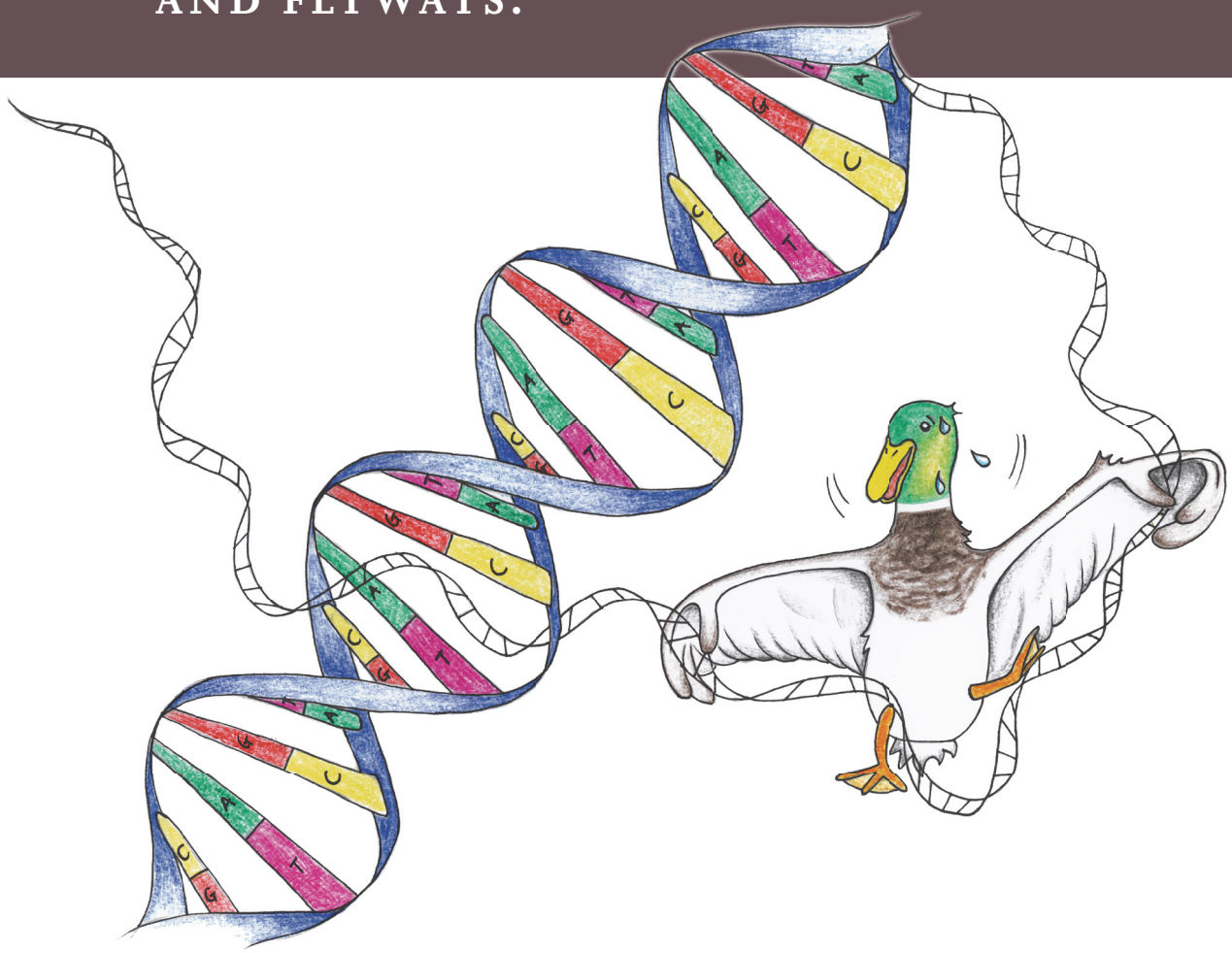
The Role of Mallard (*Anas platyrhynchos*) in the Spread of Avian Influenza:

GENOMICS, POPULATION GENETICS, AND FLYWAYS.



Robert H. Kraus

The Role of Mallard (*Anas platyrhynchos*) in the Spread of Avian Influenza:

GENOMICS, POPULATION GENETICS, AND FLYWAYS.

Robert H. Kraus

THESIS COMMITTEE

THESIS SUPERVISORS

Prof. dr. H.H.T. Prins
Professor of Resource Ecology, Wageningen University

Prof. dr. R.C. Ydenberg
Professor of Fauna Management and Conservation, Wageningen University
Professor of Behavioural Ecology, Simon Frasier University, Canada

THESIS CO-SUPERVISOR

Dr. W.F. van Hooft
Assistant professor, Resource Ecology Group
Wageningen University

OTHER MEMBERS

Prof. dr. B.J. Zwaan, Wageningen University
Prof. dr. C. Dreyer, MPI for developmental Biology, Tübingen, Germany
Prof. dr. C. Schlötterer, University of Veterinary Medicine, Vienna, Austria
Prof. dr. R. Tiedemann, University of Potsdam, Golm, Germany

This research was conducted under the auspices of the C.T. de Wit Graduate School Production Ecology & Resource Conservation

The Role of Mallard (*Anas platyrhynchos*) in the Spread of Avian Influenza:

GENOMICS, POPULATION GENETICS, AND FLYWAYS.

Robert H. Kraus

THESIS

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 December 2011
at 1.30 p.m. in the Aula

Robert H. Kraus
The Role of Mallard (*Anas platyrhynchos*) in the Spread of Avian Influenza:
Genomics, Population Genetics, and Flyways.
143 pages.

Thesis, Wageningen University, Wageningen, NL (2011)
With references, with summaries in Dutch and English

ISBN 978-94-6173-028-2

Table of Contents

CHAPTER 1

General Introduction

Birds, in particular poultry and ducks, are a source of many zoonotic diseases, such as those caused by corona and influenza viruses[1,2]. These viruses are a threat not only to these birds themselves but also to poultry farming and human health, as forms that can infect humans are known to have evolved[1,2]. It is believed that migratory birds, and water birds in particular, play an important role in the global spread of Avian Influenza (AI)[3]. However, it is still debated how large this role precisely is and whether other modes of spread may be more important[2,4]. Migratory birds with separate breeding and wintering grounds are interesting model species for studying disease transmission.

The mallard

The mallard (*Anas platyrhynchos*) is a member of the order of the Anseriformes (ducks, geese and swans)[5,6], and is generally bound to open waters and wetland habitats. Mallards are omnivorous, and their diet consists not only of small invertebrates, which they collect with their bill by "dabbling" under water with their tail up (the characteristic feeding behaviour of the sub-family dabbling ducks, the Anatini[7]), but also tadpoles, small fish, or all sorts of plant material. In the family of ducks (Anatidae) the mallard is the largest bird with a weight of around 1 kg and a body size of 50 cm in both females and males[8]. Mallards display sexual dimorphism: males have bright plumage to display to females during courtship[9], females are dull-brown camouflaged to facilitate secretive life-style, especially during breeding season.

The mallard is the most numerous and well-known waterfowl (Anseriformes) species with a Holarctic distribution. For instance, in Europe it is absent only in January from upland areas and those areas affected by prolonged freezing[10]. Most mallards are migratory without clear geographic directionality, and spring and fall flights can exceed thousands of kilometres[11]. Northern breeding birds are mostly migratory, wintering much further south, while birds breeding in temperate regions (most of Western Europe) are resident or merely dispersive[12]. Additionally, the level of "abmigration" (the switching of flyways) is thought to be very high in ducks species[13] and thus a dense network-like connectivity between populations may exist.

The large-scale migration systems of temperate waterfowl have been extensively studied using ringing, telemetry, morphometrics, radar tracking and isotope analysis[14]. Migration routes are clearly defined[12] for most waterfowl species and usually follow north-south directions, with populations travelling between northerly breeding areas and more southerly non-breeding areas. Many species follow similar routes and decades of studies on bird migration have led to the delineation of major waterfowl "flyways"[12,15-17]. In North-America especially, flyways are managerial units created by agreements between adjoining states and provinces, and thus are bounded by management requirements. In a population ecological sense true migratory pathways are much fuzzier. Populations and individuals within species may occupy different flyways, and many migrants are flexible in migration routing[18], even though these migration routes have been in place for relatively long periods of time[19,20]. Mallards seem to display an extreme flexibility in their migration behaviour, when compared to the related species in their bird order, and it is unclear if biological reality would support their placement into a global waterfowl flyway system. Migration in mallards is heavily studied but delineations of populations have been more tentative than in other waterfowl[12,16,17].

The foraging habitat of the mallard in shallow waters brings it into contact with a wide variety of pathogens and it may act as a reservoir and disperser for many of them. The mallard is considered the primary natural reservoir of avian influenza due to its wide range and large population sizes[21]. Moreover, together with the black-headed gull (*Larus ridibundus*) it was identified as the species bearing the highest risk to transmit AI to farm-birds due to frequent contact with poultry[10]. Post-breeding pre-migratory staging areas are thought to be important locations for viral acquisition[22].

Avian Influenza

The Influenza viruses are a genus within the Orthomyxoviridae. They have a segmented negative sense single-stranded RNA-genome [(-)ssRNA]. Forms of the Influenza A virus can infect a wide range of host species including among others birds, pigs, horses, seals, whales and humans. There are eight RNA segments. Viral subtypes are classified using two of the encoded genes: the hemagglutinin (HA) gene and the neuraminidase (NA) gene. These genes code for surface proteins that play a key role in host recognition and initial infection[23]. Sixteen HA-types[24] and nine NA-types are recognised, giving rise to 144 (= 9 × 16) possible subtypes. These are described as, e.g., H5N1 (type 5 of HA, and type 1 of NA). The classification used to rely on immunoassays using standard procedures[25-27]. Nowadays it is also possible to sequence the genes of a virus using retrograde transcriptase PCR (RT-PCR) with a set of universal primers for all genes and all subtypes[28] and compare the obtained cDNA sequence with databases such as GenBank[29].

Humans can be infected with all three species of Influenza viruses: A, B and C. Influenza A viruses are those referred to as Avian Influenza. In humans, the HA proteins H1-3 and NA types 1 and 2 of Influenza A viruses usually cause seasonal Influenza, but Influenza B and, less frequently, Influenza C viruses also circulate in humans[30]. However, human pandemics for which the medical histories have been reconstructed were all caused by Influenza A viruses[31,32].

It has been proposed that wild birds in general, and especially migratory waterfowl, are not only the reservoir, but also the vectors for the spread of Avian Influenza over large distances[2]. Highly pathogenic types of AI are thought to be easily spread through the flyways of waterfowl, because of these birds show no clinical signs of infection but transmit a form highly pathogenic to poultry or humans[33-36], although this is not generally true for every highly pathogenic strain[37-39]. However, the evidence for this possible route of transport in cases of highly pathogenic strains is equivocal[40-43]. Alternatively, highly pathogenic strains might evolve from low pathogenic ones directly in poultry farms where the selective regime is different from the wild[44].

The molecular ecological approach

The scientific discipline of 'molecular ecology' is characterised by the use of a certain type of data: information about the genetic constitution of an organism. Of course, in molecular ecological studies a whole range of biological data may be collected, as diverse as behavioural observations[45], physiological parameters[46] or abiotic conditions, such as geographical coordinates[47], landscape features[48] or physical parameters[49] – but the commonality between all such studies is that genetic information is employed to answer ecological questions in a broad sense. Since information

about the whole genome of an organism, or from a sample of several individuals, is difficult to obtain, molecular ecologists seek to use representative sets of genetic markers, either targeted at specific regions in the genome[50] or selected to represent the history of a population or species as neutrally as possible (i.e., not impacted by natural selection[51]). Analyses of genetic markers predate molecular techniques such as DNA sequencing[52] or polymerase chain reaction (PCR[53]) which are routinely used nowadays in laboratories throughout the world. Genetic markers can actually already be typed from phenotypic traits, such as colour variation[54], and used for genetic studies if they behave in a Mendelian fashion[55]. However, in the late 19th and first half of the 20th century the discovery of the macromolecule DNA by Friedrich Miescher in 1869, documented in a series of reports reviewed by Dahm[56], resolution of its chemical[57] and structural[58] characteristics, and finally the recognition of its role as the agent of inheritance[59,60] sounded the bell for the development of molecular DNA markers. Since the first studies, which were carried out on polymorphisms in the products of genes, the proteins[61], an arsenal of direct measures of variation in DNA has been developed[62].

Mallards have been investigated using molecular ecological tools, but rarely on a continental or global scale. The sole example of a continent-wide mitochondrial DNA (mtDNA) study shows that ecological estimates of extensive dispersal among mallard populations are confirmed by genetic data. This recent phylogeographic study showed little genetic structure within Russia, with only 0.18% of the total mtDNA variation between western and eastern Russia[63]. However, except for western Russia this study did not include European populations. In contrast to populations from western Russia and Asia many European populations are resident, which is expected to increase the degree of population genetic structuring because movements of individuals are more limited (reduced gene flow). Indeed, it has been concluded in ecological studies that the existence of differences in population structure between (western) Europe and Asia is plausible[12]. It is not clear what the possible implications of this pattern would be for the spread of a pathogen such AI between western Russia and Europe.

The possibility of spread of AI by the mallard can be studied by population genetic analyses of the host species, and this thesis focuses on patterns in the mallard. A molecular ecological tool to investigate allele frequencies could be the use of single nucleotide polymorphisms (SNPs). SNPs are comprised of nucleotide positions in the genome of an organism where there is a stable polymorphism within a population. It is formed by two alleles, one of which has a frequency of at least 1%[64]. The use of SNPs in molecular ecology has been strongly proposed by several authors (for instance Morin *et al.*[65]) because of their superior features compared to microsatellites, such as known and predictable mutation behaviour, high abundance throughout the whole genome, and the ease of comparing or combining data from different laboratories.

Outline of the thesis

The overall aim of this thesis is to study the mallard *Anas platyrhynchos*, which is possibly the most important vector for Avian Influenza[10,66], by genetic tools, with respect to its movement ecology, geographic structuring, and evolution. In Chapter 2 I conduct a study with mtDNA sequences to generate a first draft of a world-wide picture of population genetic structure in the mallard.

Further, I am able to model the post-glacial demographic history as well as to infer current population sizes of the mallard. With my mtDNA sequences I confirm a previous mtDNA study[63] that shows hardly any genetic structure within continental land masses. Hence, I developed an additional nuclear genetic marker set which I describe in Chapter 3. There, I present a generally applicable improved analysis pipeline to develop genome-wide SNP sets for non-model organisms In mallards, this resource of >100,000 SNPs is carefully evaluated and a sub-set of 384 SNPs for large-scale genotyping is devised. I genotyped nearly 1,000 ducks, mainly mallards but also other duck species, with this mallard SNP set. In Chapter 4 I deal with the refinement of the findings from Chapter 2, concerning global genetic population structure. In Chapter 5 I report on a study of the evolutionary genetic history of speciation and hybridisation in the duck genus *Anas*, to draw further conclusions about the mallard's population sizes as well as to offer new hypotheses concerning the mallard's and other waterfowl species' ability to live co-adapted to Avian Influenza. Mallards' and other *Anas*-ducks' whole continental to global distribution brings them together in sympatry. I show that a combination of sympatric distribution, conflicting genetically determined and learned mate recognition mechanisms, and genomic compatibility between species helps to explain the long-standing puzzle of waterfowl hybridisation and introgression of genes from one duck species into another. I propose that this fact can be part of the explanation as to why ducks are so adaptable and successful, as well as why they show extraordinary abilities to withstand AI infections, or its consequences for health status. From my findings of extensive omni-directional gene flow I conclude that surveillance programs cannot be optimised by targeting specific migration routes, but rather that broad sampling is needed. To address infrastructural challenges emerging from this claim, I offer pilot studies on the feasibility of FTA cards[67] for solving biosafety, transportation and analysis issues in Chapters 6 and 7. Finally, the synthesis chapter (Chapter 8) reviews the results of all previous chapters and creates links among them to describe the world's mallard population in its various facets, from phylogeography to migration biology, from population genomics to evolution.

