Supplementary Information for

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

Marja E. Heikkinen, Minna Ruokonen, Thomas A. White, Michelle M. Alexander, İslam Gündüz, Keith M. Dobney, Jouni Aspi, Jeremy B. Searle, Tanja Pyhäjärvi

Co-corresponding authors:

Marja E. Heikkinen
Email: marja.e.heikkinen@oulu.fi

Tanja Pyhäjärvi
Email: tanja.pyhajarvi@oulu.fi

This PDF file includes:

Supplementary text
Figures S1 to S4 and S6 to S16
Legend for Figure S5
Tables S1, S5 and S8
Legend for Tables S2-S4 and S6-S7
SI References

Other supplementary materials for this manuscript include the following:

Figure S5
Tables S2-S4 and S6-S7

Extended methods.

GBS pipeline and SNP calling

Raw sequence reads from Illumina were run through the Command Line Interface of the Tassel 5 GBS v2 Discovery and Production pipelines (Glaubitz *et al.* 2014) on the Taito supercluster maintained by the CSC - IT Center for Science in Finland. Figure S1 illustrates the workflow of the pipeline. The GBSSeqToTagDBPlugin was run with the default setting except for the minimum quality score set to 20. This plugin identifies good quality reads with the barcode and cut site from the raw sequence data and trims off barcodes and truncates sequences if another cut site is found in the sequence. The reads that were pulled from the raw data were trimmed to 64 bp (base pairs) but if a second cut site was found in the read, the sequence was truncated and kept if the remaining sequence was longer than 20 bp. The good quality reads are recorded as tags and along with individuals in which they appear, the tags are stored in the local database. Given the parameters used, the length of each good quality tag ranged from 20 to 64 bp and each position in the sequence had a minimum quality score of 20 from Illumina sequencing. After this, the tags were aligned to the *Anser cygnoid domesticus* GenBank assembly (AnsCyg_PRJNA183603_v1.0 GCF_000971095.1) (Lu *et al.* 2015) using the Burrows-Wheeler Aligner with default settings (Li and Durbin 2009). Then, the SAMToGBSdbPlugin was run with default settings to determine the potential positions of tags in the reference genome and the position information was recorded to the local database. Altogether 285,760 tags were mapped on the reference genome and 66,163 tags were left unmapped. In the next step the DiscoverySNPCallerPluginV2 was used to align the tags positioned in the same physical location with each other and called single nucleotide differences between aligned tags as SNPs. The SNP position and allele data were stored in the local database. The DiscoverySNPCallerPluginV2 was run with the default settings with the following change: the proportion of individuals with the genotype in the locus, the minimum locus coverage, was set to 0.8. Finally, the ProductionSNPCallerPluginV2 was used to convert the data from the local database to VCF format. The Tassel-GBS pipeline does not filter for sequencing depth per se because it is optimized for large numbers of markers in a large sample of individuals at the expense of sequencing depth (optimization of the pipeline is done for sequencing depth of 0.5 -3 x) (Glaubitz *et al.* 2014). Therefore, the mean sequencing depths in our data ranged between 1.2 – 1310 x for each SNP averaged across individuals. The low coverage in some of the SNPs was compensated by the fact that the minimum locus coverage was set to 0.8 meaning that each SNP was genotyped in at least 80% of the samples. After running the raw data through Discovery and Production pipelines, the resulting number of SNPs was 69,865.

After that, the SNPs were subjected to additional filtering using VCFtools (Danecek *et al.* 2011). We removed indels, loci with more than two alleles and invariant loci that differed from the reference. However, these invariant sites were retained in the phylogenetic tree construction as they were informative about the divergence from the swan goose. After preliminary analyses we also removed loci with observed heterozygosity over 0.75, because they were potential paralogs mapping to the same reference locus. We applied a filter that removed individuals that showed more than 20% missing data across loci. After these filtering steps, we had a dataset that consisted of 33,527 biallelic SNPs and 133 individuals that were successfully genotyped for at least 80% of the loci of which 58 were wild graylags and 75 were domestic geese.

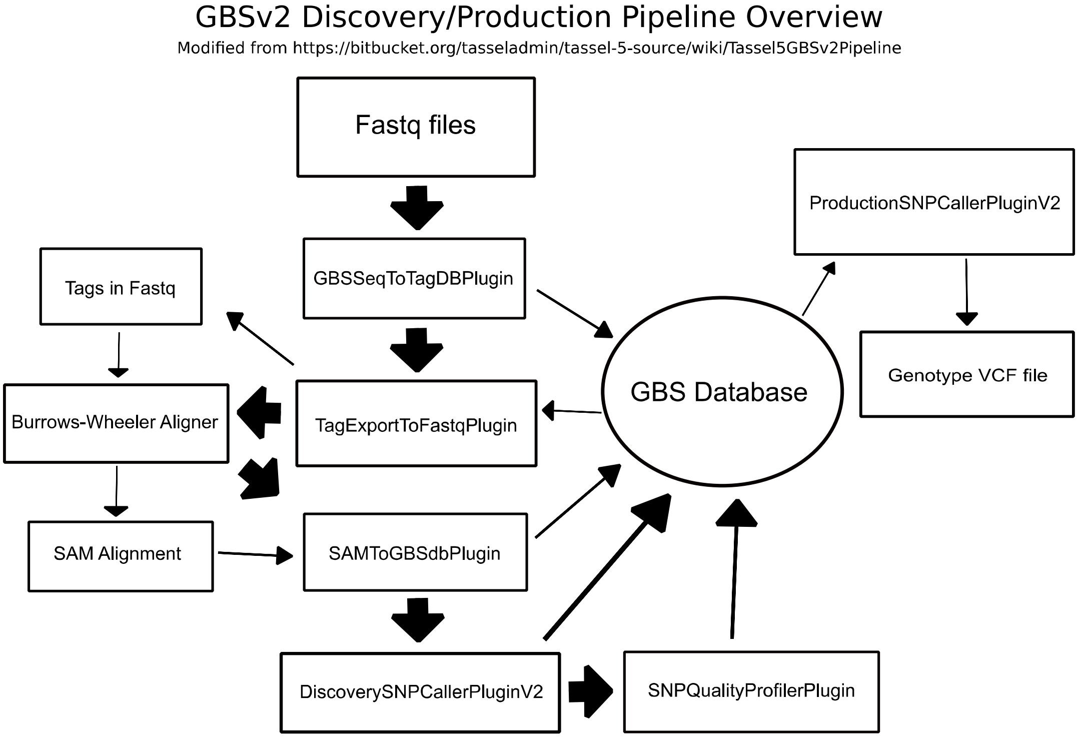

GBSv2 Discovery/Production Pipeline Overview
Modified from https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline
Fastq files
GBSSeqToTagDBPlugin
Tags in Fastq
ProductionSNPCallerPluginV2
GBS Database
Genotype VCF file
Burrows-Wheeler Aligner
TagExportToFastqPlugin
SAM Alignment
SAMToGBSdbPlugin
DiscoverySNPCallerPluginV2
SNPQualityProfilerPlugin

Figure S1. Workflow for the GBS pipeline.

Population structure analyses

Population clustering and structure at the individual level was analyzed with STRUCTURE 2.3.4 (Pritchard *et al.* 2000) and Principal Component Analysis (PCA) (Menozzi *et al.* 1978; Patterson *et al.* 2006) for the whole dataset and within graylags and domestic geese. The Bayesian STRUCTURE approach aims to find an optimal number of clusters (*K*) from a given dataset without prior population/group information by assuming that loci are in linkage equilibrium and each population is in Hardy-Weinberg equilibrium. The clustering of populations is done by considering the individual genotypes and estimating the allele frequencies in populations. For the whole dataset, STRUCTURE was run with 1,000 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; this appeared to be enough to reach convergence. For the STRUCTURE analyses done separately on graylags and European domestic geese we increased the number of burn-in steps to 10,000 and number of MCMC to 50,000 and set the *K* to 1-7.

