Long-term reciprocal gene flow in wild and domestic geese reveals complex domestication history

Marja E. Heikkinen[1,2], Minna Ruokonen[1§], Thomas A. White[1,3], Michelle M. Alexander[4], İslam Gündüz[5], Keith M. Dobney[6,7,8], Jouni Aspi[1], Jeremy B. Searle[2]*, Tanja Pyhäjärvi[1,9]*

§ deceased

* contributed equally to this paper

[1] Department of Ecology and Genetics, PO Box 3000, Fi-90014 University of Oulu, Finland

[2] Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA

[3] CMPG, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[4] University of York, BioArCh, Environment Building, Wentworth Way, Heslington, York, YO10 5NG

[5] Department of Biology, Faculty of Arts and Sciences, University of Ondokuz Mayis, Samsun, Turkey

[6] Department of Archaeology, Classics and Egyptology, University of Liverpool, 12–14 Abercromby Square, Liverpool L69 7WZ, UK

[7] Department of Archaeology, University of Aberdeen, St Mary's, Elphinstone Road, Aberdeen, AB24 3UF, UK

[8] Department of Archaeology, Simon Fraser University, Burnaby, B.C. V5A 1S6, 778-782-419, Canada

[9] Biocenter Oulu, PO Box 5000, Fi-90014 University of Oulu, Finland

Author for correspondence:

Marja E. Heikkinen

email: marja.e.heikkinen@oulu.fi

ORCID:

MEH, 0000-0002-7051-5654; MMA, 0000-0001-8000-3639; İG, 0000-0002-6436-5397; KMD, 0000-0001-9036-4681; JA, 0000-0002-2451-3201; JBS, 0000-0001-7710-5204; TP, 0000-0001-6958-5172

Abstract

Hybridisation can substantially impact genetic diversity of domesticates, but this situation has not been widely studied. We analysed the impact of hybridisation on the genome-wide genetic variation in current populations of the European domestic goose and its wild progenitor: the greylag goose (*Anser anser*). Our dataset consisted of 58 wild greylag and 75 domestic geese genotyped for 33,527 single nucleotide polymorphisms (SNPs) identified by genotyping-by-sequencing (GBS). Demographic reconstruction and clustering analysis suggested that divergence between wild and domestic geese around 5,300 generations ago was followed by long-term genetic exchange, and that greylag populations have 3.2–58.0% admixture proportions with domestic geese, that also varies geographically. Many modern European breeds share considerable (> 10%) ancestry with Chinese domestic geese (derived from the swan goose *A. cygnoid*). This study shows a domestication process progressing despite continued and pervasive gene flow from the wild form.

Keywords: *Anser anser*, domestication, domestic goose, genotyping-by-sequencing, hybridisation, population genomics

1. Introduction

Reproductive isolation is a defining feature of speciation and yet hybridisation between species is an important general phenomenon in trait evolution [1,2]. Among birds, the Anseriformes (ducks, geese and swans) show particularly pervasive hybridisation, 41.6% to > 60% of species hybridising with each other [3,4]. Domestication generates differentiated gene pools and reproductive isolation between domestics and their wild progenitor, but hybridisation between domestic and wild forms has been well demonstrated in both plants [1,5] and animals [6,7]. The impacts include genetic and trait enrichment of domestics, for instance, in chicken the acquisition of a yellow skin phenotype is a result of past mating between red junglefowl and grey junglefowl [8]. In geese, the high tendency for hybridisation among wild species is also shown among wild and domestic forms [9,10], creating an exciting opportunity to study the complex dynamics of hybridisation and domestication.

The domestic geese of the world (European and Chinese forms) are derived from two different wild species: the greylag (*Anser anser*) and the swan goose (*A. cygnoid*) [11,12]. *A. anser* and *A. cygnoid* shared a common ancestor about 3.4 Mya [13], but are still able to hybridise [4], and some domestic breeds are reportedly hybrid [14]. The greylag has been divided into the western, nominate subspecies *A. a. anser* (Linnaeus, 1758) with a European breeding range and the eastern subspecies *A. a. rubrirostris* (Swinhoe, 1871) breeding further east, although the subspecific boundary is not well defined, and mitochondrial DNA has not been found to distinguish them [10]. Of these subspecies, *rubrirostris* is larger and lighter coloured than *anser* [15] and has a pink bill and cold pink legs in contrast to the orange bill and flesh-coloured legs of *anser*, the bill colour used as primary evidence in favour of the original domestication of *rubrirostris* [16]. As with all domesticates, domestic geese varieties are morphologically more diverse than their wild counterparts, particularly in plumage and body size [14].

The current knowledge about goose domestication relies largely on ancient texts and archaeological evidence. Questions about where and when domestication took place, the genetic changes associated with it and the later history of domestic geese, however, remain largely unresolved [10]. There are depictions from the New Kingdom of Egypt that suggest geese were already fully domesticated by the 18th Dynasty (1450-1341 BCE). The earliest reliable reference to domestic geese in western Eurasia is Homer's Odyssey (first half of 8th century BCE) and geese were certainly well-established poultry by Roman times [17].

Genetic diversity in the mitochondrial DNA (mtDNA) of greylag and European domestic geese showed reduced diversity in the domestics [10] which may result from an early domestication bottleneck or, alternatively, later breed formation. There is a particular mitochondrial haplogroup common in the domestics [10], and archaeological domestic goose bones from the High Medieval (11th-13th century CE) of Russia belonged to that haplogroup [18].

77 MtDNA relationships between extant Chinese and European domestic goose breeds confirm that the
78 former, excluding one breed, have swan goose ancestry, whereas European domestic goose and the
79 Chinese Yili breed have greylag ancestry [12,19,20]. However, Chinese mtDNA haplotypes may occasionally
80 occur in European domestics, and vice versa [10,19].
81 Genomic data can be much more powerful than mtDNA in hybridisation inference, and has been
82 applied, for instance, to New World cattle, which along with their taurine ancestry were shown to have a
83 greater proportion of indicine ancestry than previously assumed [21]. Interpretation is still challenging and
84 it is best to infer jointly the genetic impact of initial domestication and subsequent hybridisation of wild and
85 domestic populations, as the latter can obscure domestic-wild genetic relationships and may also give a
86 false impression of the number of times a species has been domesticated [22–24].
87 Here we investigate goose domestication history using genome-wide single nucleotide polymorphism
88 (SNP) data from thousands of loci, obtained by genotyping-by-sequencing (GBS). We used 56 and 50
89 samples of greylag and domestic geese from a previous mtDNA study [10], together with 2 new Turkish
90 greylag and 25 new domestic specimens. We studied the interplay between domestication and
91 hybridisation by addressing the following questions: i) what is the extent of genetic differentiation amongst
92 wild and domestic geese? ii) what is the approximate time of domestication? and iii) what is the role of
93 intra- and interspecific hybridisation in goose domestication history and the current genetic makeup?