CHAPTER 2

Evolution and connectivity in the world-wide migration system of the mallard: Inferences from mitochondrial DNA

Robert HS Kraus, Anne Zeddeman, Pim van Hooft, Dmitry Sartakov,
Sergei A Soloviev, Ronald C Ydenberg, Herbert HT Prins

Article submitted for publication

Abstract

BACKGROUND

Main waterfowl migration systems are well understood through ringing activities. However, in mallards (*Anas platyrhynchos*) ringing studies suggest deviations from general migratory trends and traditions in waterfowl. Furthermore, surprisingly little is known about the population genetic structure of mallards, and studying it may yield insight into the spread of diseases such as Avian Influenza, and in management and conservation of wetlands. The study of evolution of genetic diversity and subsequent partitioning thereof during the last glaciation adds to ongoing discussions on the general evolution of waterfowl populations and flyway evolution. Hypothesised mallard flyways are tested explicitly by analysing mitochondrial mallard DNA from the whole northern hemisphere.

RESULTS

Phylogenetic analyses confirm two mitochondrial mallard clades (female proportion of the population). Genetic differentiation within Eurasia and North-America is low, on a continental scale, but large differences occur between these two land masses ($F_{ST} = 0.51$). Half the genetic variance lies within sampling locations, and a negligible portion between currently recognised waterfowl flyways, within Eurasia and North-America. Analysis of molecualr variance (AMOVA) at continent scale, incorporating sampling localities as smallest units, also shows the absence of population structure on the flyway level. Finally, demographic modelling by coalescence simulation proposes a split between Eurasia and North-America 43,000 to 74,000 years ago and strong population growth (~100fold) since then and little migration (not statistically different from zero).

CONCLUSIONS

Based on this first complete assessment of the mallard's world-wide population genetic structure we confirm that no more than two mtDNA clades exist. Clade A is characteristic for Eurasia, and clade B for North-America although some representatives of clade A are also found in North-America. We explain this pattern by evaluating competing hypotheses and conclude that a complex mix of historical, recent and anthropogenic factors shaped the current mallard populations. We refute population classification based on flyways proposed by ornithologists and managers, because they seem to have little biological meaning. Our results have implications for wetland management and conservation, with special regard to the release of farmed mallards for hunting, as well as for the possible transmission of Avian Influenza by mallards due to migration.

Background

The large-scale migration systems of temperate waterfowl (Anseriformes:Anatidae) have been extensively studied using ringing, telemetry, morphometrics, radar tracking and isotope analysis[14]. In general, migration routes are clearly defined[12]. Most have north-south bearings, with populations travelling between northerly breeding areas and more southerly non-breeding areas. Many species follow similar routes and decades of studies on bird migration have led to the delineation of major waterfowl 'flyways'[12,16,17]. Especially in North-America, flyways are managerial units

created by agreements between adjoining states and provinces, and thus are bounded by management requirements. The boundaries of true migratory pathways (in a population ecological sense) are much fuzzier. Populations and individuals within species may occupy different flyways, and many migrants are flexible in migration routing[18], even though these migration routes have been in place for relatively long periods of time[19,20]. Especially in ducks, 'irregularities' in migration routes have been described, such as individuals switching migratory trajectories, termed 'abmigration'[68] or 'flyway permeability'[69].

The mallard (*Anas platyrhynchos*) is the most numerous Holarctic waterfowl species and distributed widely over the whole Northern Hemisphere. Northerly breeding birds are mostly migratory, wintering much further south, while birds breeding in temperate regions, especially in parts of Western Europe, can be resident[12]. Migratory mallard populations often do not exhibit clearly defined routes, even when breeding and non-breeding destinations are thousands of kilometres distant[63]. Mallards play a significant role in the management and conservation of wetland habitats[70], are a very common bird in recreational wetland parks as well as the major game species in wetland systems. Hunting mallards is facilitated in many countries by supplementary restocking wild populations with farmed mallards[71], possibly with large-scale consequences for the population genetic structure, genetic integrity and fitness of the wild populations[72].

Due to its wide range and large population sizes, the mallard is considered the primary natural reservoir of avian influenza (AI)[21]. It has been identified (along with the black-headed gull *Larus ridibundus*) as the species posing the highest risk to transmit AI to farm-birds[10]. Mallards may (among other bird species) contribute to the spread of AI from Eurasia (Old World; OW) into North-America (New World; NW). Some studies propose trans-hemispheric movements of dabbling ducks between these land masses[73] and that these facilitate AI transmission[74,75] (but see [43,76]). The mallard is thus a prime candidate for spreading AI in the wild[66], and perhaps to humans, and it is important to learn more about its ecology, movements and dispersion[2,23]. Recently it became easier to sample AI from wild birds and ecology and virology should be integrated into each other further[77,78].

Analyses of mitochondrial DNA (mtDNA) show just two genetic clades: clade A, mainly found in Asia, and clade B, found in North-America[63,79], indicating widespread mixing within but not between the Old and New World. However, to date no genetic studies have been carried out for the complete native range of mallard, as data from Europe and eastern North America were missing. A range-wide survey of genetic diversity and connectivity between proposed mallard flyways could thus be useful to finally generalise findings of small-scale population genetic and phylogeographic studies. Such a study would also add substance to an increasing body of literature on migration systems, flyways concepts, and genetic population structure in other waterfowl.

In this study, we attempt to close existing gaps in the knowledge of the genetic structure of female mallards by providing large numbers of mtDNA sequences from previously unsampled regions throughout the whole native distribution of the mallard. Due to the mallard's high mobility and examples of flyway permeability in closely related duck species, we hypothesise weak barriers to dispersal between Asia and Europe. Thus, in Europe we expect to find clade A haplotypes, as in Asia, and Europe and Asia to form a joint Eurasian 'Old World' meta-population, with

possibly extensive east-west gene flow. Additionally, we present many more samples from central and eastern Canada, which lacked in the most recent analysis[63], to complete an mtDNA sequence data set comprising samples from the whole Holarctic. A combination of population genetic analyses and coalescent simulations in an isolation-with-migration framework[80,81] enables us in this to measure genetic diversity and geographic partitioning thereof, as well as population histories and migration rates between Old World and New World mallards, and to test the currently hypothesised flyways of mallards (see Table 2.1 and refs [12,16,17]).

Results

MTDNA CONTROL REGION SEQUENCING

Our data set is comprised of mtDNA sequences from 346 mallard ducks around the world consisting of 195 newly sampled mallards and 151 sequences from previous studies[63,82] (for details on the sequences and localities see Table 2.1 and methods section). 155 different haplotypes were found in this data set, of which 101 were already contained in the data set of Kulikova *et al.*[63,82], and 54 were novel. Of the 622 aligned nucleotide positions 93 (15%) were variable, 73 (11%) sites were parsimony informative and four sites showed gaps. The additional file "haplotypelist.xls" lists the haplotype names with corresponding sample IDs. Of the 155 haplotypes, 108 haplotypes were only represented by a single individual, 44 haplotypes by 2-7 individuals, one haplotype by 16 individuals (Hap 56), one haplotype by 20 individuals (Hap 57), and the most frequent haplotype (Hap A7) was represented by 71 individuals (Figure 2.1). This A7 haplotype is found in both Old and New World, in almost all localities. All new sequences obtained in this study are deposited in GenBank[29] with accession numbers JN029963-JN030157.

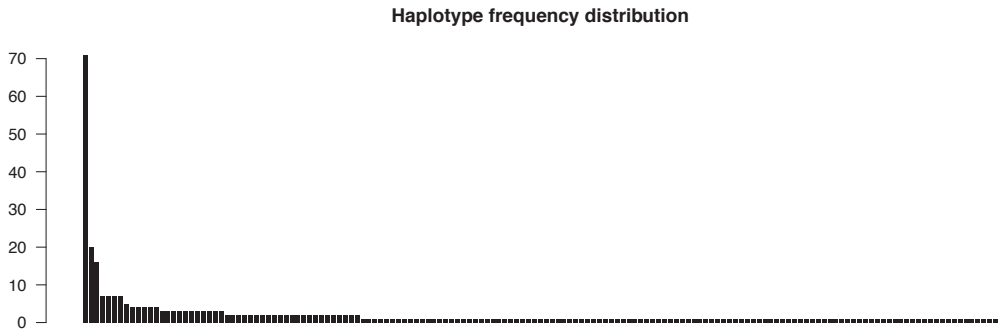


FIGURE 2.1: ***Frequency distribution of mtDNA control region haplotypes.***
Haplotypes are sorted on the x-axis descending by their frequency in the full control region data set. Haplotype A07, for instance, has the highest count (71 times), followed by haplotype 57 (20 times). See additional file "haplotypelist.xls" for details.

In Table 2.1 we further list nucleotide diversity (π) and haplotype diversity (dH). Nucleotide diversity is similar in all Old World flyways (π between 0.003 and 0.005, and 0.008 in Eastern Asia) but consistently higher in the New World ($\pi > 0.013$). Samples from the Aleutians have

TABLE 2.1: *Sampling localities and genetic diversity.*

land mass[1)]	region	flyway[2)]	π (± SD)	dH (± SD)	locality	*N*	lat.	long.
OW	Europe	North-West EU	0.00344	0.696	GBAB	21	57.433	-2.393
			(0.00038)	(0.05)	NLFR	23	53.035	5.574
					NOBE*	22	60.355	5.345
		Central EU	0.00470	0.929	ATHO	18	48.607	16.905
			(0.00034)	(0.018)	DEWU	24	50.036	11.972
					EETA*	22	58.345	27.154
					RUKR	10	60.999	38.556
		East EU	0.00536	1.000	RUAS	3	46.217	47.767
			(0.00182)	(0.272)				
	Asia	Central Asia	0.00438	0.833	KZAO	3	44.893	75.122
			(0.00112)	(0.127)	RUOM*	6	55.845	71.853
		East Asia	0.00835	0.987	MNDA	1	47.000	119.367
			(0.00093)	(0.005)	RUKK*	4	50.388	136.996
					RUMA	1	59.631	149.115
					RUPR	82	45.007	132.432
NW	Alaska	Pacific NA	0.01323	0.985	USFI	13	64.825	-147.584
			(0.00151)	(0.015)	USIZ	6	55.358	-162.728
					USJU	1	58.364	-134.572
					USKB	2	60.545	-151.148
					USKI	4	57.491	-153.495
					USSF	1	61.216	-149.884
					USYD	1	61.367	-163.717
					USYR	1	65.821	-149.733
	Canada	Central NA	0.01364	0.971	CARM	20	50.628	-101.159
			(0.00175)	(0.021)	CASL*	3	49.666	-112.704
					USPI	1	72.677	-99.469
		Atlantic NA	0.01430	0.934	CACO	3	45.579	-64.345
			(0.00224)	(0.061)	CAJC*	2	42.321	-82.385
					CALM	9	43.962	-80.400
N/A	Aleutians	N/A	0.00870	0.824	USAD	2	51.762	-176.622
			(0.00243)	(0.084)	USAI	8	52.905	172.906
					USSI	7	52.723	174.112
	Greenland	N/A	0.00042	0.177	GLNU	22	64.190	-51.708
			(0.00027)	(0.106)				

[1)] land masses are defined as Old World (OW, which is Eurasia) and New World (NW, which is the North-American continent). The Aleutian Islands and Greenland are not assigned to either of these land masses

[2)] flyway definitions based on references 2-4

π is nucleotide diversity, with standard deviation (± SD), dH is haplotype diversity, N gives the sample size of each locality, geographical latitude and longitude are abbreviated lat. and long.

*coordinates are averages of several near-by places, where ducks have been sampled and combined into one sampling locality. Full details for each individual can be found in the supplementary file "samples.xls".

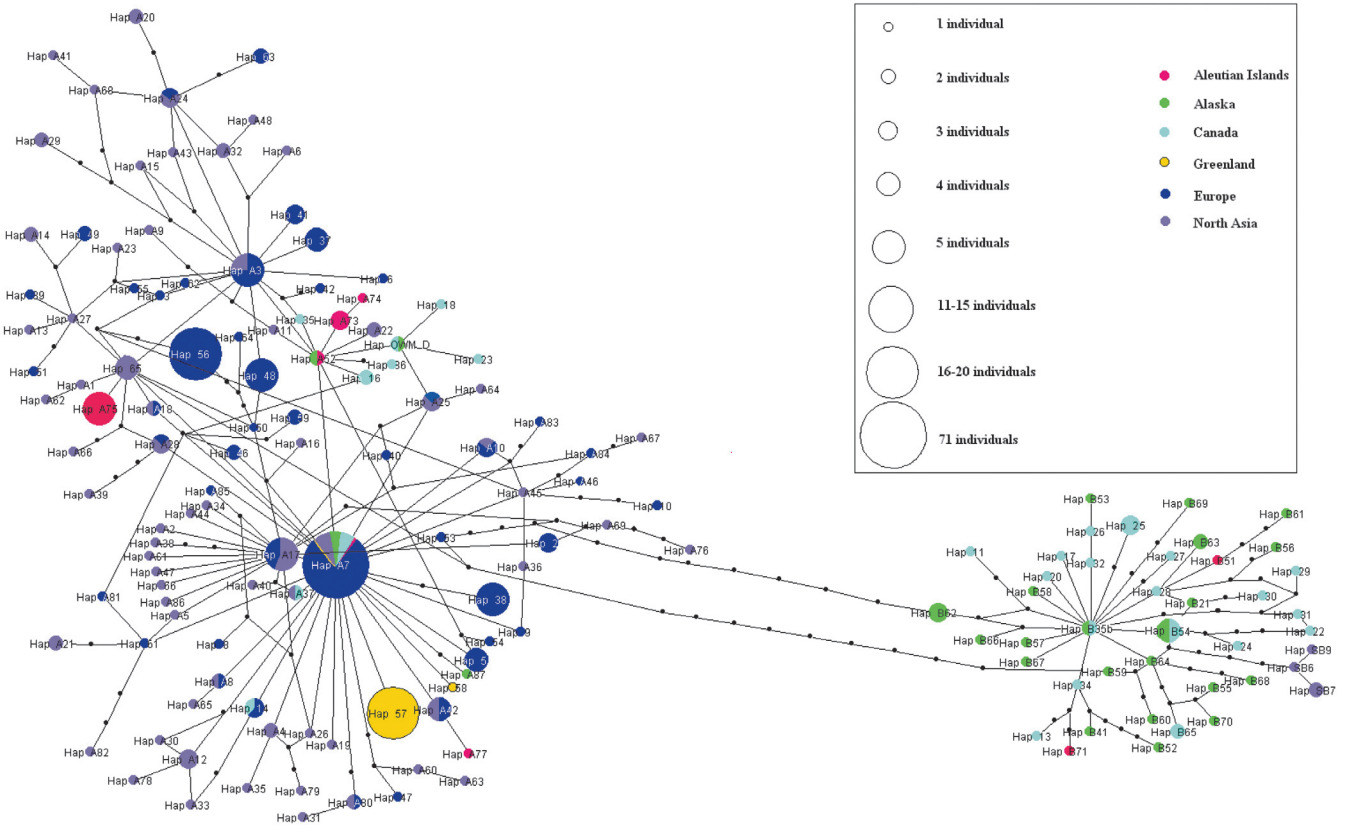


FIGURE 2.2: *Unrooted network illustrating the phylogenetic relationships of the mtDNA haplotypes.*
Haplotype names correspond to the ones in additional file "haplotypelist.xls" and colours to regions as indicated in Table 2.1. Circle sizes scale to indicate the number of individuals harbouring each haplotype. Small black dots indicate unsampled intermediate haplotypes. Note that distances between circles are not proportional to the genetic distance but are arranged for better visibility.