We also tested the impact of sample size to our results; therefore, we made some additional analyses with STRUCTURE and PCA by subsetting the data. Firstly, we took a random subsample of 58 European domestic geese to match with our 58 graylag samples. The STRUCTURE analysis and PCA were then carried on with the same settings as with the whole data and the Chinese domestic geese were also included in the analyses. Secondly, we omitted the Chinese domestic geese from the data and repeated the analyses with 58 graylag and 58 European domestic goose samples. After some experimenting, we settled on burn-in length of 10,000 steps and MCMC of 50,000 steps. Lastly, we took a random subset of 4 individuals of both graylag and European domestic geese and analyzed them together with Chinese domestic geese. For this last analysis, we increased the number of burn-in steps to 20,000 and kept the number of MCMC iterations in 50,000 steps. As with the whole dataset, each run was repeated five times. An admixture model with correlated allele frequencies among populations (Falush *et al.* 2003) was used in all the STRUCTURE analyses.

The iterations for STRUCTURE analyses were automated with the script StrAuto 1.0 (Chhatre and Emerson 2017). We used both likelihood of *K* and Evanno's Δ*K* (Evanno *et al.* 2005) of successive *K* values to determine the optimal number of clusters, as implemented in STRUCTURE HARVESTER (Earl and VonHoldt 2012). CLUMPP 1.1.2 (Jakobsson and Rosenberg 2007) was used to align the assignments from different replicates of *K* and the results were used as an input for visualization with the program DISTRUCT 1.1 (Rosenberg 2003). A PCA was performed with prcomp function in R (R Core Team 2017) and the significance of the eigenvalues was determined based on the Tracy-Widom distribution (Patterson *et al.* 2006; van Heerwaarden *et al.* 2011).

To visualize the genetic differences and distance between the reference genome and our data, a neighbor-joining tree was constructed. The tree estimation was performed based on a pairwise distance matrix computed between individuals with the R package ape (Paradis *et al.* 2004). The *A. cygnoid* reference genome was included in the construction of the neighbor-joining tree and the invariant sites that differed from the reference genome within our data set were also included, thus the tree was constructed with 40,191 loci.

Tests for admixture and simulations of demographic history

The history of admixture was tested with the 3-Population test f_3(C; A, B) implemented in AdmixTools 4.1 (Patterson *et al.* 2012). This method offers a formal test of admixture that can be used to explain the observed patterns of admixture in a target population and does not require an outgroup. The f_3 test allows separation of ancient polymorphisms from the effects of true admixture, which may be confounded in STRUCTURE. The admixture model is simple, with two source populations contributing to single target population. For identification of admixture between Chinese and European domestics, Grey and White Chinese were combined to represent the Chinese domestic source population and Landes breed that had minimum indication of admixture in STRUCTURE was chosen to represent the European domestic geese source population. We made several analyses of type f_3(graylag pop2; graylag pop1, European domestic pop) to detect admixture in graylag populations, and f_3(European domestic pop2; European domestic pop1, graylag pop) to detect admixture in domestic populations. We also tested scenarios were Chinese domestic goose was one of the source populations f_3(graylag pop2; graylag pop1, Chinese domestic pop) and f_3(European domestic pop2; European domestic pop1, Chinese domestic pop).

Different models of demographic history were tested with fastsimcoal2 ver 2.6 (Excoffier *et al.* 2013). Only sites without missing data were used for demographic analyses. We excluded all the SNPs that had missing data within the whole dataset and executed the analyses with a site frequency spectrum (SFS) that contained 6,229 polymorphic SNPs (Figure S2). The model estimation utilizing the SFS also requires information on the number of monomorphic sites. As there are no estimates of the genetic diversity per base pair for graylags, we made a rough estimation of the proportions of variable and monomorphic sites in our data. The number of bases covered by the GBS tags was calculated from BAM file with –depth option available in SAMtools 1.7 (Li *et al.* 2009). No threshold value was used for this, all the sites that were covered with the tags were recorded, regardless of their sequencing depth or quality. This resulted in 9,801,382 bases covered with tags. After this, we mimicked the filtering steps done for the biallelic SNPs to reduce the total number of sites in equivalent proportions. We removed the same number of sites that corresponded to the number of SNPs that were removed because they were indels, had more than 2 alleles or had heterozygosity over 0.75. Since a proportion of the SNPs were removed from this analysis due to missing data in some of the individuals, we removed an equal proportion of sites from the total number of sites as well. The final SFS had 1,681,316 sites of which 1,675,087 were monomorphic and 6,229 polymorphic.

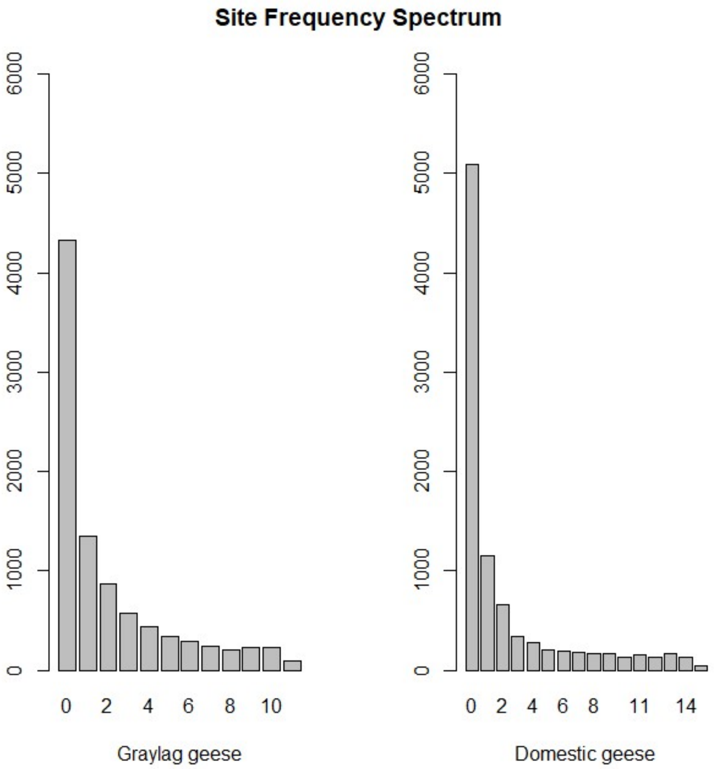
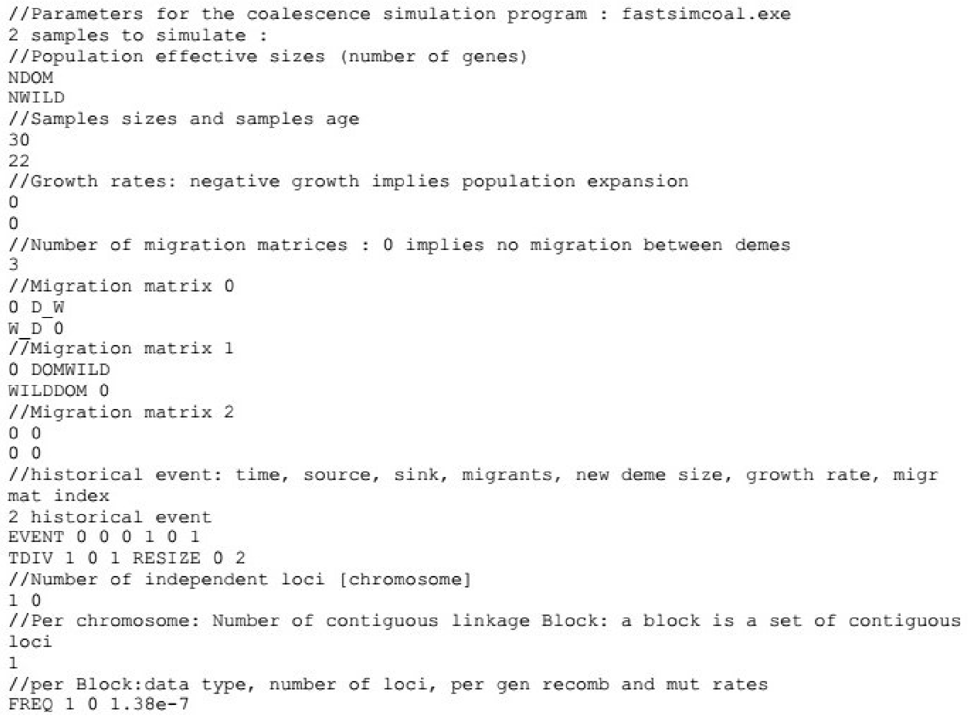