94 **2. Materials and methods**

95 a) Sampling

96 The greylag wild-collected samples derive widely from Eurasia (Figure 1, Table S1). The European domestic
97 goose samples represented 14 different breeds (Table S1) and individuals that did not belong to any
98 recognised breed or were presumed to be crosses between European and Chinese domestic geese. Some
99 specimens were reported to be Chinese domestic geese. The domestic samples were obtained from local
100 breeders in Denmark, Sweden and the UK and those from Turkey were collected directly by the authors.

101 b) DNA extraction and GBS library construction

102 GBS [25] libraries were constructed at the Cornell Biotechnology Resource Center (BRC) following DNA
103 extraction with the DNeasy Blood and Tissue Kit (QIAGEN) with RNase treatment. Each individual DNA
104 sample and an adaptor with a unique barcode were combined in a 96-well plate along with a common
105 adaptor. Samples were treated with the EcoT-22I (ATGCAT) restriction enzyme to create fragmented DNA.
106 Barcoded adapters and common adapters with matching sticky ends were ligated to each sample with T4
107 DNA ligase. The samples were pooled and purified with a QIAquick PCR Purification Kit (QIAGEN). PCR

amplification of the library used primers complementary to barcoded and common adapters with products purified as above, and the samples were 100 bp SE-sequenced with Illumina HiSeq 2000/2500 at the BRC.

c) GBS pipeline and SNP calling

Raw sequence reads were run through the Command Line Interface of the Tassel 5 GBS v2 Discovery and Production pipelines [26]. Details about the pipelines and SNP calling are in the Electronic Supplementary Material, ESM. Good quality reads were recorded as tags and aligned to the *A. cygnoid domesticus* GenBank assembly (AnsCyg_PRJNA183603_v1.0 GCF_000971095.1) [27] using the Burrows-Wheeler Aligner with default settings [28]. After running the raw data through the pipelines, 69,865 SNPs were obtained.

The SNPs were subjected to additional filtering using VCFtools [29]. We removed indels, loci with more than two alleles and invariant loci. However, invariant loci were retained for phylogenetics, informing about greylag-swan goose divergence. After preliminary analyses loci with observed heterozygosity over 0.75 were removed as potential paralogs. Individuals with more than 20% missing data across loci were removed. The final dataset consisted of 33,527 biallelic SNPs and 133 individuals (58 wild and 75 domestic).

d) Population structure analyses

Population clustering and structure was analysed with STRUCTURE 2.3.4 [30] and Principal Component Analysis (PCA) [31]. For the whole dataset, STRUCTURE was run with 1000 burn-in steps followed by 10 000 iterations of MCMC for data collection for K = 1-10 allowing admixture with five replicates of each run to reach convergence. For the STRUCTURE analyses done separately on greylags and European domestic geese, see ESM. An admixture model with correlated allele frequencies among populations [32] was used in all STRUCTURE analyses and the iterations were automated with StrAuto 1.0 [33]. We applied both likelihood of K and ΔK [34] of successive K values to determine the optimal number of clusters, using STRUCTURE HARVESTER [35]. CLUMPP 1.1.2 [36] was used to align the assignments from different replicates of K and DISTRUCT 1.1 [37] for visualisation. A PCA was performed with the prcomp function in R [38] and the significance of eigenvalues determined based on the Tracy-Widom distribution [22,31].

A neighbor-joining tree was constructed for phylogenetic analysis, with pairwise distance between individuals obtained with the R package ape [39] based in 40,191 loci. The *A. cygnoid* reference genome and the invariant sites that differed from it were included in the tree construction.

e) Tests for admixture and simulations of demographic history

The history of admixture was tested with a 3-Population test f_3(C; A, B) implemented in AdmixTools 4.1 [40, ESM]. This method offers a formal test to explain observed patterns of admixture in a target population without an outgroup. For identification of admixture between Chinese and European domestics, Grey and White Chinese were combined to represent the Chinese, and the Landes breed that had minimum indication of admixture in STRUCTURE was chosen to represent the European domestic source population.

141 Different models of demographic history were tested with fastsimcoal2 ver 2.6 [41]. We excluded all
142 SNPs that had missing data within the whole data set and executed the analyses with a site frequency
143 spectrum (SFS) based on 9,212 SNPs. As there are no estimates of the genetic diversity per base pair for
144 greylags, we estimated the proportions of variable and monomorphic sites in the data. From the BAM file
145 with –depth option in SAMtools 1.7 [42], we estimated 9,801,382 bp covered with GBS tags. We then
146 mimicked the filtering steps done for the biallelic SNPs to reduce the total number of sites in equivalent
147 proportions. We removed the same number of sites that corresponded to the number of SNPs that were
148 removed because they were indels, had more than 2 alleles or had heterozygosity over 0.75. Since some of
149 the SNPs were removed from this analysis due to missing data in some individuals, we removed an equal
150 proportion of sites from the total number of sites as well. The final SFS had 1,681,316 sites.
151 To infer the demographic history, we chose a subset of individuals from both wild-collected greylags
152 and domestic geese to represent the genetic variation in both groups. Therefore, 11 greylags with > 90.8%
153 of greylag ancestry and 15 domestic geese with > 91.4% of European domestic goose ancestry were
154 selected for the analysis. The mutation rate for the simulations was 1.38×10^{-7} [43]. The parameter
155 estimation for each model tested involved 100,000 simulations and 40 conditional maximization (ECM)
156 cycles. The parameters for each model were estimated with 100 independent runs to obtain the global
157 maximum. The models tested were i) simple divergence of two populations with no migration, ii)
158 divergence of two populations with migration and iii) divergence of two populations with two different
159 migration matrices (Figure 2). The best model was selected based on Akaike's weight of evidence as in [41].
160 For parametric bootstrapping 100 SFS were simulated with the parameter estimates obtained from the real
161 SFS, followed by maximum likelihood estimation with 50 independent runs for each bootstrap SFS. The 95%
162 confidence intervals were obtained from the bootstrap data for each estimated parameter.

163 f) The estimation of genetic diversity

164 Genetic diversity and pairwise F_{ST} values were investigated with the *hierfstat* R package [44]. Expected
165 heterozygosity (H_E) was calculated for each locus and population and averaged across loci. Difference in
166 average H_E between greylags and European domestics was tested with a two-sample t-test with the Welch
167 correction for non-homogeneity of variance [45]. For comparing the genetic diversity among wild and
168 domestics, only pure greylag populations (< 10% admixture with domestic geese) and pure European
169 domestic geese (< 10% admixture with Chinese domestic geese) were used to avoid hybridisation effects on
170 the estimates. The admixture proportions were obtained from STRUCTURE.
171 The variance components across loci for hierarchical F-statistics for pure greylags and pure European
172 domestics were estimated using locus-by-locus analysis of molecular variance (AMOVA) implemented in
173 Arlequin 3.5.2.1 [46]. The significance was tested with 16 000 permutations.