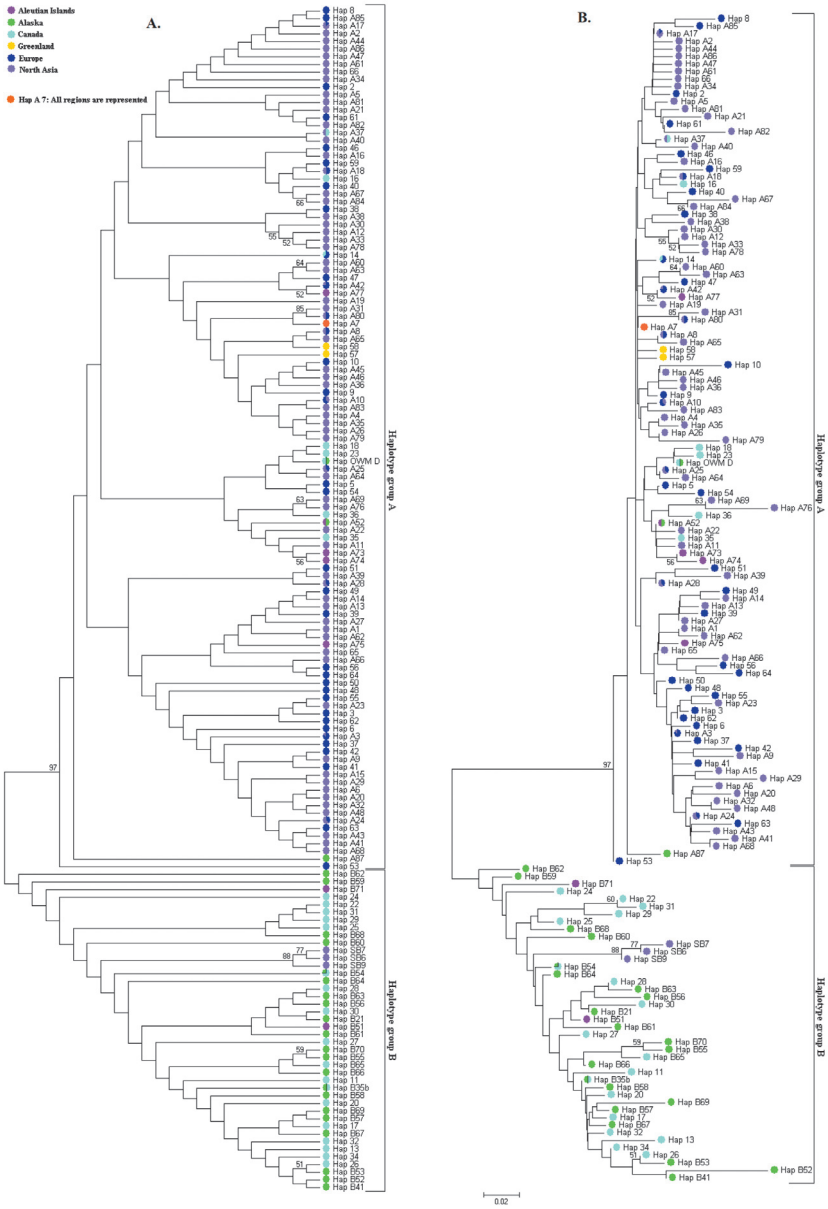


FIGURE 2.3: ***Unrooted Neighbour Joining tree illustrating clade A and B mtDNA control region haplotypes.*** Bootstrap values (500 replicates) are shown next to the branches if > 50%. Positions containing gaps are removed by pairwise deletion. Haplotype names and colours of the dots correspond to the ones in Figure 2.2. A: Tree ignoring branch lengths. B: Branch lengths scaled to evolutionary distances (sum of branch lengths = 2.65).

intermediate nucleotide diversity somewhere between Old and New World ($\pi = 0.009$). The Greenland samples are very low in both nucleotide diversity ($\pi = 0.0004$) and haplotype diversity ($dH = 0.2$), whereas haplotype diversity is high ($dH > 0.8$) in all other flyways except in the North-West Europe flyway ($dH = 0.7$).

Testing neutrality based on the frequency spectrum of the haplotypes (Figure 2. 1) with Tajima's D[83] and Fu's FS[84] revealed negative test statistics for both measures when samples were partitioned into Eurasia ($D = -2.03$, $p = 0.001$; $FS = -25.52$, $p < 0.001$) vs. North-America ($D = -0.68$, $p = 0.272$; $FS = -24.63$, $p < 0.001$), but Tajima's D was not significant for North-America. Splitting Eurasia into Europe ($D = -1.85$, $p = 0.007$; $FS = -26.41$, $p < 0.001$) and Asia ($D = -1.93$, $p = 0.006$; $FS = -25.38$, $p < 0.001$) leads to statistics significantly smaller than zero for both continents.

TWO DISTINCT MALLARD CLADES

An unrooted haplotype network (Figure 2.2) shows the presence of two distinct clades. These clades correspond well to a classification of sampling locations into Old World (clade A haplotypes, mainly Eurasia) and New World (clade B haplotypes, mainly North-America). Within these clades, a great variety of haplotypes was sampled with relatively few missing intermediate ones, represented by small black dots in Figure 2.2. Clade A and clade B are separated by about ten substitutions. Phylogenetic inference by building a Neighbour-Joining[85] tree supported the split between clade A and clade B haplotypes with 97% of the bootstrap replicates (Figure 2.3).

TABLE 2.2: *Pairwise F_{ST} values for all flyways.*

	1	2	3	4	5	6	7	8	9	10
1) Pacific NA	–			*	*		*	*	*	*
2) Central NA	–0.01	–		*	*		*	*	*	*
3) Atlantic NA	0.19	0.15	–	*	*			*	*	
4) North-West EU	**0.65**	**0.63**	**0.33**	–	*			*	*	*
5) Central EU	**0.62**	**0.60**	**0.28**	**0.08**	–				*	*
6) East EU*	0.44	0.39	0.03	0.05	–0.09	–			*	
7) Central Asia	**0.50**	**0.47**	0.16	0.08	0.00	–0.07	–		*	
8) East Asia	**0.52**	**0.49**	**0.16**	**0.06**	0.01	–0.12	–0.01	–	*	
9) Greenland	**0.64**	**0.64**	**0.41**	**0.36**	**0.30**	**0.71**	**0.54**	**0.18**	–	*
10) Aleutians	**0.43**	**0.40**	0.10	**0.17**	**0.09**	–0.06	0.05	0.05	**0.37**	–

Above the diagonal we indicate statistical significance (after Bonferroni correction), below the diagonal F_{ST} values are shown. Those printed bold-face are statically significant.

* note that the east European flyway is only represented by three samples, and hence the presented values are very crude!

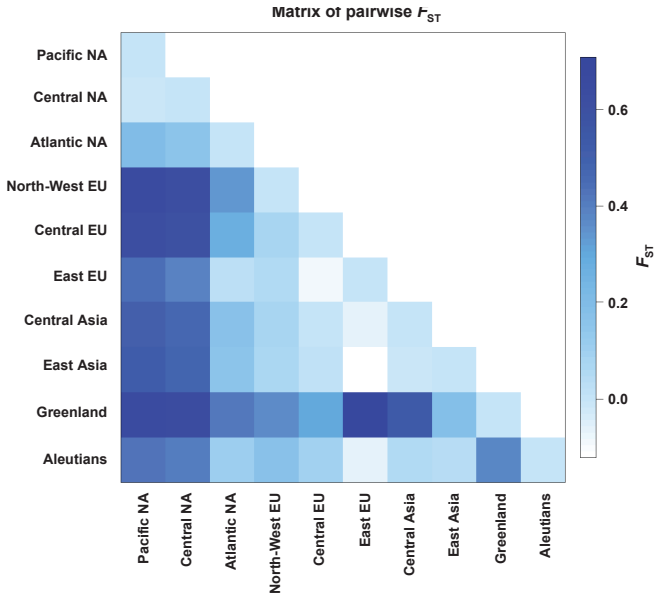


FIGURE 2.4: *Heat map visualising pairwise F_{ST} relationships between the mallard flyways.* Darker shades of blue indicate higher values of F_{ST}. The matrix corresponds to the numerical values given in Table 2.2.

LOW GENETIC DIFFERENTIATION WITHIN EURASIA AND NORTH-AMERICA, BUT LARGE DIFFERENCES BETWEEN THEM

Genetic differentiation between Eurasia and North-America (Aleutians and Greenland excluded, see methods section) and between all individual flyways (flyways definitions are in Table 2.1) was assessed by Wright's *F*-statistics[86]. The F_{ST} value between the two land masses was 0.51. Pairwise F_{ST} values on the flyway-level are given in Table 2.2 and visualised in Figure 2.4. Not all flyways were significantly differentiated from each other. In case of the Eastern Europe flyway sample size was too low (N=3) to make any meaningful statement about the amount of differentiation. None of the intra-North-American flyways was significantly differentiated from the others on that continent. The same was found between the two Asian flyways. In Europe, however, with the exception of the Eastern European flyway which suffered from low sample size, we observed significant structure, albeit low in magnitude (North-West vs. Central Europe: $F_{ST} = 0.08$). Notably, the Aleutian sample was most differentiated from the North-American Pacific and Central flyways but to a lesser extent from the Eurasian ones (in some cases not even significantly; Table 2.2). Finally, the Greenland sampling location displayed relatively high and significant differentiation with each of the other flyways.

An analysis of molecular variance (AMOVA) revealed little genetic variation within the various Eurasian or North-American flyways compared to the amount of variation between these two

land masses. 50.2% of the genetic variation lies between North-America and Eurasia, and 46.6% within the sampled localities. Only 3.2% of the genetic variation was partitioned between flyways within the continents (Table 2.3). Performing an AMOVA in which localities are not pooled into flyways does not lead to qualitative differences of this outcome (Table 2.3; values in brackets).

TABLE 2.3: *AMOVA analysis of flyway genetic variance in the land masses Eurasia and North America.*

		d.f.		SS		VC		% var
Between land masses	1	(1)	288.03	(288.03)	2.67	(2.66)	50.20	(49.94)
Between flyways (localities) within land masses	6	(26)	50.71	(168.71)	0.17	(0.43)	3.22	(8.1)
Within flyways	299	(279)	740.26	(622.74)	2.48	(2.23)	46.57	(41.96)
Total	306		1079		5.32		100	

Values in brackets were obtained when treating sampling localities directly as groups (i.e., not pooled into flyways). All variances are statistically significant ($p<0.05$).

Abbreviations: d.f.: degrees of freedom, SS: sum of squares, VC: variance components, % var: percentage of variation.

When we assigned localities to flyways within continents, per-continent AMOVA showed that genetic variation was to a large extent partitioned within localities: 85% in North-America, 76.8% in Europe, 98.7% in Asia and if combined into one land mass: 87.7% in Eurasia. In none of these analyses there was a significant amount of genetic variation between flyways ($p>0.2$). In Europe, there was a significant ($p < 0.001$) genetic variance component of 22.9% between the localities within the flyways, in the other continents this was not statistically significant.

DEMOGRAPHIC HISTORY AND GENE FLOW

The runs of the coalescent simulation program IMa2 converged quickly in an optimal parameter space. Effective sample sizes for all parameters exceeded 300,000 and each newly sampled genealogy was independent from the previous one (autocorrelations in all parameters were 0.05 and less). Pairwise correlations of parameters were between -0.1 and +0.1, except between q0 and q1 (the estimates of population size for Old World (OW) and New World (NW)) where it was -0.19.

All parameter estimates had relatively sharp unimodal posterior density distributions (Figure 2.5), except the migration rates which were poorly estimated, with wide ranges covering three orders of magnitude ($m_{OW\rightarrow NW}$ and $m_{OW\rightarrow NW}$: $5.1\times10^{-9} - 4.7\times10^{-6}$ and $5.1\times10^{-9} - 1.8\times10^{-6}$, see Figure 2.5). Table 2.4 summarises the results in demographic units. Although 95% highest posterior densities (HPD95) for population sizes of OW and NW samples overlapped IMa2 calculated that there was a 66 % probability that the effective population size of the NW mallards (1.3 – 4.2 million) is larger than the OW population (1.5 – 2.9 million). Further, even though the HPD95 ranges of the migration rates peaked almost at zero, there was a 76% probability that the migration rate from OW into NW is higher than the other way around in the few cases in which it may occur.

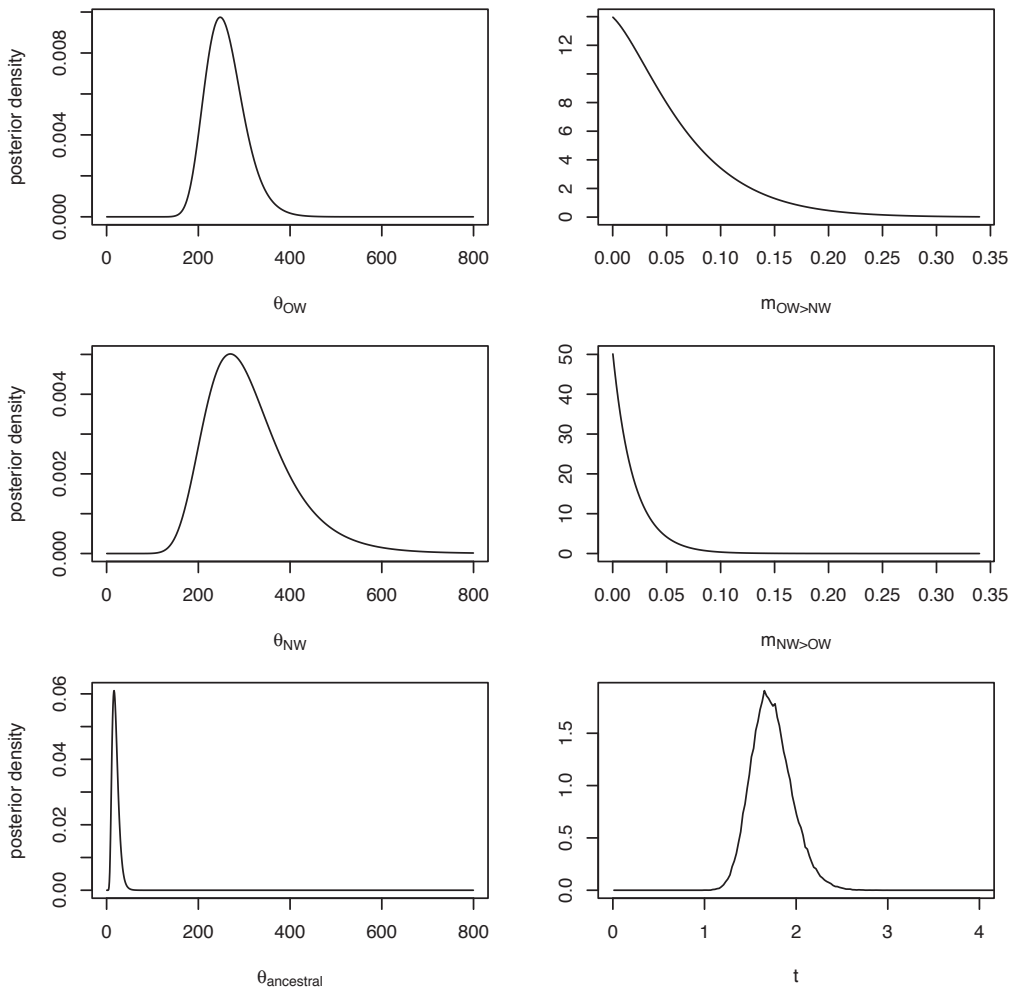


FIGURE 2.5: ***Density plots of the posterior probability distribution for each of the six estimated demographic parameters.***

Separate estimates for the effective population size (theta; θ) are given for Old World (OW), New World (NW) and the ancestral population, as well as the splitting time (t). Migration rates from OW into NW (abbreviated as '$m_{OW \rightarrow NW}$') and the opposite direction are also given. Note that all estimates are scaled to mutation rate and not in demographic units; migration rates are displayed in the direction of individual movements (i.e., not in the coalescent notation as presented in the raw output of the IMa2 program).