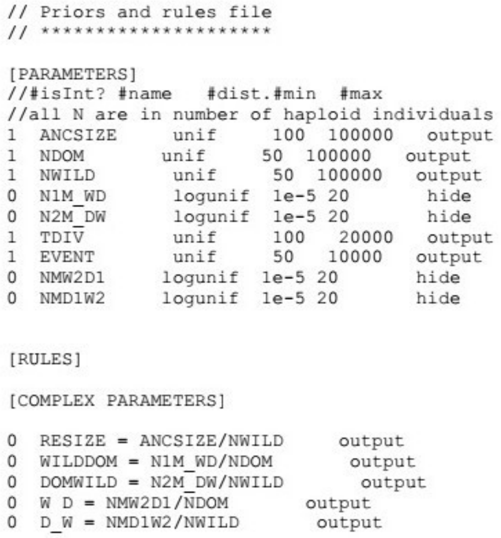
Figure S2. A site frequency spectrum for graylag and domestic geese.

For inferring the demographic history, we chose a subset of individuals from both wild graylags and domestic geese to represent the genetic variation in both groups. This selection was done based on their admixture coefficients from the STRUCTURE analysis so that each wild and domestic population, excluding those that were clearly of a hybrid origin, were represented by an individual with the least amount of admixture from other groups. Therefore, 11 graylags with > 90.8% of graylag ancestry and 15 domestic geese with > 91.4% of European domestic goose ancestry, were selected for the analysis. By doing this, we wanted to minimize the effect of recent admixture on the estimation of divergence time of these two groups, essentially to get the most accurate estimate of the domestication time available. To simulate possible population histories, the parameter estimation for each model involved 100,000 simulations and 40 conditional maximization (ECM) cycles. The parameters for each model were estimated with 100 independent runs to obtain the global maximum. The examples of input files for parameter estimation where parameters and priors are specified are presented in Figure S3 and Figure S4, respectively. The models tested were i) simple divergence of two populations with no gene flow, ii) divergence of two populations with continuous gene flow and lastly, iii) divergence of two populations with changing gene flow patterns (Figure 2). The best model to represent our data was selected based on Akaike's weight of evidence as in Excoffier *et al.* (2013). For parametric bootstrapping 100 SFS were simulated with the parameter estimates obtained from the real SFS, followed by maximum likelihood estimation with 50 independent runs for each bootstrap SFS. The 95% confidence intervals were obtained from the bootstrap data for each estimated parameter.



```
//Parameters for the coalescence simulation program : fastsimcoal.exe
2 samples to simulate :
//Population effective sizes (number of genes)
NDOM
NWILD
//Samples sizes and samples age
30
22
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
3
//Migration matrix 0
0 D_W
W_D 0
//Migration matrix 1
0 DOMWILD
WILDDOM 0
//Migration matrix 2
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, growth rate, migr
mat index
2 historical event
EVENT 0 0 0 1 0 1
TDIV 1 0 1 RESIZE 0 2
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous
loci
1
//per Block:data type, number of loci, per gen recomb and mut rates
FREQ 1 0 1.38e-7
```

Figure S3. An example of an input file that specifies the model for fastsimcoal2.



```
// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name   #dist.#min  #max
//all N are in number of haploid individuals
1  ANCSIZE     unif     100  100000   output
1  NDOM       unif     50  100000   output
1  NWILD       unif     50  100000   output
0  N1M_WD      logunif  1e-5 20       hide
0  N2M_DW      logunif  1e-5 20       hide
1  TDIV        unif     100   20000   output
1  EVENT       unif     50   10000   output
0  NMW2D1     logunif  1e-5 20       hide
0  NMD1W2     logunif  1e-5 20       hide


[RULES]

[COMPLEX PARAMETERS]

0  RESIZE = ANCSIZE/NWILD     output
0  WILDDOM = N1M_WD/NDOM       output
0  DOMWILD = N2M_DW/NWILD       output
0  W D = NMW2D1/NDOM       output
0  D_W = NMD1W2/NWILD       output
```

Figure S4. An example of an input file that specifies the priors for fastsimcoal2.

The estimation of genetic diversity

Genetic diversity and pairwise F_{ST} values were investigated using the *hierfstat* R package (Goudet 2005). Expected heterozygosity (H_E) was calculated for each locus and population and averaged across loci. Difference in average H_E between graylags and European domestics was tested with a two-sample t-test with the Welch correction for non-homogeneity of variance (Welch 1938). To compare the genetic diversity among wild and domestics, only pure graylag populations (defined as having <10% admixture with domestic geese) and pure European domestic geese (defined as having <10% admixture with Chinese domestic geese) were used to avoid hybridization effects on the estimates. The admixture proportions were obtained from STRUCTURE analysis detailed above. Therefore, the graylag populations in the Netherlands and Turkey as well as the domestic populations Diepholzer, Crested Faroese, Sebastopol, Toulouse cross, Domestic NY, Buff, Steinbacher and Kholmogory were excluded from the group estimates.

The variance components across loci for hierarchical F statistics for pure graylags and pure European domestics were estimated using hierarchical locus-by-locus analysis of molecular variance (AMOVA, (Excoffier *et al.* 1992)) implemented in Arlequin 3.5.2.1 (Excoffier and Lischer 2010). The significance was tested with 16,000 permutations.

Extended results

Genetic structure

The neighbor-joining tree confirmed the major patterns that were observed in STRUCTURE and PCA (Figure S5). In addition, it offered some insight into population structure within the groups that was not visible in STRUCTURE and PCA when the whole data was analyzed together.

It was clear that the sample size did not affect the main signal from the data. When we analyzed 58 graylags with 58 European domestic geese and 4 Chinese domestic geese, we observed essentially the same pattern as with the whole data. Evanno and likelihood methods both supported $K = 3$ (Figure S6). The PCA supported the STRUCTURE result (Figure S7).