174 ***3.* Results**

175 a) Population structure

176 There was clear divergence between greylags and both European and Chinese domestic geese according to
177 STRUCTURE and PCA (Figure 3a-b). At *K* = 2, populations/breeds are clustered based on their status (wild or
178 domestic) and, at *K* = 3, domestic geese are further separated into European and Chinese. At *K* = 4, the
179 fourth cluster is admixed within greylag populations but none of the individuals are unanimously assigned
180 to that cluster. The likelihood was highest for *K* = 3. These results were supported by PCA as the first two
181 PCs out of 14 significant PCs ($p < 0.05$) were enough to separate the three groups (wild, European domestic,
182 Chinese domestic) from each other (Figure 3a). Overall the greylag populations showed 3.2% - 23.5%
183 admixture proportions with European domestic geese when *K* = 3. In contrast, not all European domestic
184 geese showed admixture with greylags and the admixture percentages ranged from 0.0 to 8.4%. At *K* = 3
185 many European domestic goose breeds showed mixed ancestry with Chinese domestic geese (0.0-27.1%).
186 The neighbor-joining tree repeated the major patterns observed with STRUCTURE and PCA revealing
187 a star shaped phylogeny and confirming that the domestic and greylag geese largely form different clades
188 (ESM, Figure S1). Surprisingly, the Chinese domestic geese were closer to European domestic geese and
189 greylags, than to the swan goose reference genome. In addition, one greylag from Turkey was more closely
190 related to the Chinese domestic geese than other greylags, also indicated by admixture proportions from
191 STRUCTURE. Further, two Crested Faroese individuals and four domestics from the USA (2 unknown and 2
192 Toulouse crosses) were closer to Chinese than European domestic geese. These six individuals also showed
193 high proportions of admixture with Chinese domestics in the STRUCTURE analysis.
194 Some further population structure was observed within both greylags and domestic geese, when
195 analysed separately with STRUCTURE and PCA. Geographically, greylags differentiated by subspecies (ESM,
196 Figures S2-S3). STRUCTURE indicated little differentiation among European domestic geese, but the PCA
197 revealed separation between the European breeds and the Turkish domestic geese (ESM, Figures S4-S5).

198 b) Admixture and the time of domestication

199 Although, STRUCTURE implied considerable mixed ancestry from multiple genetic clusters for Dutch and
200 Turkish greylags, the f_3 analysis did not confirm admixture for these populations. The two Turkish greylag
201 samples came from the same area as our NW-Turkish domestic population, which among Turkish domestic
202 geese showed highest admixture with greylags (2.2%), but admixture was not confirmed with the f_3 test.
203 The f_3 analysis confirmed admixture of domestic geese in line with STRUCTURE results. Most notably,
204 the African breed is a hybrid between European and Chinese domestic geese (Z-score -6.399), unexpected
205 as this breed has been assumed to have originated solely from swan goose. The European-Chinese hybrid
206 status of the Kholmogor and Steinbacher breeds was also confirmed (Z-scores of -8.933 and -5.349,

respectively). The Kholmogor breed also fell halfway between European and domestic geese both in STRUCTURE and PCA, whereas the Steinbacher was genetically closer to European domestic geese in the PCA. However, the Diepholzer breed, which reportedly is also a hybrid, was not confirmed as such in our analysis. Other domestic breeds/groups with admixture status in STRUCTURE were also confirmed to have a European-Chinese admixture when a Z-score threshold of -3 (roughly corresponding to $p < 0.01$) was used: Sebastopol, Toulouse cross, Domestic NY, Embden, Tufted Roman (Table S2, Figure 3c). The Crested Faroese breed gave indication of admixture based on STRUCTURE analysis and the f_3 test supported this (Z-score of -2.228, $p < 0.05$). Surprisingly, the Northern Turkish domestic population was not admixed with Chinese domestic geese in STRUCTURE but f_3 analysis gave a contrasting signal (Z-score -2.459, $p < 0.05$).

The demographic model that best fit our data suggested divergence of greylag and domestic geese with a recent migration rate change (Table 1, Table S3). The model suggested divergence around 5319 generations ago (95% confidence intervals (CI): 1538-8225) with asymmetric but close to equal migration rates from greylags to domestic geese following divergence. About 159 (79-462) generations ago, the migration matrix changed, suggesting higher gene flow from domestic geese to greylags towards modern times. Given an estimated generation time for these geese of about 3 years, this suggests divergence about 14 000 BCE and gene flow shift about 480 years ago.

c) Genetic diversity

An AMOVA was used to partition genetic diversity among greylag vs. domestic (group level), and among populations (greylag) and among breeds (domestics), and within population levels (Table 2). The fixation index between greylag and domestic geese was 0.158 and there was also significant differentiation among greylag populations/domestic breeds. The average pairwise F_{ST} between greylag populations and domestic breeds was 0.197, among greylag populations 0.088 and among domestic breeds 0.174 (Table S4).

The genetic diversity measured as average H_E was higher in pure greylags (0.146) than pure European domestic geese (0.096) (Welch's t-test, degrees of freedom (df) = 10.594, $p = 3.91\times10^{-5}$). The difference in average H_E remained when non-pure greylag and non-pure European domestics were included in the comparison (0.156 vs. 0.107; Welch's t-test, df = 19.28, $p = 0.000418$) (Table S1). The Steinbacher and African breeds (Table S1) were excluded from both estimates because the Steinbacher is known to be a European-Chinese cross and the African has swan goose ancestry. The average H_E ranged from 0.140 (Denmark) to 0.150 (Kazakhstan) in pure greylags and from 0.047 (Landes) to 0.123 (Domestic N-Turkey) in pure European domestics. The average H_E was higher in greylag populations that showed high admixture proportions with domestic geese, Netherlands (0.157, 13.8% European domestic, 2.2% Chinese domestic) and Turkey (0.236, 23.5% European domestic, 34.5% Chinese domestic), but not significantly so (Welch's t-test, df = 1.002, $p = 0.421$, average H_E: pure greylags 0.146 vs non-pure greylags 0.196). The average H_E was also higher in European domestics that showed high admixture proportions with Chinese domestics, i.e.

241 Crested Faroese (0.117, 17% Chinese domestic), Sebastopol (0.133, 11.1% Chinese domestic), Domestic NY
242 (0.143, 78.6% European domestic, 21.4% Chinese domestic) and Toulouse cross (0.151, 72.9% European
243 domestic, 27.1% Chinese domestic), and the difference was also statistically significant (Welch's t-test, df =
244 9.0991, p = 0.0039, average H_E: pure European domestics 0.096 vs non-pure European domestics 0.136).

245 **4. Discussion**

246 We studied the dynamics of domestication and hybridisation in grey (*Anser*) geese using genome-wide SNP
247 data. The results demonstrated genetic divergence between Eurasian wild greylag and European domestic
248 geese with long-term genetic exchange between them. We also inferred temporal changes in the direction
249 of gene flow. The degree of hybridisation between greylag and domestic geese also varied geographically.
250 Surprisingly, several domestic goose breeds also showed a substantial genetic contribution of Chinese
251 domestic geese. We also provide insights about the origin and the timing of goose domestication.