Discussion

CONCLUSIVE ASSERTION: THERE ARE TWO DISTINCT MALLARD CLADES

Since Avise *et al.* first studied genetic structure in mallards in 1990[79] it is hypothesised that no more than two mitochondrially distinct clades exist among them. Our phylogenetic analyses confirm this hypothesis for mallard samples collected and analysed from the numerically largest and geographically most complete set of mallard mtDNA sequences to date. In addition to previously analysed North-American and Asian samples[63], we added substantial numbers of European and Central and Eastern North-American mallard samples and thus completed a data set spanning the whole Northern Hemisphere. Through large-scale sampling of mallards from all over their distribution range we can now conclusively state that no more than two main clades exist in mallard, also not in the previously unsampled European populations.

TABLE 2.4: *Demographic parameters as estimated under an isolation with migration model.*

parameter	scaled to mutation rate			in demographic units		
	mode	HPD95L	HPD95H	mode	HPD95L	HPD95H
θ_{OW}	247.6	181.2	345.2	2,073,285	1,517,283	2,890,541
θ_{NW}	270.0	154.8	499.6	2,260,852	1,296,222	4,183,414
$\theta_{ancestral}$	16.4	6.8	35.6	137,326	56,940	298,098
t	1.7	1.3	2.2	55,366	42,002	74,993
$m_{OW \rightarrow NW}$	0.00017	0	0.1586	5.1×10^{-9}	0	4.7×10^{-6}
$m_{NW \rightarrow OW}$	0.00017	0	0.06069	5.1×10^{-9}	0	1.8×10^{-6}

For calculation of demographic units a generation time of one year was used. The mutation rate per locus per year was set to 2.9856×10^{-5}. See main text for further explanations. The demographic units are: population size (θ in effective individuals, splitting time (t) in years, migration rates (Nm) in effective individuals per year. Estimates of the peak of the posterior parameter distributions (mode) are presented along with low and high bounds of the 95% highest posterior densities (HPD95L/H).

GENE POOL DIFFERENTIATION AND DIVERSITY

Mallards are highly mobile and some studies suggest that natal philopatry (i.e., philopatry in the sense of ref [87]) is less pronounced - even in female mallards - than it is in other waterfowl species[12,69,88] (but see[89-91]). We found strong support for the hypothesis of 'flyway permeability'[68,69] in mallards. The analyses of F_{ST} between intra-continental flyways strongly promote the conclusion of female inter-flyway mixing. Neither within North-America nor in Asia did we detect differentiation between the subcontinental flyways. We note that for some comparisons the sample size was on the low side. However, if this was causal to not finding significant differentiation between flyways within a continent, we would also not have found significant differentiation between those flyways with lower sample sizes with flyways on different continents. Hence, we believe our finding of a lack of differentiation is valid. Yet, there is population structure in Europe based on F_{ST} analysis (see below).

Confirmative evidence for the pattern of genetic population structuring found above stems from our analysis of molecular variance (AMOVA) which applies F-statics to hierarchical population models. Almost no genetic variation in our world-wide data set is explained between flyways within continents. On this global scale, about half the variation lies between North-America and Eurasia, and the other half within the flyways. When analysing genetic variance per continent, in North-America and Asia the largest proportion of variance resides on the level of sampling localities but not within or between flyways. However, notably, in Europe there is an unexpected significant variance component between localities within flyways, pointing at more genetic structure than in the other continents.

In general, these findings confirm the absence of strong female philopatry in mallards. The population structure is not shaped by female mallards returning to the place where they were born, and hence lifetime dispersal seems high. In that sense, currently recognised waterfowl flyways do not depict well the movements and dispersal of mallards – a pattern similar to other duck species[88,92,93] – although it cannot be generalized for all ducks[94,95]. However, our findings show that indeed the European mallard population seems to be structured – in contrast to their conspecifics in other parts of the world – and may have some weak philopatric tendencies. Alternatively, geographic, micro-climatic and urbanisation structuring in Europe may be different in Europe when compared to Asia and North-America, shaping mallard population structures in different ways. A further possibility to explain European mallard population structure may be extensive release of farmed mallards for hunting purposes in many countries in Europe. These originally wild mallards are bred in captivity for many generations, and even though looking quite similar to their wild relatives they can differ subtly in morphology[71]. Such differences have been proven to leak into wild populations by interbreeding with escaped farmed mallards[96]. Obviously, this introgression of genetic material of farmed animals has consequences for the genetic identity of wild populations, especially when translocation and release at distant localities takes place[97]. Genetic introgression has been demonstrated for mallards in Italy already[98]. Such manipulation of local population structure may well explain our finding of increased genetic structure in Europe.

DEMOGRAPHIC HISTORY

Mallards from the Old World are different from those in the New World, which is attributable to the differences in the distribution of clade A (Eurasia) and clade B (North-America) haplotypes. In Europe and the largest part of Asia we only found clade A haplotypes with the exception of a handful clade B haplotypes in the far east of Asia (Figures 2.2, 2.3). Nevertheless, in a phylogenetic assessment in the original study in which these sequences were determined, they were shown to be characteristic for introgressed eastern spot-billed ducks and likely do not indicate original North-American mallard clade B individuals[63,82] (removing these three curious sequences from the demographic modelling analysis did not alter the results, though; data not shown). On the other hand, in North-America, where clade B dominates, also clade A haplotypes are relatively common, which is reflected in higher nucleotide and haplotype diversities. This suggests two factors shaping the current genetic structure in mallards. From the deep split into clade A and B haplotypes we infer that Old World and New World have become separate with no or little gene

flow since then. Further, if genetic exchange occurs – as indicated by clade A haplotypes occurring both in Old and New World – it seems directional with larger rates from Old World into New World. However, this directionality is not statistically robust because in the IM analysis migration rates were only estimated poorly. The split between mallards from Old World and New World was estimated to have occurred between 43,000 and 74,000 years ago. This dating coincides with the drop in sea level during the last glacial period about 110,000 – 10,000 years ago. During this time the Beringia land bridge fell dry and extensive exchange of floras and faunas became possible. When the glaciers melted again, this Beringia land bridge connection got cut off and Eurasian and North-American faunas and floras became separated more strictly. But also during the glacial maximum the continental ice sheet developed in North-America separated northern from southern regions of this continent[99]. The estimated population size at that time ($\theta_{ancestral}$) was about 30 times smaller than it is today (the sum of θ_{OW} and θ_{NW}). Such high growth rates in waterfowl populations since the last glaciation has already been demonstrated in several other studies, often proven to be exponential[63,73,88,100] and connected to population size bottlenecks. This could explain the excess of rare haplotypes (Figure 2.1) as demonstrated especially by negative Tajima's *D* and Fu's *FS* in Eurasia, but also in North-America where, however, *D* was not significantly different from zero. The present-day situation is a deep split into an Old World and New World population, with migration rate estimates statistically not significantly different from zero, indicating that the rise of sea levels after the last glacial maximum results in a potent barrier to female gene flow between OW and NW. Numerically, the peak of the density distribution of migration rates suggests 0.04 – 0.05 effective migrants per year in each direction. However, the posterior density plots of the migration rates (Figure 2.5) are not bell-shaped as they should be. This indicates that IMa2 did not find enough information in the data to reliably estimate migration rates. If prior ranges are too wide, software like IMa2 may estimate parameters to be zero, but in our analysis the range of the prior distribution (0 – 0.34) was not much wider than the range of the estimated posterior distribution (HPD95) for migration of Old World mallards to the New World (0 – 0.16) (Table 2.4). Re-running the IMa2 analysis with a narrower prior (0 – 0.2) did not resolve the issue (data not shown). We believe this is due to a combination of very large extant population sizes as a result of strong population growth and divergence. Based on HPD95 migration rates could be as low as zero, or as high as 23 (NW OW) or 80 (OW NW) effective individuals. The failure of IMa2 to quantify migration rates may be related to this issue. Irrespective of this, the sharing of identical haplotypes indicates the possibility that migrants do travel between the land masses and establish genetic traces in the receiving population. The nucleotide diversity of the Aleutian samples, being intermediate between that of Eurasia and North-America, is a good indicator for this. Even though migration rates as such were poorly estimated with our data, the shapes of the migration rate curves and the implemented probability test in IMa2 point at directionality of migration from Old World into New World.

GENETIC EXCHANGE BETWEEN OLD WORLD AND NEW WORLD

Based on the sharing of haplotypes between Old and New World mallards and the results of the demographic modelling, we propose that some genetic exchange of mtDNA between the two

land masses possibly occurs. IMa2's estimated migration rates, being not statistically different from zero, may be biased downwards due to a lack of statistical power for the migration rate parameters, or model violation of migration/drift equilibrium, but indicate directionality. Genetic differentiation between Old and New World was high ($F_{ST} = 0.51$) indicating limited gene flow at most. We propose that the route of this migration is mainly, if not completely, via the Aleutian Islands. On these islands individuals with clade A and B haplotypes intermix and form a haplotype cline[63], and genetic diversity (π) is intermediate between Eurasia and North-America. F_{ST} values between the Aleutians and the flyways indicate a closer relationship between the Aleutians and Old World populations than with the New World. This may be explained by the Westerly winds (more regularly blowing from the west into the east), governing the Ferrel cell of the global climatic circulation system. The Aleutian Islands are all well within the Ferrel cell that reaches up to higher than 60° north[101,102], and characteristic winds are especially strong during the time in which mallards migrate[103]. A note of caution is warranted here, because smaller scale terrestrial weather conditions can be more important than the large atmospheric systems. Even if weak when compared to the strength of flight abilities this effect can have a profound impact on the regular "drifting off" of migrating birds in general[103] and waterfowl specifically[104].

The alternative route, from Europe via the Atlantic Ocean, along Iceland and Greenland, could be deemed less likely on basis of this "drifting by wind" proposition. Mallards travelling to North-America by the Atlantic route would face headwinds more regularly[103]. However, the mallards sampled from Greenland all bear clade A haplotypes implying a Eurasian origin. Additionally, genetic differentiation between mallards from the Atlantic part of Canada and European mallards is only half of what we measured between central or western Canada and Europe. Once arrived in Greenland it would not be hard to imagine crossing the last few hundred kilometres to Canada. Unfortunately, we were not able to analyse more samples from Greenland, especially from different localities. Interpretations based on the current Greenland data can only be tentative. Just two different haplotypes were found there and the sampling location was a pond near the city of Nuuk, likely inhabited by closely related mallard families. Little is known about the origin and movements of Greenland mallards[12,105], and hardly anything (except from data in this study) about their genetics (for a preliminary report, see[106]). More studies into this population are certainly needed to draw firm conclusions. Sampling of mallards from Iceland and the Faroe Islands would make it possible to perform detailed analyses of a potential Atlantic route of genetic exchange of mallard populations.

Note, however, that all results presented in this paper are based on the analysis of mtDNA which is a maternally inherited marker. Patterns discerned from our data are thus only valid for the female part of the population. Males are sometimes suggested to also possess a homing instinct[107] but in contrast to females this is generally believed to be much less pronounced[108]. It will thus be important to study nuclear markers in mallard[98,109-112] on large geographical scales which depict the genetic structure of both sexes (Kraus *et al.*, manuscript in preparation). Some sex-bias in migration is not unusual[113] but can be extreme if males are much more dispersive than females, as demonstrated in a famous example of the white shark *Carcharodon carcharias*[114].

POST-GLACIAL COLONIZATION

Clade A is characteristic for Eurasia, and clade B for North-America, although many individuals in North-America belong to clade A. Kulikova *et al.*[63] proposed two phylogeographic hypotheses to explain such a haplotype distribution, termed "Asian invasion" and "incomplete lineage sorting" (initially suggested by Avise *et al.*[79]). In an Asian invasion scenario the occurrence of clade B haplotypes in mallards is explained by Asian mallards moving into a previously mallard-free North-America, followed by acquiring B-haplotypes by introgressive hybridisation with closely related indigenous duck species (such as the black duck *A. rubripes*[115]) resulting in frequently observed mtDNA paraphyly[5,82,116]. The incomplete lineage sorting hypothesis rests on the proposed occurrence of a polymorphic ancestral gene pool (at least with respect to mtDNA clades) which is facilitated further by large populations[117,118]. These two hypotheses offer different predictions about the expected distribution of clade A and clade B haplotypes in North-America: As a result of an Asian invasion one would expect to find a gradual decline of clade A haplotypes from western to eastern North-America if we assume that all immigration from Eurasia took place via the Aleutian Islands and continues so. We cannot confirm this scenario; clade A haplotypes today occur in the whole of North-America. This, however, can also be the result of large dispersal abilities of mallards in North-America, which can diminish genetic clines relative quickly. Hence, an Asian invasion cannot be clearly rejected, either. Kulikova *et al.*[63] imagine this further to be due the impact of mallard farming, resulting in anthropogenic translocation also of clade A mallards. This explanation resembles one of the possibilities we propose to explain genetic structure in Europe.

As an alteration of the Asian invasion scenario we offer additional thoughts based on the extent of the most recent continental glaciation in North-America. The opening of the Beringia land bridge on the one hand enabled exchange of fauna and flora between East Asia and Alaska because these regions were to a large extent ice-free. Coincident with the emergence of this connection, though, the considerable North-America ice sheet built up south of Alaska, what is today southern Canada. An Asian invasion thus took place probably only in the very north of North-America. The clade B haplotype may have become fixed south of that ice sheet due to the population bottlenecks. After the retreat of the glaciers clade B mallards from the south got into secondary contact with northern clade A mallards. It would thus be interesting to sample mallards from central and southern North-America with this hypothesis in mind. To account for and further study mtDNA paraphyly in duck species such a study would have to include samples especially from Black Ducks[79,115], and if possible from other *Anas* species harbouring clade B haplotypes, too[5,116].

Avise *et al.*'s incomplete lineage sorting hypothesis[79] would naturally account for a more even distribution of clades and is consistent with the idea that a dichromatic (and mitochondrially polymorphic) mallard population gave rise to its monochromatic sister species by peripatric isolation[119]. But as Kulikova *et al.*[63] point out it seems unlikely that only the clade B haplotype gets fixed in all sister species but not the "original" mallard. We thus concur with Kulikova *et al.*[63] in concluding that a complicated mix of historical, recent and anthropogenic factors shaped the current world-wide mallard population structure. The results we can add from analysing >100 additional European mallards and the outcome of our demographic modelling study substantiate this claim.

Conclusions

Many aspects of the biology of ducks are known when it concerns ecological and management parameters (see ref [70] and references therein for a recent overview), and also their phylogenetic placement has frequently been assessed, e.g. refs [5,6]. Surprisingly, only few studies have contributed to discern population genetic patterns for ducks, especially mallards. The information collected in this study is essentially representing the first complete assessment of the world-wide mallard population genetic structure in this duck. With more than 300 samples from the whole distribution range of the mallard we do not find intermediate haplotypes between clade A and B, or additional discernable clades. Our data and analyses corroborate the conclusion of Kulikova *et al.*[63] that a complicated mix of historical, recent and anthropogenic factors shaped the current world-wide mallard population structure. Further, we offer an additional hypothesis on how the current haplotype distribution emerged: a partial Asian invasion that took place only in northern North-America, which was ice-free at the last glacial maximum, followed by secondary contact with the southerly mallard population in North-America in which clade B haplotypes could have been fixed. A study to address this hypothesis would need to additionally analyse mallard samples from several well-spaced localities throughput central and southern North-America, as well to pay attention to *Anas* species closely related to mallards, often bearing clade B haplotypes themselves[5,116].