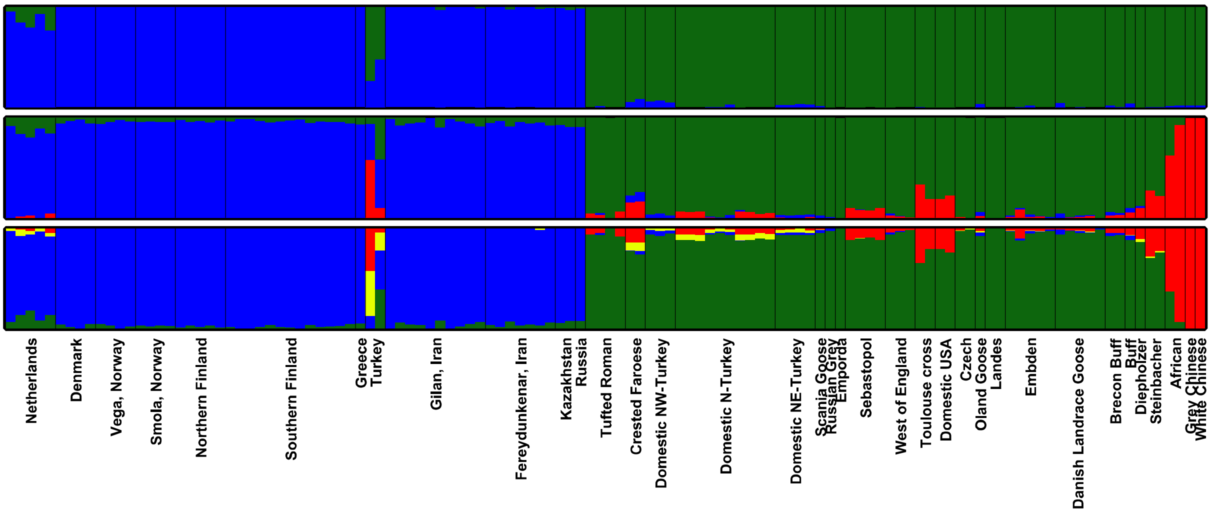


Figure S6. STRUCTURE assignment plots for when the sample sizes for graylags and European domestic geese are equal. Each vertical bar represents one individual with *K* number of colors indicating proportion of ancestry from the inferred clusters, and populations are separated by black vertical line.

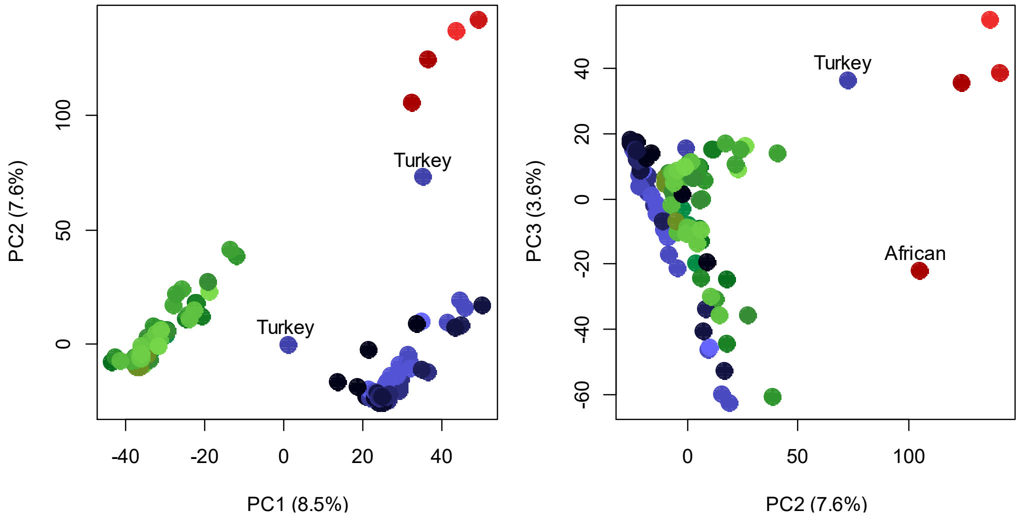


Figure S7. The first three principal components summarizing the genetic variation when sample the sample sizes for graylags and European domestic geese are equal. The colors are associated to different groups as follows: graylags (blue), European domestics (green) and Chinese domestics (red). Different shades refer to different populations. The percentages explained by each PC are shown on the X and Y axes.

When we analyzed only four individuals from each group (graylag, European domestic, Chinese domestic) we observed that the graylags and European domestics were inseparable in STRUCTURE (Figure S8). The Evanno method suggested the optimal number of *K* to be 5, but the likelihood was highest for $K = 2$. $K = 2$ appeared to detect the relevant groups in this dataset. The PCA was in accordance with that assessment and the Tracy-Widom distribution suggested that only the first PC was significant, thus separating the *A. anser* ancestry from *A. cygnoid* ancestry (Figure S9).

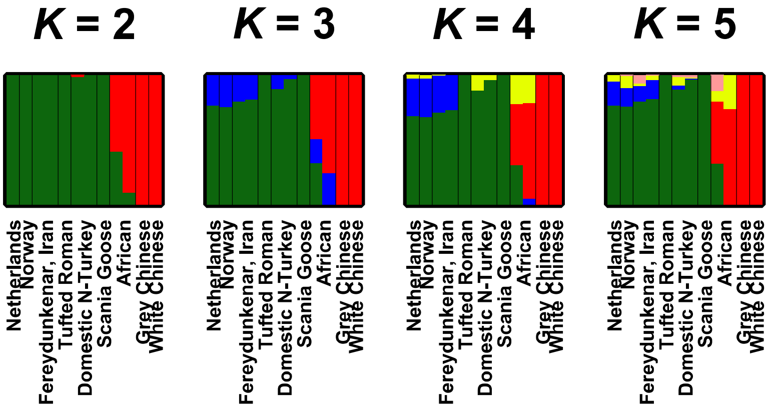


Figure S8. STRUCTURE assignment plots *K* = 2-5 when sample size is 4 for graylag, European, and Chinese domestic geese. Each vertical bar represents one individual with *K* number of colors indicating proportion of ancestry from the inferred clusters, and populations are separated by black vertical line.

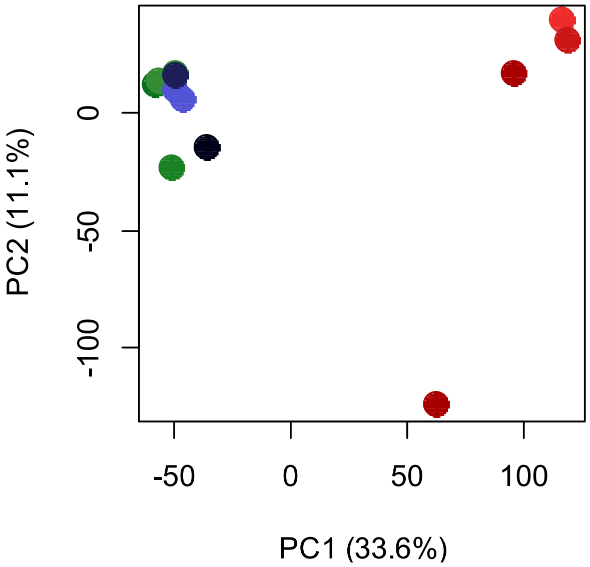


Figure S9. The first two principal components summarizing the genetic variation in geese when sample sizes are equal (percentage explained by each PC is shown). The colors are associated to different groups as follows: graylags (blue), European domestics (green) and Chinese domestics (red).

However, it was obvious that the small sample size was unable to detect all the variation within graylags and European domestic, because when the 58 graylags and 58 European domestic geese were analyzed without the Chinese domestic geese, the groups were clearly separated both in STRUCTURE and PCA (Figure S10 and Figure S11, respectively). Both Evanno and likelihood methods suggested that $K = 3$ was the optimal number of clusters. This was reasonable as the third cluster corresponded well with the Chinese domestic goose ancestry when Figure S10 was compared to Figure 3B and Figure S6.