252 a) Genetic diversity and differentiation of greylag and European domestic geese

253 Domestic species often show reduced genetic diversity compared to their wild ancestor, attributable to
254 genetic drift during population bottlenecks of initial domestication, combined with subsequent artificial
255 selection associated with breed formation [47]. Domestic geese appear to follow the same trend. We found
256 European domestic geese to have lower H_E than wild greylags.
257 European domestic geese are genetically distinct from their wild progenitor but no more so than for
258 other domestic birds. The average pairwise F_{ST} values between greylag populations and domestic goose
259 breeds were lower than between red junglefowl and domestic chicken populations [48], and domestic
260 geese are less distinctive than domestic pigeons [49]. Among domestic geese, the Turkish are particularly
261 interesting. From mtDNA, the Turkish domestic geese stand out as the most genetically variable group [10],
262 and although this is less evident from GBS, among the pure European domestic geese the Northern Turkish
263 showed the highest average H_E. The f_3 analysis indicates a history of admixture with Chinese domestics for
264 this population, which may explain its high genetic diversity.
265 We found a genetic separation between European and Near Eastern populations of greylags that
266 aligned with the western and eastern subspecies (*A. a. anser* and *A. a. rubrirostris*) [50], a distinction which
267 could not be made based on mtDNA [10]. Hybridisation between the western and eastern subspecies is
268 suggested from admixture in Dutch and Danish greylags in STRUCTURE. There is historical evidence for the
269 introduction of *rubrirostris* to Belgium in 1954 and to Netherlands in 1960s [9,51]; thus, *rubrirostris* genes
270 may have originated from the recently introduced gene pool spreading to Denmark.

271 b) When and where were geese domesticated?

272 Traditional views on goose domestication claim it first occurred in the eastern Mediterranean (possibly
273 Egypt) around the 3rd Millennium BCE [17,52]. Domestication of chicken and perhaps pigeon took place
274 earlier, but domestication of duck later, at least in Europe [23]. Demographic modelling suggests that the
275 wild greylag and related domestic lineages split approximately 5,300 generations ago placing domestication
276 origins at 14 000 BCE assuming a 3-year generation time [15]. This estimated genetic divergence time is,
277 admittedly, considerably earlier than any evidence for animal domestication except dog. It is important to
278 note that the estimated divergence times have large confidence intervals and merely indicates the split
279 between the ancestors of contemporary wild and domestic lineages. It is most likely that our demographic
280 modelling reflects the early divergence of different lineages of greylags, only one of which contributed to
281 later domestication. The subsequent reduction or even disappearance of that wild lineage means that,
282 despite wide geographical sampling, the possible modern wild population(s) of the greylag progenitor to
283 domestic geese was not sampled in this study and must, therefore, be sought by ancient DNA approaches.
284 It is, however, tempting to make our own determination of the location of goose domestication from
285 the modern data presented here. Given that genetic diversity would be expected to be highest in the
286 'domestication centre' and reduce with increasing distance from there, the high mtDNA diversity of Turkish
287 domestic geese means the eastern Mediterranean remains a strong candidate for the origin of goose
288 domestication, and our nuclear data accord with that. However, as we have shown, hybridisation between
289 wild and domestic geese can also generate high genetic diversity both within and outside the original
290 domestication location.

291 c) The role of intra- and interspecific hybridisation in goose domestication history

292 i. Evidence of current hybridisation

293 Domestic animals and their wild relatives are often observed to interbreed, and this is also true for geese.
294 Both field observations and mtDNA results [9,10] suggested some current hybridisation between domestic
295 and greylag geese. Genome-wide analysis covering multiple greylag populations and domestic breeds
296 revealed a considerable impact of hybridisation on genetic diversity of both wild and domestic geese.
297 Hybridisation is particularly prevalent in certain geographical regions. Dutch and especially Turkish
298 wild greylag samples had more shared genetic affiliation with domestics than Scandinavian and Finnish
299 greylag populations (Figure 3b). Some regions may offer more hybridisation opportunities, e.g. climate may
300 allow greylags to be sedentary year-round and be favourable for keeping domestic geese. The Netherlands,
301 for instance, lies on the Atlantic flyway offering breeding, staging and wintering areas for greylags [53,54].
302 Since pair-bonding of geese generally occurs on wintering grounds [55], hubs for migrating geese such as
303 the Netherlands may permit population mingling. Nevertheless, the f_3 test did not support a simple history
304 of admixture for the Netherlands. Patterson et al. [40] have stated that population-specific drift may mask
305 the signal of admixture in such analyses, leading to a non-negative f_3. The f_3 model is relatively simple, with

306 only two sources, and may not catch the signal of admixture in the Dutch greylag population because of the
307 previous contribution of *rubrirostris*, which was not included in the model.
308 Based on ringing data most greylag populations in Scandinavia follow the Atlantic flyway - some of
309 the geese wintering in the Netherlands and others in southwest Spain. However, Finnish greylags favour
310 the Central European flyway and winter in North Africa, with a minority of Finnish greylags using the
311 Atlantic Flyway [53,54]. The Finnish populations of greylag showed the lowest admixture proportions with
312 domestic geese (S-Finland 3.2% domestic goose, N-Finland 3.3% domestic geese) among greylag
313 populations. Rearing geese is not a popular practice in Finland and they constituted less than 5% of poultry
314 kept in Finland in 2014 [56]. The Norwegian populations showed only slightly higher admixture proportions
315 with domestic geese, although domestic mtDNA haplotype ANS19 was detected from a wild greylag
316 collected in Finnmark, Norway [57]. This haplotype is a partial sequence of the D5 haplotype identified by
317 Heikkinen et al. [10], and identical to that found in White Roman domestic geese [58].
318 Inferring the hybridisation patterns in the Turkish greylags is more complicated, as Turkish greylags
319 indicate hybridisation with both Chinese and European domestics. Both greylags sampled in Turkey showed
320 considerable admixture with domestic geese. One of them appeared genetically as a hybrid of European
321 and Chinese domestic goose with only a small proportion of greylag ancestry, whereas the other one was a
322 more equal mix of European domestic goose and greylag supplemented by a considerable Chinese
323 domestic goose ancestry. There is some indication of hybridisation between greylags and domestic geese
324 within that area as the domestic geese sampled from the same area showed some admixture with greylags,
325 but this was not confirmed with f_3 analysis. These results may reflect a local practice of keeping captive
326 greylags within a flock of domestic geese. Another possibility is that the Turkish greylags have hybridised
327 with some unsampled distinct greylag population and simply appear genetically like domestic geese due to
328 lack of representation of the unsampled wild population. The greylag population breeding and wintering in
329 the Black Sea region is not well monitored [59].