Mallards form an enormously large population. In the northern hemisphere, the population is structured deeply into two major clades, but within these landmasses the populations are very homogenous (and perhaps panmictic) in which the concept of flyway do not contribute to a further understanding of mallard population genetic structure. Homogeneity over thousands of kilometres facilitates the spread of diseases such as Avian Influenza. Between the two landmasses there may be a little gene flow, apparently in west - east direction across the Bering Strait. The genetic distinction between the OW and the NW is not eroded away by this gene flow, and apparently arose during the last Glacial Maximum.

Our analysis does not undermine the notion that waterfowl can very well be managed on the basis of a flyway concept: the North American success story of waterfowl population increases over the last century underpins that success. In that sense, flyways ought to be viewed as 'problemsheds' in the sense of the Ecosystem Approach of the Convention on Biological Diversity and not so much as biological realities.

Methods

SAMPLE COLLECTION

Blood from 195 mallards was collected on FTA cards[67], in most cases by hunters. Exceptions are the localities from Greenland and Norway. There, mallards were trapped, blood drops on FTA cards were sampled from the wing or foot vein, and mallards were released again (animal handling approved by the animal ethical committee of Wageningen University – DEC, the Greenland Home Rule and the Norwegian Food Safety Authority – Forsöksdyrutvalget). Details on sampling localities and samples can be found in Table 2.1 and the supplementary file "samples.xls". A cautionary note is needed for the Greenland samples. The trapping of these animals most likely resulted in the catching of a few ducks with several of their chicks which often entered the trap

as a group. This may have affected some measures of population genetic parameter (see results and discussion).

For phylogenetic analyses samples were assigned to continents to visualise their geographic region of origin. Greenland and the Aleutian Islands were not assigned to a specific continent, and Alaska and Canada were treated as separate regions within North-America. For population genetic analyses, sampling localities were pooled into biological populations based on hypothesised mallard flyways in Europe[12], Asia[16], and North-America[17] (Table 2.1), allowing us to test our data against these human-made classifications. Samples from Greenland and the Aleutian Islands were classified as separate populations for their intermediate status.

DNA ISOLATION AND SEQUENCING

DNA was extracted from FTA cards using the Gentra Systems 'Puregene DNA purification Kit' (Qiagen, Valencia, California). The manufacturer's instructions were followed with slight modifications for handling the FTA cards: up to a quarter of the encircled area of the FTA cards (depending on how much blood was preserved) was cut into small pieces of approximately 2 mm^2 and digested with 60 µg Proteinase K (Sigma-Aldrich, St. Louis, Missouri) in 600 µl Cell Lysis Solution (Gentra Systems) at 55°C over night. Subsequently, proteins were precipitated with 200 µl Protein Precipitation Solution (Gentra Systems) and spun down together with the FTA card material. DNA in the supernatant was precipitated with isopropanol and washed with 70% ethanol. Quantity and purity of the DNA were measured using a Nanodrop ND1000.

PCR amplification targeted the 5'end of the mtDNA control region which is homologous to positions 79-773 in the chicken (*Gallus gallus*) mitochondrial genome[120]. In some duck species the presence of 'numts' (nuclear copies of mtDNA) was proposed in this region[121] but previous examination of mallard sequences did not reveal evidence for this[63]. Reactions were performed in 12 µl containing 30 ng genomic DNA as template, 3 µl STE buffer, 5.5 µl Abgene Mastermix (ThermoScientific) and 0.25 µl of each primer (10mM): L78[121] (forward) and H774[122] (reverse). PCR amplification was done in a BioMetra Thermocycler (Biometra, Göttingen, Germany) under the following cycling conditions: 7 minutes initial denaturation at 94 ºC, followed by 45 cycles of 20 seconds at 94 ºC, 20 seconds at 49 ºC and 1 minute at 72 ºC, completed by 7 minutes final elongation at 72 ºC. Quality and quantity of the PCR product was determined by gel electrophoresis and the product was purified by vacuum filtration on a Millipore Multiscreen PCR plate. Forward and reverse DNA strands were cycle-sequenced using the ABI Big Dye Terminator Cycle Sequencing Kit 3.1 under the following cycling conditions: 1 minute initial denaturation at 96 ºC followed by 25 cycles of 10 seconds at 96 ºC, 5 seconds at 49 ºC, and final elongation of 4 minutes at 60 ºC. The sequencing reaction products were precipitated by sodium acetate and ethanol to purify the product, followed by capillary sequencing on an ABI 3730 DNA Analyzer. The forward sequences were verified with the sequence of the reverse strand in MEGA4[123]; some manual corrections where needed. If the forward sequence was absent, or only partially resolved, the reverse strand was used and aligned with the other sequences for verification. Additionally, 151 published sequences from studies of Kulikova *et al.*[63,82] were downloaded from GenBank[29] (accession numbers: AY506868-AY506870; AY506873-AY506901; AY506904-AY506908; AY506910-

AY506917; AY506919-AY506944; AY506974-AY506984; AY928831-AY928899). Altogether, these sequences were aligned in MEGA4[123] using the ClustalW algorithm[124] under default settings. Names of haplotypes defined in the studies by Kulikova *et al.*[63,82] were preserved.

PHYLOGENETIC ANALYSES

A phylogenetic tree of the haplotype sequences was constructed in MEGA4[123], using the Neighbour-Joining algorithm[85] with 500 bootstrap replicates. Evolutionary distances were computed under the Tajima Nei model[125]. All positions containing alignment gaps and missing data were eliminated from the dataset for tree construction (complete deletion option). Further, a phylogenetic network was constructed in TCS[126] (version 1.21) by statistical parsimony, here, treating alignment gaps as fifth state.

POPULATION GENETIC ANALYSES

The basic population genetic parameters nucleotide diversity (π) and haplotype diversity (dH) for each flyway was calculated with DnaSP[127]. Tajima's *D*[83] and Fu's *FS*[84] were calculated in Arlequin 3.5.1.2[128], and Fu's *FS* values were evaluated for statistical significance by 16,000 simulated samples in order to guarantee a less than 1% difference with the exact probability in 99% of the cases[129]. Population differentiation was assessed by Wright's *F*-statistics[86], and partitioning of genetic variance among and within groups was investigated by analyses of molecular variance (AMOVA[130-132]). Calculations were also performed in Arlequin 3.5.1.2[128] from pairwise nucleotide differences, and statistical significance tested by 16,000 permutations.

DEMOGRAPHIC MODELLING

To make inferences about the extent of migration between Old World (OW) and New World (NW) we modelled the demographic history of mallards by coalescent simulations under an "Isolation with Migration" (IM) model[80,81], as implemented in the program IMa2 (Linux version 10.13.10). OW and NW samples were treated as belonging to distinct populations based on their sampling locality. Greenland and Aleutian samples were excluded because of their possible intermediate status. Upper bounds for parameter priors were estimated during consecutive preliminary runs of the program, based on initial estimates of theta as advised in the manual of IMa2. The final values used for population size, migration rate and splitting time were: -q 800, -m 0.34, -t 23.448. We ran 60 Markov chains in parallel under a geometric heating scheme (option -hfg), with the hottest chain being $\beta = 0.5$ and the coldest chain $\beta = 0.975$. Estimated parameters in IMa2 are scaled to the mutation rate. To convert them into demographic units we used a mutation rate of 4.8×10^{-8} (confidence interval $3.1\text{-}6.9 \times 10^{-8}$) substitutions per site per year initially published for a wood duck[94]. This rate also produced sensible results in a study of two other ducks of the genus *Anas*[133,134] and needs to be multiplied by the number of nucleotides in the sequence alignment (here, 622) to obtain the substitutions per locus per year to be used for IM analysis. From the two sequence mutation models available in IMa2 we chose HKY[135] which is the applicable model for mtDNA control region sequences[136]. Several run time settings with different heating schemes and durations were explored, all yielding essentially the same outcome. The final

simulation, for which we report the results here, was run for a burn-in period of 360,000 steps (they reached convergence already after a few 10,000 steps), and afterwards 26,000 genealogies were sampled every 100 steps from a total 2,600,000 steps.

Acknowledgements

The following people contributed samples (in alphabetical order of the respective sample IDs): Ernst Niedermayer, Hans Jörg Damm (Stiftung Fürst Liechtenstein, Austria), Severin Wejbora (Lehr- und Forschungsrevier des Landesjagdverbandes Bayern, Germany), Urmas Võro (Estonia), Andy Richardson (Safari in Scotland, Scotland), Herman Postma (The Netherlands), Jan Bokdam (Nature Conservation and Plant Ecology, Wageningen University, The Netherlands), Alf Tore Mjøs (Museum Stavanger, Norway), David Lamble, Garry Grigg, Steven Evans, Thomas Kondratowicz, Henry J. Bruhlman, Darren Hasson, Aaron Everingham, Andrew Iwaniuk (hunters from Canada). Charles Bull (Northmore, Britain), Alyn Walsh and Dominic Berridge (Wexford Wildfowl Reserve, Ireland) helped in sampling set-up and coordination in Great Britain. Holly Middleton helped in coordinating sampling efforts in Canada. We would like to thank the staff of the Greenland Institute of Natural Resources for their support during our Greenland expedition, especially Aili Lage Labansen for organising our stay, and Carsten Egevang for hosting us in his laboratory. Hans Geisler supported our trapping activities at the sampling site in Nuuk. The Animal Breeding and Genomics Group (Wageningen, The Netherlands) generously hosted us in their molecular laboratory. Sylvia Kinders, Tineke Veenendaal and Bert Dibbits are thanked for helping with lab work. This work was financially supported by the KNJV (Royal Netherlands Hunters Association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik trusts (both in The Netherlands).

Additional files

upon request from robert.kraus@senckenberg.de

ADDITIONAL FILE 1: **haplotypelist.xls**
MS Excel sheet giving information which samples had which haplotype.

ADDITIONAL FILE 2: **samples.xls**
MS Excel sheet with all details for each mallard samples and sequenced for this study. Details given include internal sampling ID, sampling date, sampling country, names of sample collectors, name of sampling location, decimal latitude and longitude coordinates and determined sex of the sampled mallard.

CHAPTER 3

Genome wide SNP discovery, analysis and evaluation in mallard (*Anas platyrhynchos*)

Robert HS Kraus, Hindrik HD Kerstens, Pim van Hooft, Richard PMA Crooijmans, Jan J van der Poel, Johan Elmberg, Alain Vignal, Yinhua Huang, Ning Li, Herbert HT Prins, Martien AM Groenen

Article published: BMC Genomics. 2011. 12:150

Abstract

BACKGROUND

Next generation sequencing technologies allow to obtain at low cost the genomic sequence information that currently lacks for most economically and ecologically important organisms. For the mallard duck genomic data is limited. The mallard is, besides a species of large agricultural and societal importance, also the focal species when it comes to long distance dispersal of Avian Influenza. For large scale identification of SNPs we performed Illumina sequencing of wild mallard DNA and compared our data with ongoing genome and EST sequencing of domesticated conspecifics. This is the first study of its kind for waterfowl.

RESULTS

More than one billion base pairs of sequence information were generated resulting in a 16X coverage of a reduced representation library of the mallard genome. Sequence reads were aligned to a draft domesticated duck reference genome and allowed for the detection of over 122,000 SNPs within our mallard sequence dataset. In addition, almost 62,000 nucleotide positions on the domesticated duck reference showed a different nucleotide compared to wild mallard. Approximately 20,000 SNPs identified within our data were shared with SNPs identified in the sequenced domestic duck or in EST sequencing projects. The shared SNPs were considered to be highly reliable and were used to benchmark non-shared SNPs for quality. Genotyping of a representative sample of 364 SNPs resulted in a SNP conversion rate of 99.7%. The correlation of the minor allele count and observed minor allele frequency in the SNP discovery pool was 0.72.

CONCLUSION

We identified almost 150,000 SNPs in wild mallards that will likely yield good results in genotyping. Of these, ~101,000 SNPs were detected within our wild mallard sequences and ~49,000 were detected between wild and domesticated duck data. In the ~101,000 SNPs we found a subset of ~20,000 SNPs shared between wild mallards and the sequenced domesticated duck suggesting a low genetic divergence. Comparison of quality metrics between the total SNP set (122,000 + 62,000 = 184,000 SNPs) and the validated subset shows similar characteristics for both sets. This indicates that we have detected a large amount (~150,000) of accurately inferred mallard SNPs, which will benefit bird evolutionary studies, ecological studies (e.g. disentangling migratory connectivity) and industrial breeding programs.

Background

The mallard (*Anas platyrhynchos*) is the world's most abundant and well-studied waterfowl species. Besides being an important game and agricultural species, it is also a flagship species in wetland conservation and restoration. Waterfowl (Anseriformes: Anatidae) and especially ducks are focal organisms in long distance dispersal of Avian Influenza in the wild[66,137-139], and the mallard has been identified as the most likely species to transport this virus[10,140].

As a general pattern, mallards breeding in temperate areas migrate from northern breeding grounds to more southerly wintering areas avoiding freezing conditions at breeding sites[141]. How-

ever, there are also non-migratory populations in Europe and elsewhere. Although some geographical patterns can be discerned from ringing recoveries on national levels, there is in Europe no clear delineation of flyways, and only little knowledge about the overall population structure from a genetic perspective[12]. This is exactly the situation for which Wink[14] proposed the use of SNPs to study bird migration in a population genetic framework. Since the number of SNPs necessary to detect low levels of differentiation is expected to be high (>80) for highly mobile organisms[142,143], we aimed at a high throughput discovery of SNPs in the mallard. Large scale discovery of SNPs in the genome of the wild mallard might also provide a useful set of markers in the descendant, closely related domesticated duck (*Anas platyrhynchos domestica*). Being the third most consumed species on the poultry market globally[144], the domestic duck provides a valuable subject for detailed genomic studies. Nevertheless, genomic information about the domestic duck is limited to a few studies providing only low resolution linkage and physical maps[109,145]. Therefore our study also set out to facilitate duck breeding objectives by providing sufficient markers for improving the duck linkage map and allowing QTL mapping using SNPs.

A general limitation in developing a SNP-set in non-model organisms has been the unavailability of extensive genomic sequence information from multiple individuals that represent a sufficient portion of the genetic variability of the population or species under study. However, the Illumina sequencing technology[146-148] coupled with the approach of generating a reduced representation library (RRL)[149] has proven an efficient approach in solving this problem in the turkey (*Meleagris gallopavo*)[150] and great tit (*Parus major*)[151]. Also in rainbow trout[152], pig[153,154] and cattle[155] next generation sequencing of RRLs has been effective in the identification of considerable numbers of SNPs.

Here, we describe the discovery of more than 180,000 novel SNPs in the genome of the mallard, which currently lacks a published sequenced genome. Lacking this reference genome we initially aimed for paired-end sequencing on an Illumina Genome Analyzer of an RRL of fragments in the size range of 110-130 base pairs (bp) and with a read length of 76 bp. This would create an overlap between the forward and reverse DNA sequence reads of continuous sequences, permitting the reads to be merged. This in turn helps in providing sufficient flanking sequence (i.e., DNA sequence on either side) of a SNP which is a requirement for genotyping and is hard to retrieve in the absence of a reference genome. However, at the time when our study had started, genome sequencing of the domestic duck genome and *de novo* assembly was in progress and almost completed by the Beijing Genome Institute (BGI). This allowed for SNP discovery by next generation sequencing of an RRL of pooled wild mallard samples and mapping locations of almost 13 million of the resulting reads to a draft mallard reference sequence. Identified SNPs were compared with those observed within the reference genome sequence of domestic duck (Huang *et al.*, in prep.) and EST sequencing (expressed sequence tags; Alain Vignal, unpublished data) resulting in more than 20,000 shared high quality SNPs. A set of putative SNPs can contain large numbers of incorrectly inferred SNPs (i.e., false positives) and thus we also aimed to estimate the quality of our set. Quality, here, is a measure of the reliability of the SNP set. This includes not only the percentage of false SNP inferences but also evaluation of the way in which these SNPs will be usable for many purposes; i.e., if they cover a large spectrum of minor allele frequencies, or if these were reliably inferred by our analyses (correlation between true allele frequencies and estimated allele counts, see below).