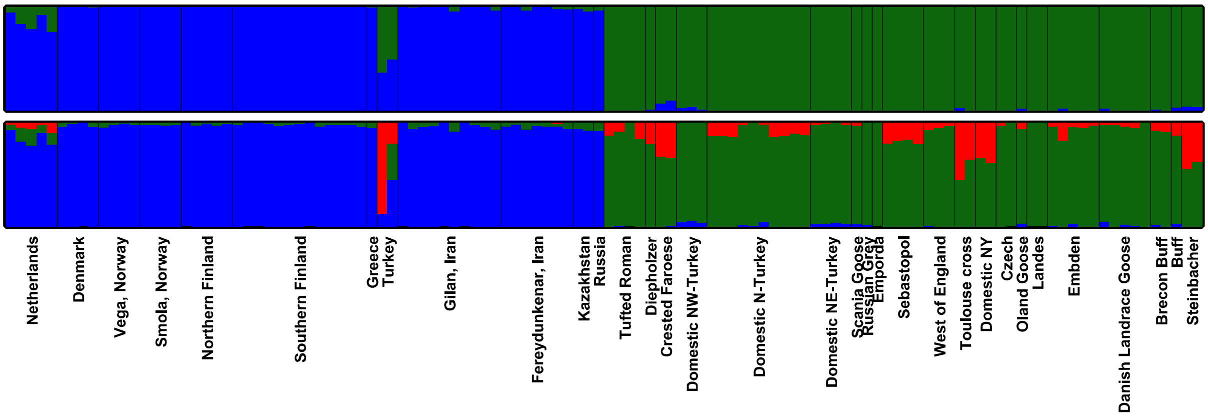


Figure S10. STRUCTURE assignment plots $K = 2$ (above) and $K = 3$ (below) when sample size was 58 graylags and 58 European domestic geese. Each vertical bar represents one individual with K number of colors indicating proportion of ancestry from the inferred clusters, and populations are separated by black vertical line.

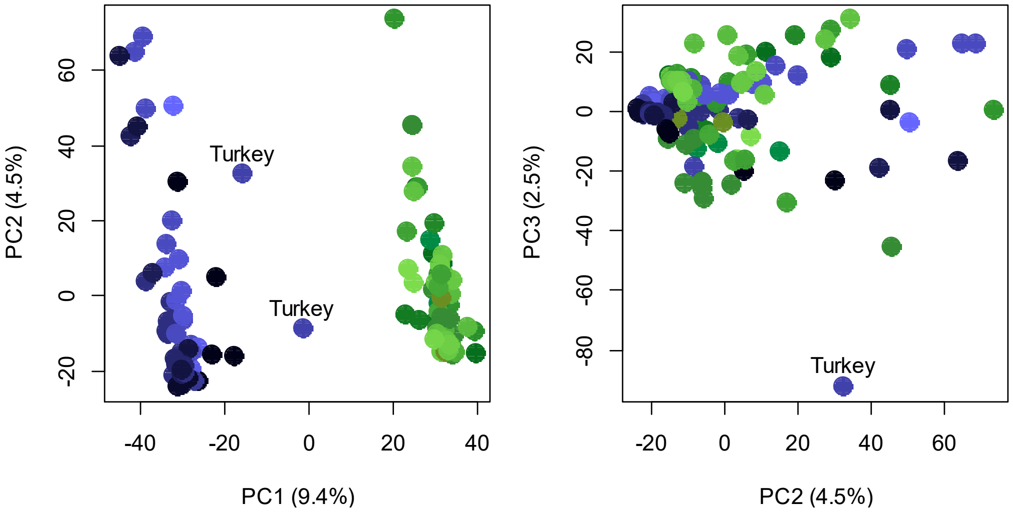


Figure S11. The first three principal components summarizing the genetic variation within wild graylags and European domestic goose when sample sizes are equal. Graylag populations are shown with shades of blue and domestic populations with shades of green. The percentages explained by each PC are shown on the X and Y axes.

Genetic structure of graylags

Among the wild graylags, the Evanno and likelihood methods indicated that the most likely number of clusters in STRUCTURE was four (Figure S12). Since the Turkish graylags showed the highest admixture with the domestic geese in the STRUCTURE analysis of the whole data, the STRUCTURE analysis on graylags was executed also without the Turkish graylags but $K = 4$ was deemed best in both analyses. The only difference between the analyses was that without the Turkish graylags, the Danish and Dutch (along with the eastern graylags) showed admixture with a cluster that was not so prominent in the analysis that included the Turkish graylags. The results suggested that there is some population structure, especially within the Iranian populations, but the major separation appears to be between eastern and western graylags. This was also visible in the PCA analysis where there was a tendency to separate the Western European populations from the eastern graylag populations of Iran, Kazakhstan, and Russia (Figure S13). We found three significant PCs ($p < 0.05$) of which the first PC explained 7.2% of the variation, the second 4.3% and the third 4.1% (Figure S13).

In the neighbor-joining tree eastern and western graylags mostly fall into separate clades (Figure S5). Most Dutch and all Danish graylags formed their own clades whereas all Finnish and most of Norwegian graylags were in the same clade (Figure S5). The Finnish and Norwegian graylags were in different branches of the tree with one individual from Norway being closely related to graylags from SW-Finland. Five samples from an island of Hailuoto in northern Finland formed a subclade that was separate from other Finnish samples that were collected in SW-Finland. The eastern graylags formed a clade with some subclades consistent with the geographical origin of the individuals, but three Norwegian individuals and one Dutch individual also fell into this clade.

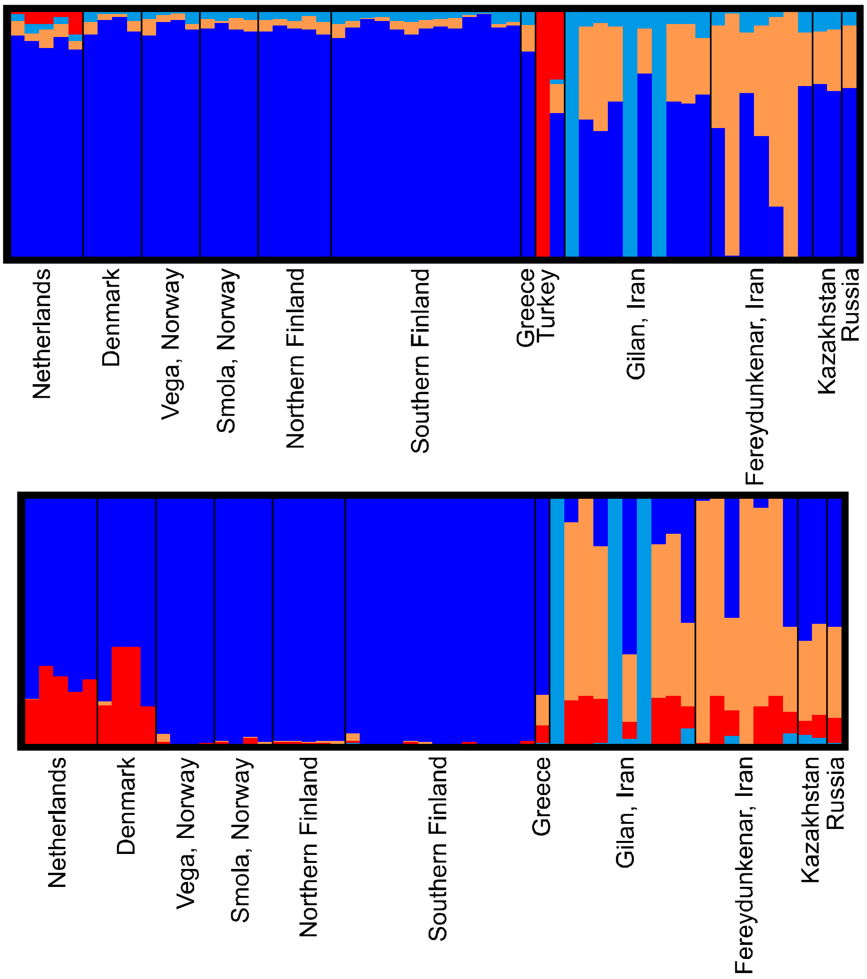


Figure S12. STRUCTURE assignment plots for graylags when *K* = 4. Each vertical bar represents one individual with *K* number of colors indicating proportion of ancestry from the inferred clusters, and populations are separated by black vertical line. The analysis was done with (top) and without (bottom) Turkish graylags because the two Turkish graylags were highly admixed with domestic geese according to STRUCTURE analysis with the whole data and we wanted to assess the impact of their inclusion.