330 ii. Long-term hybridisation

331 Domestication can be seen as an analogy of speciation where an animal population transforms to an
332 ecotype that is adapted to the human niche [23] and at later stages of domestication is perpetuated with
333 reproductive isolation in the form of selection managed by humans [60]. However, this reproductive
334 isolation may not be complete e.g., in pig [7]. While the genetic divergence of the greylag and its domestic
335 descendant is evident, our results suggest extensive long-term genetic exchange between them. In
336 addition, the demographic modelling suggests that the gene flow patterns have changed over time.
337 Initially, gene flow was greater from greylag to domestic geese. Possibly early goose herders
338 supplemented the domestic gene pool with wild-caught greylags. Alternatively, wild ganders may have
339 interbred with domestic geese and contributed to the domestic gene pool. It is unlikely that the early

stages of goose domestication were rigorously managed, allowing matings outside the domestic gene pool. Both would explain the observed higher gene flow from greylag to domestic geese in the early stages of domestication. It is in the farmers' interest to keep the domestic geese and wild geese reproductively isolated to keep control over the traits that are being selected, but artificial selection of traits would have become possible only after the domestic gene pool had been established. After that, it may occasionally be beneficial to restock the flock to maintain enough genetic diversity. The natural tendency for imprinting in geese facilitates this practice. Later, the balance of gene flow shifted towards greylags. This likely occurred after goose-keeping became well-established in the Medieval period [17]. The rise in number of domestic geese may have allowed an increase in domestic goose escapees resulting in increased gene flow from domestic geese back to greylags.

Furthermore, not only have domestic geese admixed with wild greylags but also European and Chinese domestic geese have hybridised. Hybridisation with ancestral species or closely related species is a rule rather than an exception in domestic species, e.g., the genetic makeup of chicken derives from multiple different species of *Gallus* [8]. Similarly, the genetic make-up of domestic geese seems to derive from two closely related species. This hybridisation with Chinese domestic geese may have introduced some traits not present in greylags to European domestic geese and vice versa.

5. Conclusion

The genome-wide data reported here reveals that wild and domesticated European and Chinese greylag and swan goose have intermingled relationships where selection, domestication and hybridisation all play a role. The two independent domestication events of closely related species provide an excellent system to further investigate the relative roles of mutation, selection and gene flow. This includes investigation of potentially adaptive gene flow between the domesticates and identification of shared genome regions subject to anthropogenic selection and extent of convergent evolution in a domestication setting. A more detailed analysis of the relationships of wild greylag and swan geese and their domestic descendants would require wide geographic sampling of both wild species and the domestic breeds combined with an application of ancient biomolecules and advanced biometrics applied to the remains of geese recovered from archaeological sites.

Authors' contributions.

MEH conceived the study, contributed to data collection, analysed the data and drafted the manuscript. MR acquired the funding, conceived the study and contributed to data collection. TAW contributed to data collection and participated to data analysis. MMA and İG contributed to data collection. KMD, JA, JBS and

TP conceived the study and contributed to writing and interpretation of data, with TP also participating in data analysis. All authors, excluding MR, reviewed, improved and approved the manuscript.

Competing interests.

The authors declare no conflict of interest.

Funding.

This work was supported by the Academy of Finland (grant no. 131673 to MR, grant no 283609 to JA and grant no. 287431 to TP); and the Cornell Center for Comparative and Population Genomics (3CPG) to JBS. The Emil Aaltonen Foundation and Oulun läänin talousseuran maataloussäätiö are acknowledged for the personal grants awarded to MEH.

Acknowledgements.

We thank the Ministeriet for Fødevarer, Landbrug og Fiskeri in Denmark, the Svenska Lanthönsklubben in Sweden and the Goose Club in the UK for their help in obtaining the domestic goose samples. Silvia Markova is acknowledged for her help with the laboratory work. We are grateful for Tiina Mattila for her help with the fastsimcoal2 analysis. Lastly, we thank the CSC – IT Center for Science in Finland for providing computing resources.

References

1. Arnold ML. 2004 Natural hybridization and the evolution of domesticated, pest and disease organisms. *Mol. Ecol.* **13**, 997–1007. (doi:10.1111/j.1365-294X.2004.02145.x)
2. Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, et al. 2013 Hybridization and speciation. *J. Evol. Biol.* **26**, 229–246. (doi:10.1111/j.1420-9101.2012.02599.x)
3. Grant PR, Grant BR. 1992 Hybridization of bird species. *Science* **256**, 193–197. (doi:10.1126/science.256.5054.193)
4. Ottenburghs J, van Hooft P, van Wieren SE, Ydenberg RC, Prins HHT. 2016 Hybridization in geese: a review. *Front. Zool.* **13**, 20. (doi:10.1186/s12983-016-0153-1)
5. Janzen GM, Wang L, Hufford MB. 2019 The extent of adaptive wild introgression in crops. *New Phytol.* **221**, 1173-1664. (doi:10.1111/nph.15457)
6. Godinho R, Llaneza L, Blanco JC, Lopes S, Álvares F, García EJ, et al. 2011 Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. *Mol.*

399 *Ecol.* **20**, 5154–5166. (doi:10.1111/j.1365-294X.2011.05345.x)

400 7. Frantz LAF, Schraiber JG, Madsen O, Megens H-J, Cagan A, Bosse M, et al. 2015 Evidence of long-
401 term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig
402 genomes. *Nat. Genet.* **47**, 1141–1148. (doi:10.1038/ng.3394)

403 8. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. 2008
404 Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet.* **4**,
405 e1000010. (doi:10.1371/journal.pgen.1000010)

406 9. Kuijken E, Devos K. 1996 The status of the greylag goose *Anser anser* in Flanders, Belgium. *Wetlands*
407 *International Goose Specialist Group Bulletin* **26**, 8.

408 10. Heikkinen ME, Ruokonen M, Alexander M, Aspi J, Pyhäjärvi T, Searle JB. 2015 Relationship between
409 wild greylag and European domestic geese based on mitochondrial DNA. *Anim. Genet.* **46**, 485–497.
410 (doi:10.1111/age.12319)

411 11. Delacour J. 1954 *The waterfowl of the world : Vol. 1, The magpie goose, whistling ducks, swans and*
412 *geese, sheldgeese and shelducks*. 2nd ed. London: Country Life.

413 12. Shi X-W, Wang J-W, Zeng F-T, Qiu X-P. 2006 Mitochondrial DNA cleavage patterns distinguish
414 independent origin of Chinese domestic geese and western domestic geese. *Biochem. Genet.* **44**,
415 237–245. (doi:10.1007/s10528-006-9028-z)

416 13. Ottenburghs J, Megens H-J, Kraus RHS, Madsen O, van Hooft P, van Wieren SE, et al. 2016 A tree of
417 geese: a phylogenomic perspective on the evolutionary history of True Geese. *Mol. Phylogenet. Evol.*
418 **101**, 303–313. (doi:10.1016/j.ympev.2016.05.021)

419 14. Buckland R, Guy G, editors. 2002 Goose production. In *FAO Animal Production and Health Paper -*
420 *154*, Rome: Food and Agriculture Organization of the United Nations.

421 15. Cramp S, Simmons KEL, editors. 1977 *Handbook of the birds of Europe, the Middle East, and North*
422 *Africa: The birds of the Western Palearctic. Vol. 1: Ostrich-Ducks*. New York: Oxford University Press.