Results

COMPLEXITY REDUCTION

We targeted for a sequencing depth of about 40 times at limited sequencing cost by sequencing a fraction, representing 5% of the mallard genome (reduced representation library (RRL) approach). Restriction enzymes were screened for suitability for RRL construction, with the goal of a 20-fold complexity reduction of the mallard genome within the targeted size range of 110-130 bp. Restriction enzyme analyses showed that these requirements are met by combining two RRLs, one created by enzymatic digestion with *Alu*I and one by digestion with *Hha*I, representing 4% and 1% of the mallard genome, respectively.

An *in silico* digest of the chicken genome, which is very similar[156,157], predicts similar genome fractions of the RRLs of 4.1% for *Alu*I, but only 0.2% for *Hha*I (data not shown). We prepared two pooled DNA samples of nine wild mallard individuals from three locations across Europe. To prepare the RRLs, we digested these samples with *Alu*I or *Hha*I and isolated fragments in the 110-130 bp size range from a preparative polyacrylamide gel. The genomic libraries were combined in the sequencing sample preparation procedure. Due to a lack of a reference genome we aimed for paired-end sequencing on an Illumina GAII of the combined RRLs and a sequence read length of 76 bases. This created an overlap between the forward and reverse reads of a pair which allows merging of the reads. Merging the reads helps in providing sufficient flanking sequence of a SNP. This sequence is necessary for genotyping and is hard to retrieve in the absence of a reference genome. Merged paired reads, possibly supplemented with single reads, are subsequently clustered for SNP discovery.

TABLE 3.1: *Summary of DNA sequence filtering results.*

	raw (76 bp)	l62 N . q12 o152[1]	%	paired-end	%	single	%
reads	34818352	16611852	47.7	10793170	65.0	5818682	35.0
bases	2547361732	1029934824	40.4	669176540	65.0	360758284	35.0

Paired and single sequence reads remaining after filtering raw reads.

[1]Raw sequences were filtered for length 62. Only reads without base-call errors (N or .) were considered. Singly represented reads are required to have a per base-call quality of 12. Sequences more than four times overrepresented, based on the raw RRL coverage (38X, see methods) were discarded.

ILLUMINA SEQUENCING AND SNP DETECTION

We generated 34.8 million 76 bp reads using three sequencing lanes on an Illumina GAII of which two lanes were run in paired-end mode. The raw data files from the sequencing instrument are deposited in the NCBI short read archive under accession number SRA024498. It was shown that a phred quality score[158] threshold of 12 ensures sufficient quality reads for SNP detection purposes[153,159]. Because the average base call quality score over all sequence reads dropped below 12 after read position 62, reads were trimmed to 62 bp. After trimming, we performed additional quality score based filtering (see methods) and finally we retained 16.6 million reads (47% of the raw data) of 62 bp length corresponding to a total of 1.03 billion bp of sequence information (Table 3.1). Of these reads 35% were single and 65% were paired reads. By creating RRLs 5% (69

Mb) of the mallard genome was represented (estimated size 1.38 billion bp, based on several entries in the Eukaryotic genome size databases[160]). From this we calculated that the raw sequencing data cover the sequence target 38 times (38X) whereas the quality filtered data provide a 16X target coverage. Using MAQ[161] 12,823,563 of the reads could be mapped onto the mallard reference genome (Huang *et al.*, in prep.). A total of 632,163 putative SNPs were identified by MAQ[161] of which 122,413 candidate SNPs passed our applied SNP identification quality thresholds (see methods). This set of SNPs is further referred to as duck-RRL (d-RRL) and available in the dbSNP database under accession numbers ss263068950 - ss263191362.

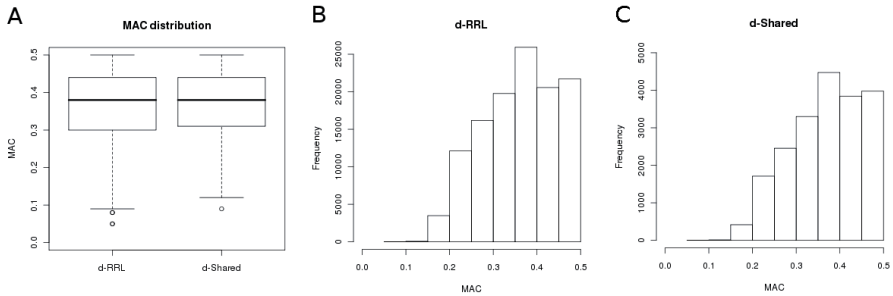


FIGURE 3.1: ***Minor allele frequency distributions.***
In boxplot A MAC distributions of d-RRL (SNPs identified in this study) and d-Shared (SNPs that d-RRL shares with d-EST or d-WGS (also see Venn diagram Figure 3.2D)) are compared. Histograms (B and C) show MAC distributions of d-RRL and d-Shared at a bin width of 0.05.

SNP USABILITY

More than 98.8% of the SNPs were flanked by at least 40 bp on either side and met the requirements for probe design constraints for all genotyping platforms whereas all SNPs met the flanking sequence requirements for an iSelect (Illumina) genotyping assay. For the 2,565 SNPs that showed more than two alleles, we only considered the most frequently observed minor allele because tri- or tetra-allelic SNPs are very rare[64] and it is likely that most other minor alleles represented sequencing errors instead of true sequence variants. Analysis of the estimated allele counts of the SNPs in our dataset (Figures 3.1A and 3.1B) showed that we obtained a majority of SNPs with a high minor allele count (MAC, used here as a predictor of the minor allele frequency (MAF) of the real population data).

SNP QUALITY ASSESSMENT

Sequencing errors are more abundant in the tails of next generation sequencing reads and are thought to cause an excess of false SNP predictions. An increase in the number of SNPs towards the end of the reads is expected if sequencing errors are the cause of a substantial number of predicted SNPs in the dataset. To validate our sequence filtering and SNP detection constraints we plotted the distribution of the SNPs over the 62 positions in the sequence reads (Figure 3.2A). Positions one, two and 62 all show an underrepresentation of SNPs whereas positions three, four and five show an

overrepresentation. SNPs are equally distributed over read positions 6 to 25 and at 26 the number of SNPs per nucleotide position drops but after this remains more or less stable until position 62.

Because of the length of the RRL fragments (~110-130bp), there is an overlap between paired forward and reverse reads (62 nucleotides each) from position 48 onwards. This overlap results in a higher sequence depth and a tiny increase in the number of SNPs being detected at these nucleotide positions (Figure 3.2A).

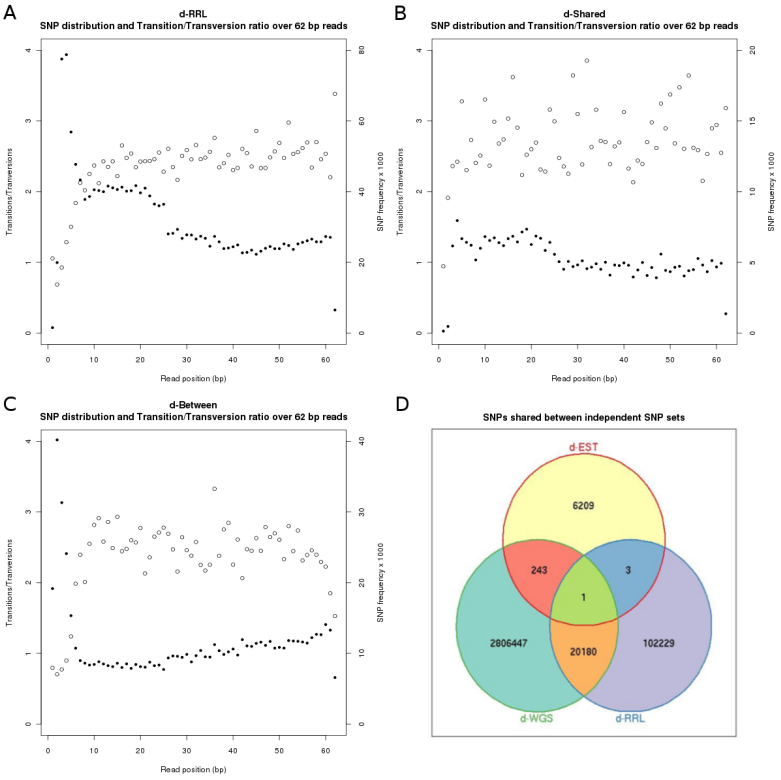


FIGURE 3.2: *SNP distributions within datasets and between datasets.*
Diagrams A-C show the distribution of SNP predictions over the nucleotide position in the sequence reads for d-RRL, d-Shared and d-Between. Each filled dot represents the cumulative number of occurrences that the read position was involved in a SNP inference. Open dots represent the average TS:TV ratio of SNPs indentified in that read position. Diagram D shows how many SNPs are shared between independent SNP sets d-EST (SNPs identified by EST sequencing of domesticated duck (Vignal, unpublished data)), d-WGS (SNPs identified in the whole genome assembly of domesticated duck (Huang et al., in prep.)) and d-RRL (SNPs identified in RRL sequencing of wild mallard (this study)).

We estimated the possible errors in SNP calling due to sequencing errors by looking at transition (TS) - C/T pyrimidine to pyrimidine or A/G purine to purine changes - versus transversion (TV) ratios, which are all the four other possible pairs of changes. Random mutations or sequence differences due to errors should give a TS:TV ratio of 1:2. In reality, a bias due to a higher rate of C→T mutations due to the deamination of methylcytosines in CpG dinucleotides induce a much higher TS rate[162-165]. For instance in chicken, the TS:TV ratio is 2.2:1, based on the analysis of more than 3 million SNPs in the dbSNP database[166]. Our results show that the number of A/G substitutions almost equalled the number of C/T substitutions in the transitions class. Also the substitutions within the transversions class occurred in comparable frequencies (Table 3.2). The TS:TV ratio for d-RRL was 2.3:1 which is very similar to the 2.2:1 ratio found in chicken.

Sequencing errors were also evaluated per read position by plotting the TS:TV ratio observed over the 62 positions in the sequence reads (Figure 3.2). We observed steady expected TS:TV ratios for positions 7-61 whereas TS:TV ratios for positions 1-6 were lower and the TS:TV ratios for position 62 was higher than expected.

TABLE 3.2: *Transition/transversion ratios in SNP subsets.*

	Transitions				Transversions		Total	TS:TV[1]
subset	R	Y	M	W	S	K		
d-RRL	42313	42602	9658	9051	9114	9675	122,413	2.3
d-Shared	7300	7442	1396	1227	1334	1484	20,184	2.7
d-Between	20156	21333	5464	5165	4804	4830	61,752	2.0

[1]=The transitions total divided by the transversions total for a data subset.

The two transitions and four transversions are abbreviated by their nucleotide ambiguity codes R, Y and M, W, S, K.

SNP BENCHMARKING

The *de novo* assembly of the domestic duck genome by the Beijing Genome Institute (BGI), covering both chromosomes of a single individual, resulted in the identification of 2,826,871 putative SNPs (further referred to as d-WGS; Huang *et al.*, in prep.). Domestic duck EST sequencing identified a total of 6,456 SNPs (further referred to as d-EST) in protein coding regions of the genome (Alain Vignal, unpublished data).

To benchmark d-RRL we compared it with these two external and independent datasets and identified SNPs that are shared with either d-WGS or d-EST. We observed 20,180 SNPs (16.5%) in common between d-RRL and SNPs in the d-WGS dataset. Furthermore d-RRL had four SNPs in common with d-EST whereas d-WGS shared 244 SNPs with d-EST (Figure 3.2D). Only a single SNP was shared between all three datasets. The subset of SNPs (n=20,184) that d-RRL shared with either of the two other SNP resources is further referred to as d-Shared. We analysed d-Shared by calculating the MACs and the TS:TV ratios (Figure 3.1C and Table 3.2). Furthermore, we plotted the TS:TV ratio per read position and the distribution of the SNPs over the 62 nucleotides of the sequence reads in the same way as was done for d-RRL. In d-Shared we observed a similar distribution of MACs compared to d-RRL (Figure 3.1C). The distribution of the SNPs in d-Shared detected on read positions 7-62 is similar to that observed for d-RRL; however, d-Shared shows

a higher variation in the amount of SNPs between the read positions (Figure 3.2B). Also, TS:TV ratios at these read positions were similar with slightly more variation per read position in d-Shared.

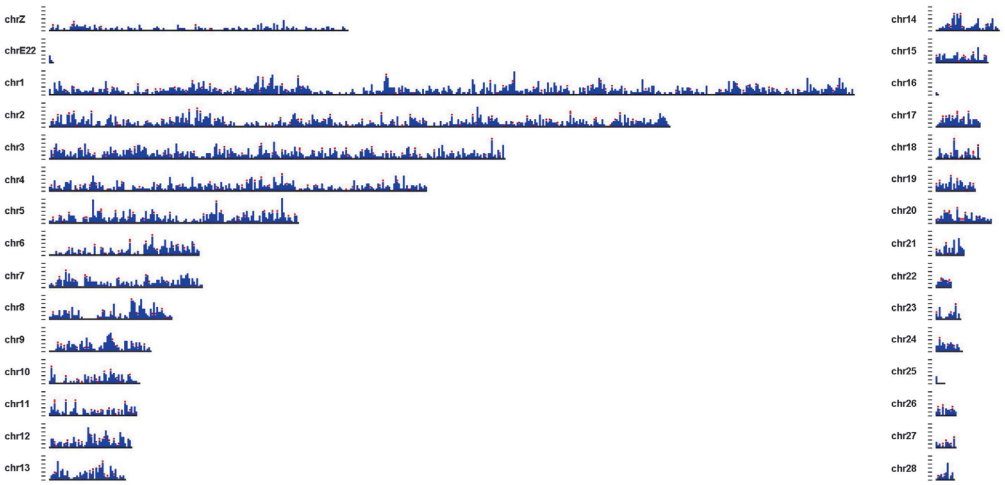


FIGURE 3.3: ***Distribution of mallard SNPs uniquely mapped on the chicken genome.***
In blue are 4272 mallard SNPs with a unique mapping position to the chicken genome (see text for mapping algorithms). 384 mapped SNPs that were selected for genotyping are in red. On the X-axis, the chicken genome in 400 kb intervals, and on the Y-axis, the frequency (0-15) of mapped mallard SNPs for a specific chicken genome interval is given.

Although reduced, also d-Shared showed a peak of the SNP distribution on read positions three to six, as we observed in d-RRL. However, TS:TV ratios for these positions were at the expected level of >2.3 indicating that most SNPs in these read positions likely resulted from true nucleotide polymorphisms. Finally, compared to d-RRL, the d-Shared subset of SNPs showed a higher average TS:TV ratio of 2.7 and indicated a relative increase of (C/T) over (A/G) transitions (Table 3.2).

DOMESTICATED VERSUS WILD MALLARD

Besides the identification of SNPs in wild mallards we also searched for nucleotide positions in the genome that show differences between the wild mallard population and the domesticated duck reference. We investigated nucleotides that where monomorphic within the wild mallard RRL consensus sequence data set but that differ from the corresponding non-polymorphic position in the domesticated duck reference. We identified 61,752 such SNPs (further referred to as d-Between) and assessed the quality of this set of SNPs by plotting the TS:TV ratio per nucleotide position and plotting the distribution of the SNPs over the 62 nucleotide positions in the sequence reads (Figure 3.2C). The distribution of SNPs predicted in the first six read positions showed

a high peak whereas from position six to 62 the number of SNPs per read position was more or less constant, only slightly increasing towards the end. The TS:TV ratios were as expected except on the first six read positions and the end, where it was lower than expected. Compared to d-RRL and d-Shared the overall TS:TV ratio of d-Between was lower, 2:1, and showed a relative increase of (C/T) over (A/G) transitions (Table 3.2).