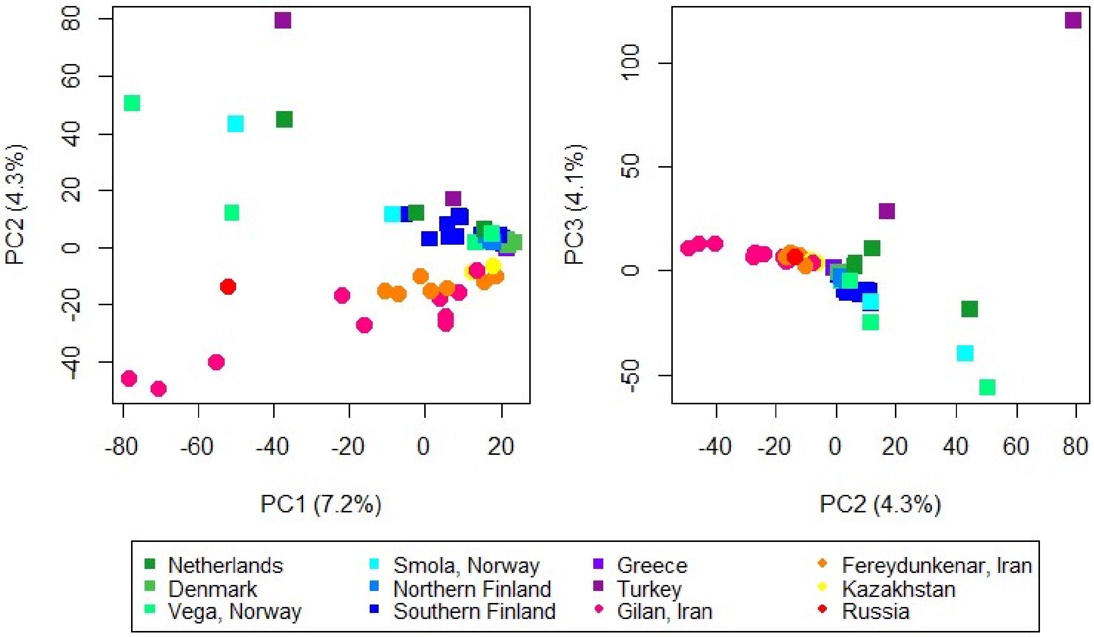


Figure S13. The first three principal components summarizing the genetic variation within wild graylags. European populations are shown with square symbols while the Asian populations are shown with round symbols. The percentages explained by each PC are shown on the X and Y axes.

Genetic structure of domestic geese

Based on STRUCTURE analysis conducted solely on domestic geese, the Evanno and likelihood methods indicated that there are two clusters within domestic geese ($K = 2$), approximately representing the European and Chinese domestic geese (Figure S14). No subsequent split of breeds to different clusters was observed with higher values of K. Nearly all European domestic goose populations showed admixture with Chinese domestic geese. The separation between European and Chinese domestic geese was also confirmed by PCA where the first PC out of seven significant PCs ($p < 0.05$) separated the Chinese and European domestic geese (Figure S15). Moreover, within the European domestic geese, the Turkish domestic geese and mostly purebred European domestic geese formed their own groups (Figure S15). This was also seen in the neighbor-joining tree (Figure S5). According to Food and Agriculture Organization of the United Nations (FAO), the Diepholzer, Steinbacher and Kholmogory breeds are crosses of the two domestic goose types and they all showed admixture proportions with both types of domestic geese, Diepholzer (88.8% European, 11.2% Chinese), Steinbacher (76.6% European, 23.4% Chinese) and Kholmogory (45.8% European, 54.2% Chinese). In the PCA, Diepholzer and Steinbacher fell into the variation within European domestic goose breeds but Kholmogory was halfway between the European and Chinese domestic geese. Even though the STRUCTURE and PCA did not split breeds into separate genetic clusters, there was a trend of individuals of the same breed forming a clade in the neighbor-joining tree (Figure S5). The Turkish domestics also formed subclades with individuals from nearby areas with some exceptions.

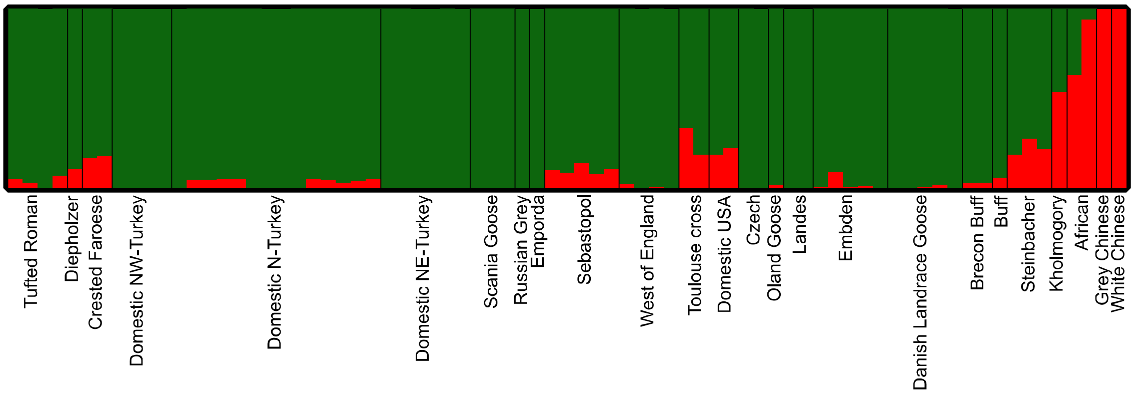


Figure S14. STRUCTURE assignment plot for domestic geese when *K* = 2. Each vertical bar represents one individual with *K* number of colors indicating proportion of ancestry from the inferred clusters, and populations/breeds are separated by black vertical line. Green indicates proportion of European domestic goose ancestry and red indicates proportion of Chinese domestic goose ancestry.

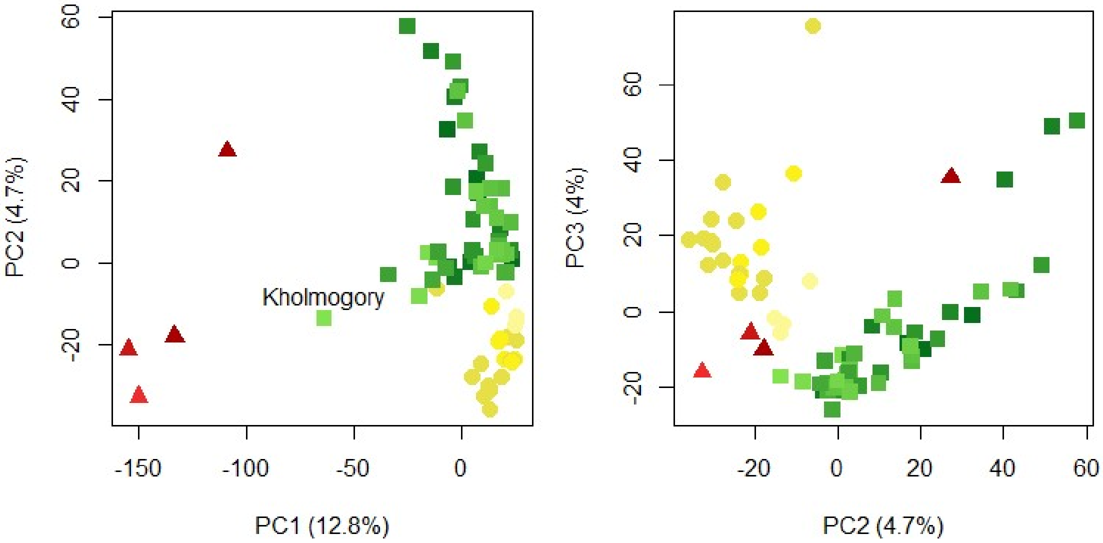


Figure S15. The first three principal components summarizing genetic variation within the domestic geese only. The European breeds are square symbols with different shades of green, Turkish domestics are round symbols with different shades of yellow and Chinese breeds are triangular symbols with different shades of red. The percentages explained by each PC are shown on the X and Y axes.