423 16. Kear J. 1990 *Man and wildfowl*. London: T & AD Poyser Ltd.

424 17. Albarella U. 2005 Alternate fortunes? The role of domestic ducks and geese from Roman to
425 Medieval times in Britain. In Feathers, grit and symbolism (ed. G Grupe, J Peters), *Documenta*
426 *Archaeobiologiae* **3**, 249–258.

427 18. Honka J, Heino M, Kvist L, Askeyev I, Shaymuratova D, Askeyev O, et al. 2018 Over a thousand years
428 of evolutionary history of domestic geese from Russian archaeological sites, analysed using ancient
429 DNA. *Genes* **9**, 367. (doi:10.3390/genes9070367)

430 19. Sun J, Zhang S, He D-Q, Chen S-Y, Duan Z-Y, Yao Y-G, et al. 2013 Matrilineal genetic structure of
431 domestic geese. *J. Poult. Sci.* **51**, 130–137. (doi:10.2141/jpsa.0120152)

432 20. Ren T, Liang S, Zhao A, He K. 2016 Analysis of the complete mitochondrial genome of the Zhedong
433 White goose and characterization of NUMTs: reveal domestication history of goose in China and
434 Euro. *Gene* **577**, 75–81. (doi:10.1016/j.gene.2015.11.018)

435 21. McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM. 2013 New World cattle show ancestry
436 from multiple independent domestication events. *Proc. Natl. Acad. Sci. USA* **110**, E1398–E1406.
437 (doi:10.1073/pnas.1303367110)

438 22. van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, de Jesus Sanchez Gonzalez J,
439 et al. 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces.
440 *Proc. Natl. Acad. Sci. USA* **108**, 1088–1092. (doi:10.1073/pnas.1013011108)

441 23. Larson G, Fuller DQ. 2014 The evolution of animal domestication. *Annu. Rev. Ecol. Evol. Syst.* **45**,
442 115–136. (doi:10.1146/annurev-ecolsys-110512-135813)

443 24. Marshall FB, Dobney K, Denham T, Capriles JM. 2014 Evaluating the roles of directed breeding and
444 gene flow in animal domestication. *Proc. Natl. Acad. Sci. USA* **111**, 6153–6158.
445 (doi:10.1073/pnas.1312984110)

446 25. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. 2011 A robust, simple
447 genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379.
448 (doi:10.1371/journal.pone.0019379)

449 26. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. 2014 TASSEL-GBS: a high
450 capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**, e90346.
451 (doi:10.1371/journal.pone.0090346)

452 27. Lu L, Chen Y, Wang Z, Li X, Chen W, Tao Z, et al. 2015 The goose genome sequence leads to insights
453 into the evolution of waterfowl and susceptibility to fatty liver. *Genome Biol*. **16**, 89.
454 (doi:10.1186/s13059-015-0652-y)

455 28. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform.
456 *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)

457 29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. 2011 The variant call format
458 and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

459 30. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus
460 genotype data. *Genetics* **155**, 945–959.

461 31. Patterson N, Price AL, Reich D. 2006 Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.

(doi:10.1371/journal.pgen.0020190)

32. Falush D, Stephens M, Pritchard JK. 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

33. Chhatre VE, Emerson KJ. 2017 StrAuto: automation and parallelization of STRUCTURE analysis. *BMC Bioinformatics* **18**, 192. (doi:10.1186/s12859-017-1593-0)

34. Evanno G, Regnaut S, Goudet J. 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620. (doi:10.1111/j.1365-294X.2005.02553.x)

35. Earl DA, VonHoldt BM. 2012 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361. (doi:10.1007/s12686-011-9548-7)

36. Jakobsson M, Rosenberg NA. 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806. (doi:10.1093/bioinformatics/btm233)

37. Rosenberg NA. 2003 distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138. (doi:10.1046/j.1471-8286.2003.00566.x)

38. R Core Team. 2017 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

39. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)

40. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. 2012 Ancient admixture in human history. *Genetics* **192**, 1065–1093.

41. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013 Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905. (doi:10.1371/journal.pgen.1003905)

42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

43. Pujolar JM, Dalén L, Olsen RA, Hansen MM, Madsen J. 2018 First de novo whole genome sequencing and assembly of the pink-footed goose. *Genomics* **110**, 75–79. (doi:10.1016/j.ygeno.2017.08.008)

44. Goudet J. 2005 hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186. (doi:10.1111/j.1471-8286.2004.00828.x)

45. Welch BL. 1938 The significance of the difference between two means when the population

493 variances are unequal. *Biometrika* **29**, 350. (doi:10.2307/2332010)

494 46. Excoffier L, Lischer HEL. 2010 Arlequin suite ver 3.5: a new series of programs to perform population
495 genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567. (doi:10.1111/j.1755-
496 0998.2010.02847.x)

497 47. Moyers BT, Morrell PL, McKay JK. 2018 Genetic costs of domestication and improvement. *J. Hered.*
498 **109**, 103–116. (doi:10.1093/jhered/esx069)

499 48. Kanginakudru S, Metta M, Jakati R, Nagaraju J. 2008 Genetic evidence from Indian red jungle fowl
500 corroborates multiple domestication of modern day chicken. *BMC Evol. Biol.* **8**, 174.
501 (doi:10.1186/1471-2148-8-174)

502 49. Stringham SA, Mulroy EE, Xing J, Record D, Guernsey MW, Aldenhoven JT, et al. 2012 Divergence,
503 convergence, and the ancestry of feral populations in the domestic rock pigeon. *Curr. Biol.* **22**, 302–
504 308. (doi:10.1016/j.cub.2011.12.045)

505 50. Scott DA, Rose PM. 1996 *Atlas of Anatidae populations in Africa and Western Eurasia*. Wetlands
506 International Publication No. 41. Wetlands International, Wageningen, The Netherlands.

507 51. Rooth J. 1971 The occurrence of the greylag goose *Anser anser* in the western part of its distribution
508 area. *Ardea* **59**, 17–27.

509 52. Zeuner FE. 1963 *A history of domesticated animals*. New York and Evanston: Harper & Row.

510 53. Madsen J, Cracknell G, Fox T, editors. 1999 *Goose populations of the Western Palearctic. A review of*
511 *status and distribution*. Wetlands International Publication No. 48. Wetlands International,
512 Wageningen, The Netherlands; National Environmental Research Institute, Rønde, Denmark.

513 54. Andersson Å, Follestad A, Nilsson L, Persson H. 2001 Migration patterns of Nordic greylag geese
514 *Anser anser*. *Ornis Svecica* **11**, 19–58.

515 55. Rohwer FC, Anderson MG. 1988 Female-biased philopatry, monogamy, and the timing of pair
516 formation in migratory waterfowl. In *Current Ornithology* (ed. RF Johnston), pp. 187-221. Boston,
517 MA: Springer US. (doi:10.1007/978-1-4615-6787-5_4)

518 56. Official Statistics of Finland (OSF): Number of livestock [e-publication]. Helsinki: Natural Resources
519 Institute Finland [referred: 17.12.2016]. Access method: http://www.stat.fi/til/klm/index_en.html
520 2016.