THE DISTRIBUTION OF SNPS OVER THE GENOME

Knowing genomic positions of SNPs as genetic markers is important. Many population genetic and genetic mapping applications rely on unlinked markers. Thus, for future use in generating a mallard linkage map and performing QTL studies in domestic and wild mallard it is essential that the SNPs are widely distributed over the genome. The domestic duck genome assembly that we used as a genome reference consists of thousands of scaffolds and contigs which are not assigned to chromosomes. Estimating the distribution of SNPs across this duck genome is therefore not possible using this sequenced reference. Consequently, the closest related available genome sequence (*Gallus gallus*, chicken; divergence time 80-90 million years ago, see discussion section) was used for estimating the physical distribution of the identified SNPs. Common and high quality mallard SNPs (d-Shared) were aligned to the chicken genome and the distribution of this SNP-set was plotted over the chicken chromosomes (Figure 3.3). A total of 4,272 SNPs could be mapped to unique locations evenly distributed over the chicken genome.

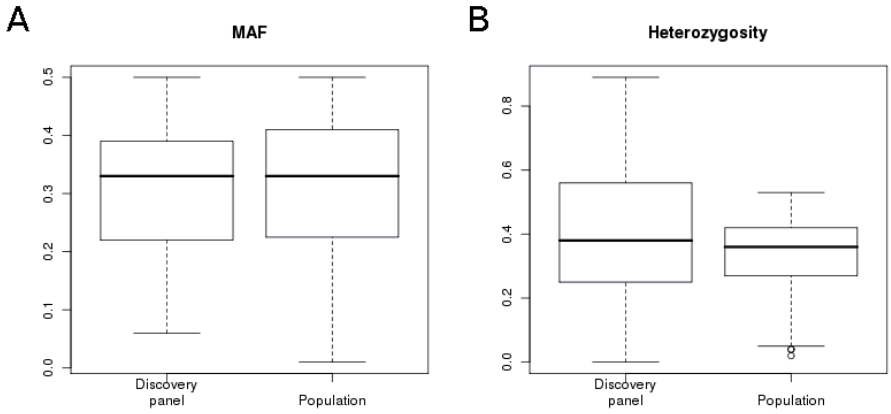


FIGURE 3.4: ***Genotyping minor allele frequency and heterozygosity distributions.***
Validation of the d-Shared subset involved genotyping of 384 selected SNPs on 765 ducks including the nine mallards that made up the SNP discovery panel. Minor allele frequency (MAF) and heterozygosity of SNPs were calculated for the discovery panel, as well as for the whole set of genotyped ducks.

SNP VALIDATION BY GENOTYPING

The d-Shared subset of SNPs was validated by genotyping an animal panel consisting of 765 mallards using 384 predicted SNPs distributed uniformly over the chicken genome (Figure 3.3). A total of 364 (95%) SNPs gave reliable genotypes in the assay, and 363 (99.7%) of these were indeed proven to be polymorphic. The average minor allele frequency (MAF) was 0.32 in the animals that made up the discovery panel and 0.31 in the whole animal panel (Figure 3.4). The average heterozygosity was 0.39 in the discovery panel and 0.34 in the whole animal panel. The allele frequencies of polymorphic genotyped SNPs in the discovery pool showed a correlation of 0.72 with those derived from the sequence data in the discovery pool of nine animals.

Discussion

This SNP study is the first large sequence variant discovery performed in mallards, as well as in any of the waterfowl. The availability of a large number of detected SNPs provides sufficient markers to study mallard population structure and migration in a population genetic framework. This large number of accurately inferred SNPs will also facilitate improved linkage maps of the mallard genome[109,145] and provide a sufficiently dense marker map to allow high resolution QTL studies in the domestic duck, further facilitating duck breeding. Furthermore, such high density linkage maps are essential for chromosomal assignment of the sequence scaffolds of the sequenced reference genome.

SNP DETECTION WITHIN A POOL OF WILD EUROPEAN MALLARDS

Initially, our study was designed to detect SNPs within a pool of wild European mallards by single-end and paired-end sequencing of a small fragment RRL. We targeted for genome libraries of sufficiently small fragments for paired reads to overlap. This allows the reads to be merged resulting in the complete sequence of the majority of the fragments in the RRL. Merged paired reads subsequently would serve as a reference genome. However, with the recent availability of a next generation sequenced domestic duck genome assembly, a reference based mapping approach became feasible, enabling a more efficient SNP identification approach. Our study shows that the overlap in generally lower-quality ends of paired-end sequence reads is beneficial in reference based SNP detection. An observed drop in the number of predicted SNPs after position 25 (Figure 3.2A) is explained by a drop in phred scores of the raw sequence data at exactly that position (data not shown). Subsequent filtering for quality scores eliminates more putative SNPs after read position 25. However, accounting for this inherent quality issue in the raw data, we observed that the number of SNPs being predicted per read position shows a tiny increase in the overlapping ends of our mate pairs whereas earlier studies[150,151,153] reported decreasing numbers of predicted SNPs per nucleotide position towards the end of sequence reads. The deamination of methylcytosines results in a thymine base. This reaction is especially frequent in CpG dinucleotides motifs, causing a much higher mutation rate from C to T than any other mutation type. As a consequence, TS:TV rations are much higher than expected, as for instance in chicken where it is 2.2:1 instead of the 1:2 ratio expected if mutations were random. A similar 1:2 for TS:TV ratio would be found in sequences if base differences were due to sequencing errors rather than

true polymorphism (same as above). The TS:TV ratio of SNPs we predicted in the overlapping ends of our sequences remains in the expected range (Figure 3.2A) suggesting that these SNPs reflect true nucleotide polymorphisms. A local decrease in TS:TV ratio would be observed if SNPs in read positions (51-61 in d-RRL and 52-60 in d-Between) were caused by randomly introduced polymorphisms (e.g., sequencing errors). Thus we expect that the predicted SNPs represent true nucleotide polymorphisms. The increased number of SNPs at the overlapping ends can be explained by local higher sequence coverage, caused by sequence overlap of paired reads, resulting in a higher representation of DNA sequence variants. A higher coverage allows for multiple observations of the variant in low quality sequences, allowing it to pass MAQ's quality thresholds to call it a true SNP[161]. As a result, even more of the rare sequence variants in these overlaps will meet the minor allele occurrence constraint in our SNP detection method. An indication that the additionally identified SNPs at the read ends involve rare sequence variants is the lower representation of these SNPs in d-Shared.

ASCERTAINMENT BIAS DUE TO LIMITED SEQUENCE DEPTH

Besides limited sequencing depth also sequence quality is a limiting factor for inferring SNPs. This is illustrated by the overall trend in the number of predicted SNPs per read position in d-RRL and d-Shared (Figure 3.2A and 2B), which mirror the decreasing trend of average base call score per nucleotide position inherently present in Illumina sequencing (as also observed in our data set, data not shown). A similar trend is not observed in d-Between because here the SNPs are predicted from differences between the reference and the discovery panel of wild mallards. Read depth is less limiting in d-Between because it is only used to provide one unambiguous (consensus) base, deviating from the reference, of sufficient quality whereas in d-RRL the read-depth has to provide sufficient base calls for both the major allele and the minor allele to be considered a SNP.

Besides the unequal distribution of identified SNPs over the read positions also the underrepresentation of SNPs with a MAC <0.2 is an indicator of a coverage limitation. Due to the limited coverage, only SNPs that are present in multiple individuals in the discovery panel have a reasonable probability to meet the minor allele representation constraint set by our SNP detection method. More common alleles will pass the representation constraint more frequently than rare alleles resulting in an overrepresentation of common alleles and an underrepresentation of rare alleles.

SNP SET QUALITY ASSESSMENT BY COMPARISON

We identified a large number of putative SNPs in the sequenced mallard discovery panel by sampling ~5% of the mallard genome. Extrapolating the total number (d-RRL + d-Between) of identified SNPs would result in a SNP every ~375bp. The actual number of true SNPs in the sets d-RRL and d-Between is expected to be lower considering the overrepresentation of predicted SNPs in the read positions one to six together with low TS:TV ratios in these read positions. Also the comparison of d-RRL with d-WGS, in which common true variants remained and false SNPs

were discarded, show that SNPs predicted in read positions one to six should be used cautiously. The distribution of d-Shared does not show overrepresentation of SNPs on position one to six. Furthermore, expected TS:TV ratios in d-Shared were observed for positions three to six and expectedly lower TS:TV ratios in position one and two due to the RRL enzyme restriction motif. Therefore we think that a considerable fraction of SNPs in read positions one to six in d-RRL and d-Between are false positives. Because standard sequencing error rates of the Illumina GAII are low (<0.5 %) in the first 20 bases of a read[167] we expect that the first six bases in our sequence dataset were affected by non-standard, systematic, sequencing errors. These are most likely resulting from a combination of inadequate separation of sequencing clusters due to the restriction tag in the RRL and an overloaded sequencing flow cell (Kees-Jan Françoijs, personal communication). This hypothesis is supported by the fact that quality scores were considered by the SNP inferring algorithm[161] and that two observations of the minor allele were required for a putative SNP making it unlikely that these numbers of false positives are due to standard sequencing errors. Low TS:TV ratios for SNPs at read position 61 and 62 in d-Between suggest that the SNPs from these positions should also be omitted. Subtracting SNPs from positions one to six (and position 61 and 62 in d-Between) results in 101,095 SNPs in d-RRL and 48,592 SNPs in d-Between that will likely yield good success rates in genotyping.

SHARED SNPS

We showed that d-RRL shares one sixth of the SNPs with d-WGS and an almost negligible number of SNPs with d-EST. ESTs only represent a few percent of the genome, of which only a fraction was sampled by the RRL. Due to this limited shared genome fraction and because SNPs in protein coding regions are rarer than in non-coding regions, a large overlap in SNPs between these sources was not expected. Between d-WGS and d-EST we observed a 2.6 times larger overlap, which can be explained by a more or less complete overlap in sampled genome fraction and a better representation of rare alleles in d-WGS. The relatively large overlap between d-WGS and d-RRL indicates a low genetic divergence between wild mallard and domestic duck. A relatively large fraction of shared SNPs between two independent studies also suggests a low false discovery rate. As stated earlier, the SNPs identified in this study will be used to study mallard population structure and movements in a population genetic framework[14]. Because the required number of genetic markers for such an analysis is small compared to the total amount of markers we generated[142], we selected SNPs from d-Shared that show an equal distribution over the chicken genome. This requirement greatly reduces the number of available markers since only a small fraction could be mapped (Figure 3.4) due to the relatively large evolutionary divergence time between chicken and ducks (80-90 million years ago, http://www.timetree.org)[168]. Genotyping of this SNP subset confirmed the expectation that SNPs that are shared between independent SNP detection studies yield a SNP set of high quality.

Conclusions

When performing SNP identification studies using next generation sequence technologies, it is important to know what limitations in sensitivity and specificity can be expected, particularly at

low sequence coverage. We show that sensitivity decreases with decreasing base calling quality towards the ends of sequence reads which can be compensated for by increasing the sequence coverage in the ends. SNP distribution and TS:TV ratio over read positions are helpful metrics for the assessment of systematic errors in the sequencing dataset in particular when statistics can be compared to a high quality subset of the data. We showed that the fairly large subset of predicted SNPs that is shared between independent SNP detection studies in wild and domestic duck is likely to represent true SNPs, and suggests a low divergence between these forms.

We present for the first time a solid and scalable genotyping environment applicable to mallards and its domestic form. Not only do we provide over 100,000 most reliable SNP markers that can be used in duck breeding and molecular genetics, we also evaluate a sub-set of 384 SNPs for use in ecological genetics. The power of this set combined with relatively low genotyping costs through down-scaling of the marker set will allow long needed studies into the molecular ecology of mallards with regard to various relevant topics, including the study of genetic variation and genetic structure, resolution of unresolved ambiguities of mallard migration systems or inference of both small and large scale movement patterns.

Methods

SAMPLE COLLECTION AND PREPARATION

Mallard DNA samples were prepared from ethanol preserved whole blood collected from nine individuals from three locations across Europe: two females and a male each from Coto de Doñana (Spain), Northern Netherlands and Ottenby (Sweden). Each of these individuals was either directly caught from the wild, or was a first generation descendant from local wild mallard parents. Ducks were sampled under the approval of the animal ethical committee of Wageningen University; the Spanish Ministry of Environment and Consejeria de Medio Ambiente of Junta de Andalucia; the KNAW (Royal Dutch Academy of Sciences) Animal Experiment Commission; and the Swedish Board of Agriculture and its Research Animal Ethics Committee. DNA extraction was performed using the Gentra Systems Puregene DNA purification Kit according to the manufacturer's instructions. Briefly, ~200µl blood was digested with 9 µg Proteinase K (Sigma) in Cell Lysis Solution (Gentra Systems) at 55ºC over night. Proteins were subsequently precipitated with Protein Precipitation Solution (Gentra Systems) and spun down. DNA from the supernatant was precipitated with isopropanol and washed twice with 70% ethanol. DNA quantity and purity were measured using the Nanodrop ND1000. Possible degradation was inspected on an agarose gel and only high quality DNA samples were used to prepare the DNA pool. Equal amounts of DNA from the nine mallards were combined into two pools of 25 µg each. Aliquots of 5 µg for each pool were digested with either *Alu*I or *Hha*I (10 units per reaction, Pharmacia). The digested pools in O'range loading dye (Fermentas) were size-fractionated on precast 10% polyacrylamide in 1xTBE with the Criterion™ Cell (BioRad). The gel was run 190 minutes at 100 volt and stained for 30 minutes in ethidium bromide solution. After staining, the target fragment size range between 110-130bp was sliced out of the gel. The gel slice was sheared by nesting a 0.5ml Eppendorf tube (with a hole in the bottom formed with a needle) containing the gel slice inside a 2ml Eppendorf tube, and centrifuged at 14000 rpm for 2 minutes. The sheared gel pieces were covered

with 300µl DNA recovery buffer (8mM Tris pH 8.0, 0.08 mM EDTA, 1.25M ammonium acetate), vortexed, and eluted at 4°C overnight, followed by 15 minutes incubation at 65°C. The slurry was divided over two Montage DNA gel extraction devices (Millipore) and centrifuged at 5000g for 10 minutes to purify the eluted gel. DNA was precipitated by adding 1/10 volume 3M sodium acetate pH 5.2, 1 volume isopropanol and 1/500 volume glycogen, washed with ethanol and resuspended in DNA hydration solution (Gentra Systems). The genomic libraries were combined and prepared using the Illumina Sample Preparation kit[169] and sequenced for 76 cycles with the Illumina GAII, Illumina Inc., USA, with a paired end module attached.

SNP DETECTION

Prior to analysis we applied quality filters to the raw reads. Due to the use of restriction enzymes *Alu*I and *Hha*I for creating the genomic libraries we expect that the sequence reads start with a 'C'. Therefore, reads not starting with 'C' were discarded as unreliable or contamination. All reads of the sequencing dataset were trimmed from the position where the average quality score dropped below 12. Reads containing a base that was called with a quality lower than 12 were discarded unless an identical copy of the read occurred in the dataset, since it is unlikely that two fragments of such a long sequence of nucleotides are identical by chance. We removed reads that - based on the theoretical raw sequencing coverage of the RRL (38X) - were more than four times overrepresented to limit the number of sequences from repetitive regions in the dataset. This is to prevent the prediction of SNPs within multi-copy genes or other repetitive regions[150,151].