Genetic diversity

There was a trend that populations with higher admixture proportions from other groups showed higher average H_E and this was visible in both graylag and domestic populations (Figure S16, Table S1). For instance, the average H_E in the Netherlands was 0.157 (13.8% European domestic, 2.2% Chinese domestic) and in 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 graylags 0.146 vs non-pure graylags 0.196). The average H_E was also higher in European domestics that showed high admixture proportions with Chinese domestics, i.e. Crested Faroese (0.117, 17% Chinese domestic), Sebastopol (0.133, 11.1% Chinese domestic), Domestic NY (0.143, 78.6% European domestic, 21.4% Chinese domestic) and Toulouse cross (0.151, 72.9% European domestic, 27.1% Chinese domestic), and the difference was also statistically significant (Welch's t-test, df = 9.0991, p = 0.0039, average H_E: pure European domestics 0.096 vs non-pure European domestics 0.136). 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. Both populations had higher than average H_E (Table S1, Figure S16).

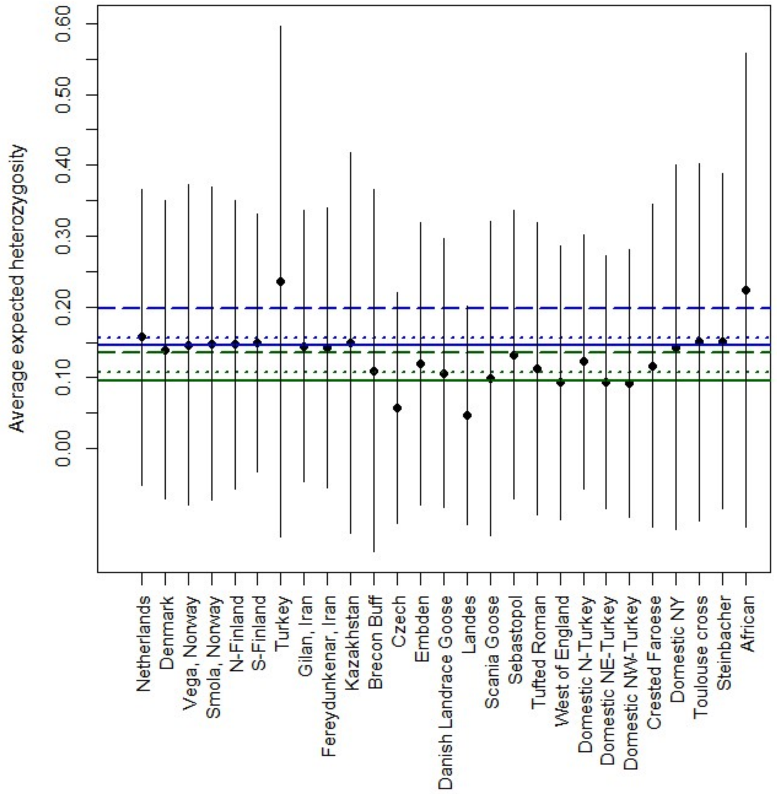


Figure S16. The average H_E estimated for different populations. The blue color represents graylags and the green color domestics. The solid line shows the average for pure populations, dotted line includes all populations within a group, and the dashed line shows the average for non-pure populations.

Table S1. Diversity and ancestry estimate in different graylag populations and breeds of domestic geese. The hybrid status of Diepholzer, Kholmogory and Steinbacher is based on Appendix 1 in Buckland & Guy (2002) Buckland and Guy (2002) (24). The admixture proportions for *K* = 3 were obtained from STRUCTURE.

Population	Status	Sample size	Expected heterozygosity	Graylag ancestry	European domestic ancestry	Chinese domestic ancestry
Netherlands	Wild Graylag	5	0.157	0.8405	0.1375	0.0220
Denmark	Wild Graylag	4	0.140	0.9600	0.0400	0.0000
Vega, Norway	Wild Graylag	4	0.146	0.9596	0.0404	0.0000
Smola, Norway	Wild Graylag	4	0.148	0.9600	0.0400	0.0000
Northern Finland	Wild Graylag	5	0.147	0.9666	0.0334	0.0000
Southern Finland	Wild Graylag	13	0.149	0.9676	0.0324	0.0000
Greece	Wild Graylag	1	NA	0.9370	0.0630	0.0000
Turkey	Wild Graylag	2	0.236	0.4206	0.2345	0.3449
Gilan, Iran	Wild Graylag	10	0.145	0.9536	0.0464	0.0000
Fereydunkenar, Iran	Wild Graylag	7	0.142	0.9448	0.0552	0.0000
Kazakhstan	Wild Graylag	2	0.150	0.9166	0.0834	0.0000
Russia	Wild Graylag	1	NA	0.9086	0.0914	0.0000
Brecon Buff	European Domestic	2	0.110	0.0284	0.9346	0.0370
Buff	European Domestic	1	NA	0.0428	0.8933	0.0640
Crested Faroese	European Domestic	2	0.117	0.0838	0.7466	0.1696
Czech	European Domestic	2	0.058	0.0044	0.9828	0.0128
Danish Landrace Goose	European Domestic	5	0.107	0.0170	0.9680	0.0150
Domestic Northern Turkey	European Domestic	14	0.123	0.0020	0.9562	0.0418

Domestic Northeast Turkey	European Domestic	6	0.094	0.0072	0.9868	0.0060
Domestic Northwest Turkey	European Domestic	4	0.092	0.0222	0.9778	0.0000
Domestic NY	European Domestic	2	0.143	0.0000	0.7862	0.2138
Embden	European Domestic	5	0.120	0.0150	0.9510	0.0340
Emporda	European Domestic	1	NA	0.0000	1.0000	0.0000
Landes	European Domestic	2	0.047	0.0000	1.0000	0.0000
Oland Goose	European Domestic	1	NA	0.0298	0.9422	0.0280
Russian Grey	European Domestic	1	NA	0.0154	0.9846	0.0000
Scania Goose	European Domestic	3	0.099	0.0266	0.9646	0.0088
Sebastopol	European Domestic	5	0.133	0.0008	0.8883	0.1109
Toulouse cross	European Domestic	2	0.151	0.0000	0.7293	0.2707
Tufted Roman	European Domestic	4	0.113	0.0070	0.9490	0.0440
West of England	European Domestic	4	0.093	0.0044	0.9808	0.0148
Diepholzer	European x Chinese Domestic	1	NA	0.0152	0.8704	0.1144
Steinbacher	European x Chinese Domestic	3	0.152	0.0002	0.7626	0.2372

Kholmogory	European x Chinese Domestic	1	NA	0.0042	0.4537	0.5421
African	Chinese Domestic	2	0.224	0.0000	0.2167	0.7833
Grey Chinese	Chinese Domestic	1	NA	0.0000	0.0000	1.0000
White Chinese	Chinese Domestic	1	NA	0.0000	0.0000	1.0000

Admixture

The Turkish population was the only one that obtained negative Z-score (Table S3), but since the two individuals were genetically very dissimilar and unlikely to come from the same population, we analyzed them separately which resulted in non-negative Z-score. However, when we used one of the Turkish graylags as a source, we obtained a negative Z-score in one of the analyses f_3(Embden; Turkey1, Landes) (Table S4). In STRUCTURE this Turkish individual showed that a considerable part of its ancestry originated from the Chinese domestic goose. None of the other graylag populations was either source or recipient of admixture involving domestic geese (Table S5-S6). However, several domestic breeds showed admixture with Chinese domestic geese both in STRUCTURE (Table S1) and the 3-Population test f_3 (Table S5). The f_3 results were quite consistent when several other pure breeds were used as a European source (Table S7).

