Norge*. Skog og landskap report 18; 102 pp.

604 Vähä JP, Primmer CR (2006). Efficiency of model-based Bayesian methods for detecting hybrid
605 individuals under different hybridization scenarios and with different numbers of loci. *Mol. Ecol.*
606 **15**, 63–72.

607 Wagner I (1996). Zusammenstellung morphologischer Merkmale und ihrer Ausprägung zur
608 Unterscheidung von Wild- und Kulturformen des Apfels. *Mitteilungen der Dtsch. Dendrol.*
609 *Gesellschaft* **82**, 87–108.

610 Wagner I, Maurer WD, Lemmen P, Schmitt HP, Wagner M, Binder M, *et al.* (2014). Hybridization and
611 Genetic Diversity in Wild Apple (*Malus sylvestris* (L.) MILL) from Various Regions in Germany
612 and from Luxembourg. *Silvae Genet.* **3**, 81–94.

613 Wang J (2017). The computer program STRUCTURE for assigning individuals to populations: Easy to
614 use but easier to misuse. *Mol. Ecol. Resour.* **17**, 981–990.

615

616

Table and Figure captions

Figure 1. Sampling locations and broad geographical regions of 332 *Malus* trees collected in the field in Northern Britain. Lake District (LD, n=24), Dumfries (DF, n=89), Central Belt (CB, n=66), Southern Highlands (n=122), West Coast (WC, n=9), Aberdeenshire (AB, n=9) and Northern Highlands (NH, n=13, including one sample from Shetland).

Figure 2. Barplot of STRUCTURE membership coefficients inferred with 14 nSSRs for all assayed *Malus* trees (n=366, six species) for K=3, the most likely number of genetic clusters according to the method by Pritchard *et al.* (2000). Red (dark grey in print version) represents the *M. domestica* cluster, blue (light grey in print version) the *M. sylvestris* cluster and black the '*M. baccata* - *M. hupehensis* - *M. orientalis*' cluster. Grey header boxes highlight the included reference *Malus* species (Domestica - *M. domestica*; Baccata - *M. baccata*; Hu - *M. hupehensis*; Or - *M. orientalis*; Si - *M. sieversii*; Syl - *M. sylvestris*) and sampling locations.

Figure 3. Principal Coordinate Analysis (PCO) of all assayed *Malus* samples using 14 nSSRs (n=366). Individuals are coloured according to the K=3 STRUCTURE groupings. '*M. domestica* group'...*M. domestica* reference samples plus individuals with genomic *M. domestica* membership coefficient ≥ 0.9 . '*M. baccata*' group'...samples with *M. baccata*/*M. hupehensis*/*M. orientalis* membership coefficient ≥ 0.9 . '*M. sylvestris* group'...samples with *M. sylvestris* membership coefficient ≥ 0.9 . 'Hybrid'...introgressed *M. sylvestris* samples with $0.1 < P_{\text{dom}} < 0.9$.

Figure 4. A. Proportion of field collected samples which were either pure *Malus sylvestris* (proportion of *M. sylvestris* genepool $P_{\text{syl}} \geq 0.9$, blue; light grey in print version) or of hybrid origin ($P_{\text{syl}} < 0.9$, red; dark grey in print version) using STRUCTURE results for K=3. **B.** Mean proportion of *M. sylvestris* (blue) and cumulative proportions of *M. domestica* and '*M. baccata* group' genetic contributions ('cultivated genome'; red) in hybrid samples. LD - Lake District, DF - Dumfries, CB - Central Belt, SH - Southern Highlands, WC - West Coast, AB - Aberdeenshire and NH - Northern Highlands. 'n' in (A) denotes the total number of samples excluding pure *M. domestica* samples (proportion of *M. domestica* genepool ≥ 0.9) and in (B) the total number of hybrid samples ($P_{\text{syl}} < 0.9$).

Table 1. Introgression rates of field collected *Malus* trees in Scotland and Northern England inferred using STRUCTURE and NEWHYBRIDS. For the STRUCTURE results three different thresholds for

the contributions from other than parental groups have been applied to call a sample pure ($P \leq 0.10/0.15/0.20$; upper, middle and lower row, respectively). NEWHYBRIDS assignments are based on a posterior probability $\geq 55\%$ of belonging to a certain class. ' P_{dom} ' denotes the combined contributions from the *M. domestica* and *M. hupehensis* - *M. baccata* - *M. orientalis* group; ' P_{dom} ' is the contribution from the *M. domestica* group; 'Syl' pure *M. sylvestris*; 'BX-Syl' backcross to *M. sylvestris*; F1/F2 first/second generation hybrid, respectively; 'BX-Dom' backcross to *M. domestica*; 'Dom' pure *M. domestica*. 'Syl/BX-Syl', 'F1/BX-Syl' and 'F2/BX-Syl' denote the percentage of samples in the NEWHYBRIDS output which had more or less equal chances of belonging to either of the two categories.

Table 2. Genetic diversity estimates of pure *M. sylvestris* populations with more than ten pure individuals (based on the STRUCTURE results for $K=3$ and a $P_{dom} \leq 0.1$ threshold for pure individuals). 'n' number of samples; ' H_o ' and ' H_e ' observed and expected heterozygosity, respectively; ' F_{is} ' inbreeding coefficient; ' A_r ' allelic richness estimated by rarefaction using the lowest sample size of $n=17$.

Table 3. Reliability of morphological field assessments judged against genetic data using a threshold of $P_{dom} \leq 0.1$ for pure *M. sylvestris* and $P_{dom} > 0.1$ for hybrid samples.

662 **SUPPLEMENTARY MATERIAL**

663 **Supplementary Figure S1.** Statistics of the STRUCTURE analysis over 20 runs for each K=2 to 10.

664 (A) Mean estimated ln probability L(K) of data (mean $\pm$ SD) (B) Rate of change of the likelihood
665 distribution (mean $\pm$ SD) (C) ΔK method for most likely K according to Evanno (2005) (D) Pr[X|K]
666 method for most likely K according to Pritchard (2000).

667 **Supplementary Figure S2.** STRUCTURE results for K=2 to 10 using the full dataset (n=366). Only
668 major modes are shown. Number of minor modes in square brackets. Reference samples of the
669 following species: 1...*M. baccata*, 2...*M. hupehensis*, 3...*M. orientalis*, 4...*M. sieversii*, 5...*M.*
670 *sylvestris*

671 **Supplementary Figure S3.** STRUCTURE results for K=2 to 6 using only pure *M. sylvestris* samples
672 (n=227, $P_{\text{dom}} \leq 0.1$). Only major modes are shown. Number of minor modes in square brackets.
673