521 57. Pellegrino I, Cucco M, Follestad A, Boos M. 2015 Lack of genetic structure in greylag goose (*Anser*
522 *anser*) populations along the European Atlantic flyway. *PeerJ* **3**, e1161. (doi:10.7717/peerj.1161)

523 58. Wang CM, Way TD, Chang YC, Yen NT, Hu CL, Nien PC, et al. 2010 The origin of the White Roman

goose. *Biochem. Genet.* **48**, 938–943. (doi:10.1007/s10528-010-9374-8)

59. Fox AD, Ebbinge BS, Mitchell C, Heinicke T, Aarvak T, Colhoun K, et al. 2010 Current estimates of goose population sizes in western Europe, a gap analysis and an assessment of trends. *Ornis Svecica* **20**, 115–127.

60. Zeder MA. 2012 The domestication of animals. *J. Anthropol. Res.* **68**, 161–190. (doi:10.3998/jar.0521004.0068.201)

61. BirdLife International and Handbook of the Birds of the World (2016) 2016. *Anser anser*. The IUCN Red List of Threatened Species. Version 2018-1.

549 Table 1 The parameter values that maximise likelihood estimates (MLE) of the preferred demographic
550 model for goose domestication history (see text) with their 95% confidence intervals (CI) and the search
551 range for the parameter values.

Model	Number of parameters	Parameter	MLE	95% CI	Search range
Divergence with 2 migration matrices	9	ANCSIZE	1112	291.95 - 9030.75	100 - 100000
		T_1	5319	1538.80 - 8225.45	100 - 20000
		$M1_{WD}$	$5.35x10^{-4}$	$2.96x10^{-4}$ - $6.75x10^{-4}$	$1x10^{-5}$ - 20
		$M1_{DW}$	$4.25x10^{-4}$	$1.73x10^{-4}$ - $6.03x10^{-4}$	$1x10^{-5}$ - 20
		T_2	159	79.65 - 462.95	50 - 10000
		$M2_{WD}$	$6.69x10^{-4}$	$4.57x10^{-4}$ - $9.69x10^{-4}$	$1x10^{-5}$ - 20
		$M2_{DW}$	$1.72x10^{-3}$	$1.33x10^{-3}$ - $2.15x10^{-3}$	$1x10^{-5}$ - 20
		N_{DOM}	959	879.95 - 1073.25	50 - 100000
		N_{WILD}	2504	2341.60 - 2695.55	50 - 100000

ANCSIZE, effective population size of ancestral population; T_1, time of divergence in generations; N_{DOM}, effective population size for domestic geese; N_{WILD}, effective population size for greylags; T_2, estimate of time in generations when the migration matrix switched; $M1_{WD}$ migration rate from wild to domestic following T_1; $M1_{DW}$ migration rate from domestic to wild following T_1; $M2_{WD}$ migration rate from wild to domestic following T_2; $M2_{DW}$ migration rate from domestic to wild following T_2.

552

553 Table 2 Hierarchical analysis of molecular variance (AMOVA) of greylags and their domestic descendants,
554 considering pure populations of greylags (first group) and pure breeds of European domestic geese (second
555 group).

Source of variation	Sum of squares	Variance components	Percentage variation	Fixation indices
Among groups	47565.119	431.4291	15.8	$F_{CT} = 0.158^{**}$
Among populations and breeds within groups	82960.489	302.51404	11.1	$F_{SC} = 0.131^{**}$
Within populations and breeds	345889.821	2003.45893	73.2	$F_{ST} = 0.268^{**}$
Total	476415.429	2737.40207		

** $p < 0.001$

556

557

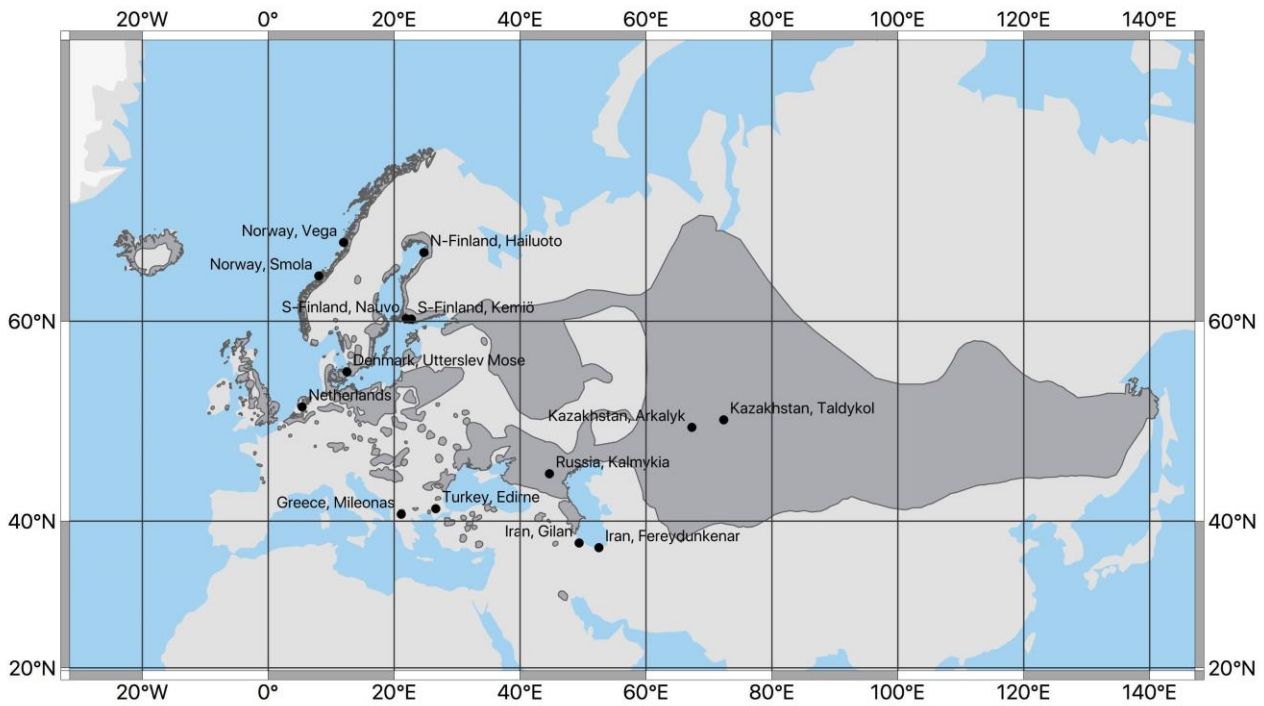


Figure 1 Map showing the sampling sites for wild greylags used in this study. The breeding area of the species is shown on darker grey. The sampling sites in Kazakhstan were combined for analyses (one sample per location) and the sampling sites in Southern Finland included combined samples from the geographically close sites of Västanfjärd, Nauvo (shown) and Kemiö (shown). The Iranian samples were collected during the wintering season. Map modified from IUCN [61].