In line with our other results, the best model to explain our data in simulations made with fastsimcoal2 included gene flow between graylag and domestic geese (Table S8).

Table S5. The Patterson's 3-Population test statistics obtained to test the history of admixture in different graylag populations and breeds of domestic geese.

Source1	Source2	Target	f_3	Standard error	Z-score
Landes	Chinese	Netherlands	0.0098	0.0135	0.727
Landes	Chinese	Denmark	0.0775	0.0141	5.501
Landes	Chinese	Vega, Norway	0.0848	0.0144	5.875
Landes	Chinese	Smola, Norway	0.0575	0.0123	4.686
Landes	Chinese	Northern Finland	0.0598	0.0132	4.550
Landes	Chinese	Southern Finland	0.0389	0.0103	3.768
Landes	Chinese	Greece	0.0706	0.0260	2.715
Landes	Chinese	Turkey1	0.1902	0.0612	3.106
Landes	Chinese	Turkey2	0.1395	0.0758	1.840
Landes	Chinese	Gilan, Iran	0.0621	0.0135	4.603
Landes	Chinese	Fereydunkenar, Iran	0.0696	0.0147	4.749
Landes	Chinese	Kazakhstan	0.0426	0.0135	3.148
Landes	Chinese	Russia	0.3555	0.0425	8.366
Landes	Chinese	Tufted Roman	-0.0543	0.0175	-3.103*
Landes	Chinese	Crested Faroese	-0.0620	0.0278	-2.228**
Landes	Chinese	Domestic NW-Turkey	0.0942	0.0195	4.818
Landes	Chinese	Domestic N-Turkey	-0.0383	0.0156	-2.459**
Landes	Chinese	Domestic NE-Turkey	0.1133	0.0205	5.523
Landes	Chinese	Scania Goose	0.0578	0.0287	2.012
Landes	Chinese	Russian Grey	0.6133	0.1574	3.897
Landes	Chinese	Emporda	0.1028	0.0539	1.909
Landes	Chinese	Sebastopol	-0.1107	0.0182	-6.089*
Landes	Chinese	West of England	0.1028	0.0228	4.500
Landes	Chinese	Toulouse cross	-0.1240	0.0209	-5.921*
Landes	Chinese	Domestic NY	-0.1190	0.0227	-5.242*
Landes	Chinese	Czech	0.2749	0.0511	5.376
Landes	Chinese	Oland Goose	0.3046	0.1280	2.380
Landes	Chinese	Embden	-0.0693	0.0164	-4.223*

Landes	Chinese	Danish Landrace Goose	-0.0041	0.0167	-0.248
Landes	Chinese	Brecon Buff	-0.0357	0.0257	-1.392
Landes	Chinese	Buff	-0.0207	0.0416	-0.497
Landes	Chinese	Diepholzer	0.0732	0.0565	1.296
Landes	Chinese	Steinbacher	-0.1149	0.0215	-5.349*
Landes	Chinese	Kholmogory	-0.1672	0.0187	-8.933*
Landes	Chinese	African	-0.1267	0.0198	-6.399*

* p < 0.01, ** p < 0.05

Table S8. Model selection results and parameter estimates for different demographic models that were tested (see text). Confidence intervals for the best model are shown at the bottom line of the table.

Model	Number of parameters	log L	AIC	ΔAIC	AIC_W	ANCSIZE	N_{DOM}	N_{WILD}	T_1	T_2	$M1_{WD}$	$M1_{DW}$	$M2_{WD}$	$M2_{DW}$
Divergence with changing gene flow patterns	9	-29766.4	59550.7	0	0.99	1112	959	2504	5319	159	$4.25x10^{-4}$	$5.35x10^{-4}$	$1.72x10^{-3}$	$6.69x10^{-4}$
Divergence with gene flow	6	-29787.7	59587.44	36.736	$1.05x10^{-8}$	756	1056	2304	5952		$8.86x10^{-4}$	$7.70x10^{-4}$		
Divergence without gene flow	4	-29919.8	59847.53	296.83	$3.50x10^{-65}$	7312	1047	2576	226					
						378.95-7990.65	833.95-1040.55	2352.4-2680.25	2014.45-6503.75	88.9-476.25	$1.21x10^{-7}$-$6.28x10^{-4}$	$2.88x10^{-4}$-$6.45x10^{-4}$	$1.30x10^{-3}$-$2.23x10^{-3}$	$4.17x10^{-4}$-$8.00x10^{-4}$

ANCSIZE: effective population size of ancestral population. N_{DOM}: effective population size for domestic geese. N_{WILD}: effective population size for graylags. T_1: time of divergence in generations. 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

Figure S5. (separate file) Neighbor-joining tree based on genetic distances between all samples of geese analyzed in this study including the reference genome *A. cygnoid domesticus* breed Zhedong that was used as a reference in SNP calling. Branches leading to graylags are blue, to European domestic geese green and to Chinese domestic geese red. Branches that lead to breeds which are crosses between European and Chinese domestic geese are colored purple. Branches are labelled with a population identifier followed by the graylag, European and Chinese domestic goose admixture proportions from STRUCTURE.

Table S2 (separate file). Pairwise F_{ST} values for each population analyzed in this study.

Table S3 (separate file). The f_3 analysis results for the Netherlands and Turkey. Significant negative Z-scores are in bold.

Table S4 (separate file). The f_3 analysis results for domestic geese. Significant negative Z-scores are in bold.

Table S6 (separate file). The f_3 analysis results for the graylags.

Table S7 (separate file). The f_3 analysis results for domestic geese with Chinese domestic as a source. Significant negative Z-scores are in bold.

References

Buckland, R., and G. Guy (Eds.), 2002 Goose production, in *FAO Animal Production and Health Paper - 154*, Food and Agriculture Organization of the United Nations, Rome.

Chhatre, V. E., and K. J. Emerson, 2017 StrAuto: automation and parallelization of STRUCTURE analysis. BMC Bioinformatics 18: 192.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Earl, D. A., and B. M. VonHoldt, 2012 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4: 359–361.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14: 2611–2620.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. 9: e1003905.

Excoffier, L., and H. E. L. Lischer, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10: 564–567.

Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.

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

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9: e90346.

Goudet, J., 2005 hierfstat, a package for R to compute and test hierarchical F-statistics. Mol. Ecol. Notes 5: 184–186.

van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman *et al.*, 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. USA 108: 1088–1092.

Jakobsson, M., and N. A. Rosenberg, 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801–1806.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Lu, L., Y. Chen, Z. Wang, X. Li, W. Chen *et al.*, 2015 The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver. Genome Biol. 16: 89.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza, 1978 Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792.

Paradis, E., J. Claude, and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

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

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

R Core Team, 2017 R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rosenberg, N. A., 2003 distruct: a program for the graphical display of population structure. Mol. Ecol. Notes 4: 137–138.

Welch, B. L., 1938 The significance of the difference between two means when the population variances are unequal. Biometrika 29: 350.