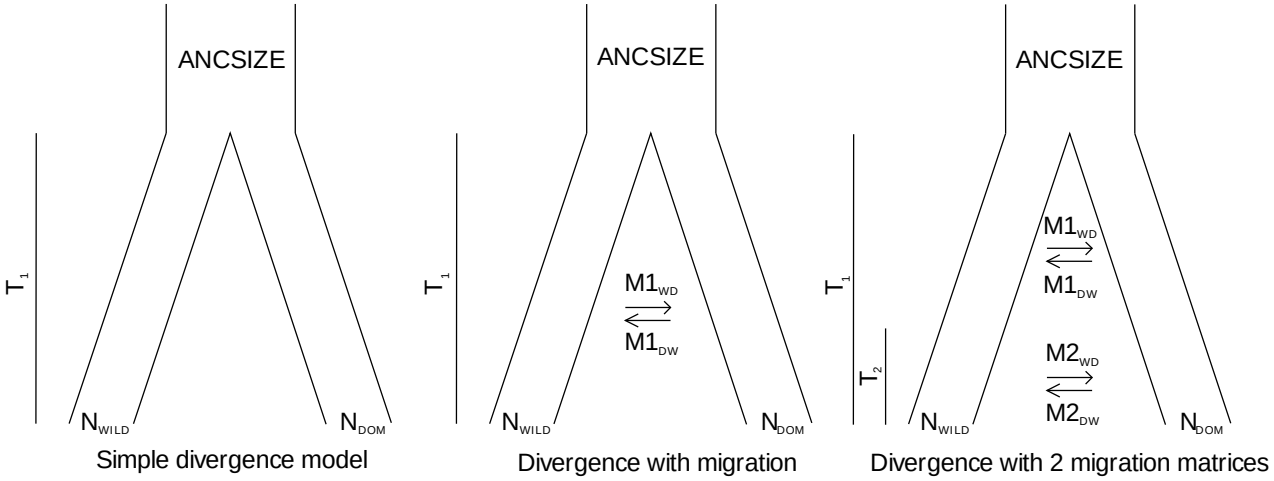
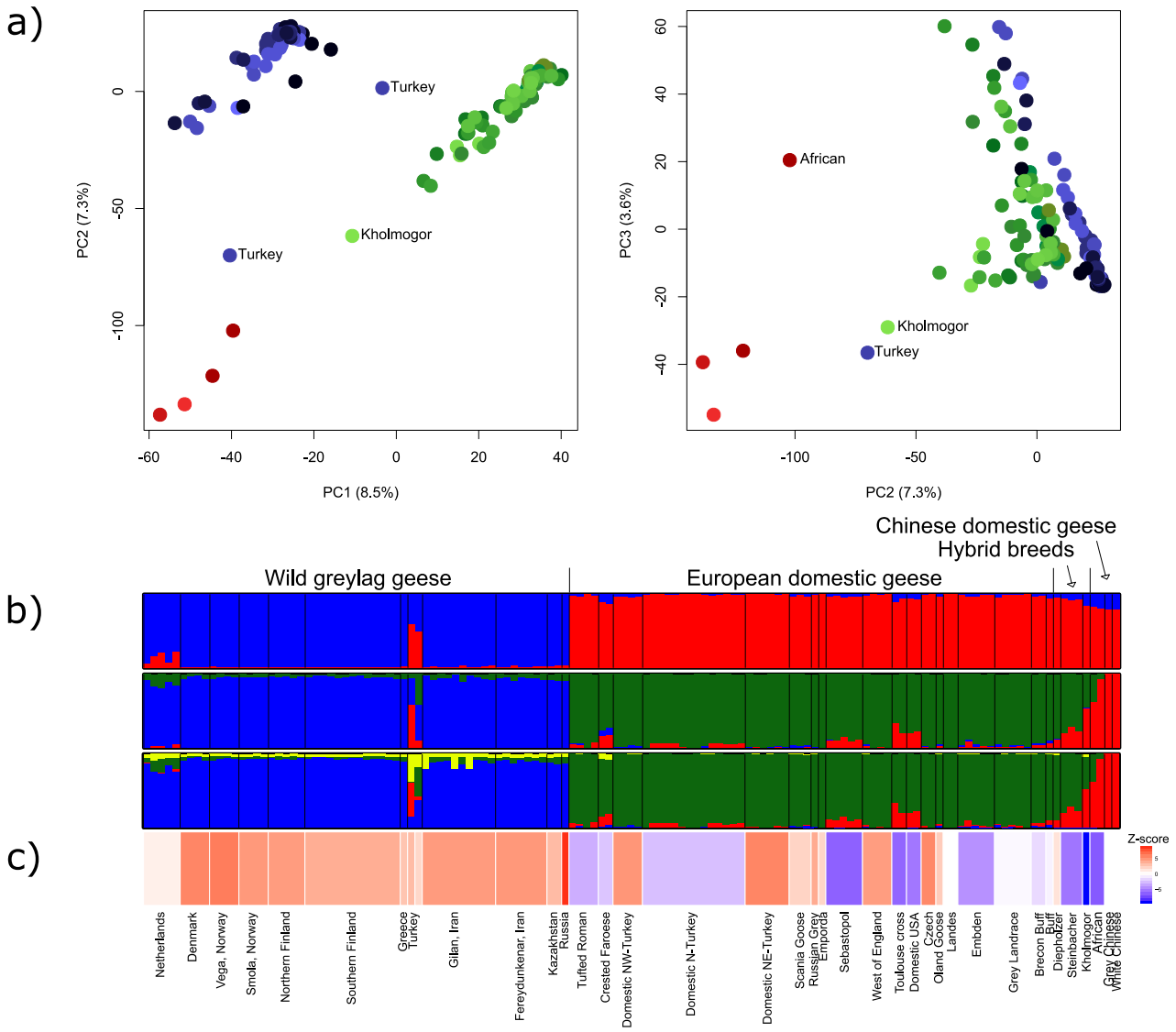


Figure 2 Demographic histories of goose domestication as tested with fastsimcoal2.



568

569 Figure 3 The genetic divergence and hybridisation patterns in greylag and domestic geese. Population
570 status and names labelled as in Table S1. The colours in a) and b) are associated to different groups as
571 follows: greylags (blue), European domestics (green) and Chinese domestics (red). a) The first three
572 principal components summarising the genetic variation in geese (percentage explained by each PC is
573 shown). b) STRUCTURE assignment plots for *K*=2, *K*=3 and *K*=4. Each vertical line represents one individual
574 with *K* number of colours indicating proportion of ancestry from the inferred clusters. c) Plot relating to the
575 f_3 (Table S2) values obtained for each population. The more negative the f_3, the more significant is Z-score
576 in favour of admixture. The f_3 values were not calculated for Landes and the Chinese geese, as they were
577 used as source populations, thus they were given an f_3 value of 0.