As reference we used a domestic duck genome sequenced by next generation technology by the Beijing Genome Institute (Huang *et al.* in prep.). MAQ[161] was employed to map the quality filtered reads to the domestic duck genome with default parameters. Putative SNPs were tagged if the reads involved were mapped unambiguously to the reference. We filtered the MAQ[161] SNP output according to several rules: minimal map quality per read: 10; minimal map quality of the best mapping read on a SNP position: 10; maximum read depth at the SNP position: four times the actual coverage after quality filtering; minimum consensus quality: 10 (ref [153]). We required that the minor allele at a polymorphic position in the reference was observed at least two times.

EST-MAPPING

We mapped d-EST SNPs on the genome reference to identify their genomic locations whereas SNPs in d-RRL and d-WGS were predicted on an identical genome reference coordinate system. Mallard SNPs (with on average 116bp of flanking sequence) being predicted in EST sequences by the group of Alain Vignal (INRA France, unpublished data) were mapped to the reference genome using GMAP[170]. Results were filtered for SNPs that aligned with 96% sequence identity.

COMPARATIVE MAPPING

To examine the distribution of SNPs over the genome, we comparatively mapped our predicted SNPs (including 100bp flanking sequence at each side) to the repeat masked chicken genome (assembly WASHUC2). Mapping was performed using BLAT[171] with parameters -oneOff=1 -minIdentity=70.

SNP VALIDATION BY GENOTYPING

SNPs were validated by genotyping an animal panel using the Illumina GoldenGate® Genotyping assay on an Illumina® BeadXpress with VeraCode™ technology. Selection criteria for the SNPs were based on the Illumina design score (above 0.8) and the assayed 384 SNPs should distribute evenly along the chicken genome to minimise the extent of linkage between neighbouring SNPs. Oligo-nucleotides were designed, synthesised, and assembled into oligo pooled assays (OPA) by Illumina Inc. The Illumina OPA file can be found as Additional file 1 "GS0011809-OPA.opa". The 384 SNPs were genotyped in 765 animals which included domesticated ducks from a French (7 individuals) and a Chinese (189 individuals) genetic mapping population, non-*Anas platyrhynchos* duck species specimens (36 individuals), ~500 wild mallards from Europe, North America and Asia and the nine mallards that made up the SNP discovery panel. Genotyping results were analysed in Genome Studio (Illumina). Using the cor-function in R[172] the Pearson correlation between allele frequency estimated by sequencing and genotyping was calculated over 361 SNP loci that were polymorphic in the discovery panel genotyping by randomly selecting the major or minor allele.

Acknowledgements

Mallard samples for the discovery pool were kindly provided by Jordi Figuerola (Biological Station Doñana, Spain), Marcel Klaassen (NIOO Nieuwersluis, The Netherlands) and Neus Latorre-Margalef (Ottenby bird observatory and Kalmar University, Sweden). The sources of samples for the genotyping are too numerous to mention, so we thank the enthusiastic wild duck community for their assistance. Technical assistance was provided by Bert Dibbits. The analysis of the EST data was made possible by Frédérique Pitel and Christophe Klopp and his colleagues from the SIGENAE (Système d'Information des GENomes des Animaux d'Elevage) bioinformatics team. We would like to thank Nikkie van Bers for helpful comments on the manuscript, and Hendrik-Jan Megens and Ron Ydenberg for valuable discussions on the subject. This work was financially supported by European Union grant FOOD-CT-2004-506416 (Eadgene), the KNJV (Royal Netherlands Hunters Association), the Dutch ministry of Agriculture, the Faunafonds and the Stichting de Eik trusts (both in The Netherlands). Computational support was offered by the Netherlands National Computing Facilities foundation grant SH-110-08 to RHSK. JE was supported by grant V-220-08 from the Swedish Environment Protection Agency. Funding bodies had no influence on any aspects of designing, carrying out and publishing of this study.

Additional files

upon request from robert.kraus@senckenberg.de

ADDITIONAL FILE 1 – GS0011809-OPA.opa

Oligo pooled assay (OPA) data file. This file was used in the genotyping method as indicated in the methods section to generate the raw data. The genotyping assay can be re-ordered from Illumina using this file. Format is plain text, comma separated.

CHAPTER 4

Global panmixia in a cosmopolitan bird? Model selection with hundreds of genome-wide single nucleotide polymorphisms reveals world-wide gene pool connectivity

Robert HS Kraus, Pim van Hooft, Hendrik-Jan Megens,
Arseny Tsvey, Ronald C Ydenberg, Herbert HT Prins

Article in preparation for submission

Abstract

Technical advances in statistical phylogeography methods rely on genome-wide marker systems to undertake population genomic analyses, making single nucleotide polymorphism (SNP) marker panels a potentially powerful technique. Knowledge about population structure of duck species, especially mallards, has become a research priority due to outbreaks of Avian Influenza in recent years. Understanding population connectivity is important to deduce large-scale movement patterns. Traditional descriptive studies (e.g., ringing) have failed to detect clearly delineated mallard populations. We employed SNP markers comprising hundreds of loci in combination with population genetics and phylogeographic approaches to conduct a population genomic test of panmixia in mallard. Basic population genetic and phylogenetic methods suggest the absence of population structure on continental scales. Nor could individual-based structuring algorithms discern geographical structuring. Model-based coalescent analyses were employed to test models of population structuring and pointed to strong connectivity among the world's mallard population. These diverse approaches, utilising a large and genome-wide genetic marker set, support each other in their basic conclusion: a lack of clear population structuring but high connectivity, compatible with panmixia. This finding shows that zoonotic diseases such as Avian Influenza could in theory be spread between distant areas solely by the movements of wild ducks.

Introduction

One of the applications of molecular ecology is the study of geographic genetic structure of species and populations: Phylogeography[173]. Newly developed methods in statistical phylogeography (e.g., refs [174-179], and references therein) based on the (structured) coalescent[180-182] have good ability to explicitly model demographic quantities such as effective population sizes and migration rates. Model-based approaches towards the structured coalescent[176,177] are currently among the most prominent methods to infer phylogeographic and demographic scenarios on a population level. Coalescent approaches are based on a likelihood function that incorporates many parameters, and accuracy increases with increasing density of the genetic markers, and, therefore, benefit relatively more from adding more genetic loci than from adding more individuals [183,184].

Since the early 1990s some of the most widely-used markers in population genetics have been mitochondrial DNA (mtDNA) and microsatellites. Although they have provided exciting new insights, they have some shortcomings: mtDNA comprises only a single, maternally inherited locus (i.e., a specific place in the genome), while microsatellites show high levels of homoplasy and uniform mutation models do not apply[65,185,186]. In contrast to these markers, mutations observed as single nucleotide polymorphisms (SNPs) are abundant and widespread in genomes and evolve in a manner that is well described by simple mutation models[164]. A SNP normally pertains to two alleles with a minimum minor allele frequency of 1%[64]; three- and four-allelic SNPs do occur as well, though relatively rarely. The use of SNPs in molecular ecology is advocated because of their superior features as compared to, for instance, microsatellites[65]. Bi-allelic SNPs have relatively low statistical power per locus, but this can easily be compensated by genotyping a much larger number of loci[62,142].

The term 'population genomics', defined by the use of hundreds or thousands of independent genetic markers across all regions of the genome, was coined a decade ago[187,188]. SNP sets containing more than 100 markers have been used in human[149] and other model organism studies for some time, but research on non-model species has suffered from a lack of genomic resources. Few studies using SNPs have been carried out until recently, mostly genotyping far less than 100 independent (i.e., physically unlinked) loci, e.g., refs [189-192]. A recent boom in sequencing technology[193,194] has enabled the development of larger resources of thousands of SNPs[110,150,151,195]. Consequently, molecular ecological and conservation studies have begun using genotyping assay panels consisting of hundreds of SNPs[20,196-198] and related technologies to generate data sets that enable population genomic analyses[199].

The migration systems of waterfowl have been extensively studied by ringing, telemetry, morphometrics, radar tracking and isotope analysis[12]. (e.g., ref [200]). In general, migration routes are longitudinal, with northern breeding and southern wintering areas. Within both North America and Eurasia, generally distinct flyways have been described[12,16,17]. Especially in duck species, irregularities in migration routes have been described, such as individuals switching migratory routes, termed 'abmigration'[68] or 'flyway permeability'[69].

The mallard (*Anas platyrhynchos*; Anseriformes: Anatidae) is the most numerous waterfowl species with a Holarctic distribution. However, extensive analyses have not been sufficient to detect any population structure in this species. Most mallards are migratory without clear geographic directionality, though both spring and fall flights can exceed many thousands of kilometres [63]. Northern breeding birds are mostly migratory, wintering much further south, while birds breeding in temperate regions, especially in parts of Western Europe, can be resident [12]. On a continental scale, their genetic population structure is known to consist of clade A mitochondrial (mtDNA) haplotypes, mainly found in Eurasia, and clade B haplotypes, found in North-America [63,79] (also see Chapter 2). No analogous large scale data on nuclear genetic markers exist. Within continents, however, no genetic structuring can be detected (Chapter 2), which matches the results obtained using traditional (ringing etc.) procedures. These results have led to the suggestion that, at least on the continental scale, mallards constitute a single large panmictic population. Panmixia is defined by complete random mating, and is a rare phenomenon on a large geographical scale in higher terrestrial organisms[201]. In practical terms panmixia is indicated by the absence of genetic structure, usually explicitly studied in a geographical context.

In this study we employ a SNP marker set comprising hundreds of loci to conduct a population genomic test of panmixia in mallards. Diverse approaches were followed to scrutinise aspects of proposed mallard migration models. Coalescent analyses, Bayesian frameworks and model selection procedures were employed.

Methods

SAMPLING

Mallard blood from 801 individuals from 45 localities throughout all of the mallard's native range on three continents was collected on FTA cards[67]; by hunters in most cases, during their regular hunting activities. Exceptions are localities from the Faroe Islands, France, Greenland, Iran,

Norway, Portugal, Sweden and USA (Alaska). There, mallards were caught in traps, blood drops on FTA cards were sampled from the wing or foot vein, and mallards were released again (sampling approved by the animal ethical committee of Wageningen University – DEC, and the respective local organisations in the course of ongoing trapping and sampling schemes). Sampling localities are abbreviated by a four letter code in this paper: Letters one and two represent the ISO code of the country[202] (e.g., 'DE' for Germany), and letters three and four abbreviate the sampled locality (e.g., for DEWU in Germany, 'WU' stands for "Wunsiedel"). More details on sampling localities and samples can be found in Table 1 and the supplementary file "samples-details.xls".
We pooled sampling localities by hand into population units for some analyses on basis of delineated mallard specific flyways in Europe[12], Asia[16], and North-America[17] (Table 4.1). These hypothesised flyways were thus tested against the genetic data, as was previously attempted in a mallard mtDNA study (Chapter 2). Together, the two localities from Greenland were classified as a separate population/flyway unit for they are thought to constitute a separate population[12].

DNA isolation, SNP genotyping and descriptive statistics

We isolated DNA and genotyped SNPs using Illumina's GoldenGate Genotyping assay on the Illumina BeadXpress as explained previously[110] (also see Chapter 5). For each mallard, we screened SNP genotypes across 384 SNPs (accession numbers ss263068950 to ss263069333 in dbSNP[166]). Raw data was analysed in GenomeStudio (Illumina Inc.) and genotypes of 363 SNPs could be scored polymorphic among all sampled mallards in this study. The SNP set contained SNPs on all of the major chromosomes of the mallard as inferred from their mapping positions in the chicken genome[110] and did not show significant departures from neutrality or linkage disequilibrium (Chapter 5).

PHYLOGENETIC ANALYSIS

With many independent nuclear genetic markers, each representing a different locus in the genome, we expected the data not to conform to a tree-like configuration since each locus can have a separate phylogenetic history due to recombination[203]. A program to take this into account is Neighbor-Net[204], which we used as implemented in SplitsTree[205,206], version 4. An earlier paper successfully adopted this method for use with genome-wide SNP data and we used the settings described there[196]. For each individual the genotype at each SNP was collapsed into a single base character and concatenated to a sequence of 363 nucleotides. Heterozygote genotypes were represented by IUPAC nucleotide ambiguity codes, and missing data denoted 'N'.

POPULATION ASSIGNMENTS

One of the most widely used programs to determine the number of genetic clusters and assign individuals to them is STRUCTURE[207]. It maximises Hardy-Weinberg and linkage equilibria of the genotypes within groups that it determines based on the data. STRUCTURE might determine spurious genetic clusters if closely related individuals are contained in the sample due to a possible disruption of Hardy-Weinberg or linkage equilibria within otherwise good panmictic genetic clusters. To avoid this we first identified closely related individuals in our data set. Sepa-

rately for each sampling locality, and only using genotype data from these individuals localities, we assessed pair-wise relatedness (r) with Coancestry[208], version 1. As estimator for *r* we chose the dyadic maximum-likelihood estimator of Milligan[209] because it produced the best correlations with known r values in a simulation study with bi-allelic SNPs in a similarly sized SNP set (unpublished data). 95% confidence intervals (CIs) were calculated by 1,000 bootstraps. If the lower bound of the 95% CI was >0.2 we excluded one of the two individuals of the tested pair from the STRUCTURE analysis. In that way we excluded individuals that are candidates for a half-sib relationship (theoretical r = 0.25). We ran STRUCTURE version 2.3.3 in 10 replicates for all values of *K* (the number of genetic clusters) from 1 to 20 (twice the number of flyways in our population model), for 1 million steps of which the first 200,000 were discarded as burn-in. We determined the most likely number of genetic clusters (value of *K*) following the Evanno method[210]. Evanno *et al.*'s method is based on an ad-hoc statistic termed 'ΔK' derived from the first and second order rates of change of the log likelihood ($L'(K)$ and $L''(K)$) of the tested model between successive values of *K*, where $L(K)$ is short for $\ln[\Pr(X|K)]$, the logarithm of the posterior probability of the data (X) given a certain *K*. It also takes into account statistical uncertainty by incorporating the standard deviation over replicated runs, and thus allows to objectively infer *K*.

Recently, a method for detecting the number of genetic clusters and assignment of individuals was developed: Discriminant Analysis of Principal Components (DAPC[211]; *adegenet*[212] package version 1.2.8 in R[172]). This method does not suffer from equilibrium assumptions like, for instance, STRUCTURE. Hence, no closely related individuals need to be excluded from this analysis. Using the function *find.clusters* we determined the most likely number of genetic clusters in the data, using all available principal components (PCs). To calculate the probability of assignment of individuals to each of these clusters using DAPC we determined the optimal number of principal components. As advised in the manual, to avoid unstable assignments of individuals to clusters, we retained only 242 PCs (sample size divided by three), but all discriminant functions (DFs) in a preliminary DAPC run. The results were then re-iterated by the *optim.a.score* function with 25 simulations to determine the optimal number of PCs, and a final DAPC was subsequently carried out with the optimal number of PCs.

MIGRATION MODELING

We used the coalescent-based program MIGRATE-N[182,213] to estimate population parameters of our hypothesised flyways. MIGRATE-N calculates the marginal likelihoods for each model[214,215]. These can be used to evaluate multiple models when based on the same data. To compare competing models in an information theoretic framework we calculated Bayes factors (BFs) which are ratios of the marginal likelihoods between two models[179]. Information from all loci was combined into a global estimate by Bezier approximation of the thermodynamic scores. The probability of a certain model is then retrieved by dividing the exponentiated (on the base of *e*) log likelihoods by the sum of all exponentiated log likelihoods[216].

The models that we tested in the current study were defined to serve two purposes: i) testing to which extent the mallard populations are panmictic, and ii) testing to which extent there are restrictions in directionality of possible gene flow. We pooled sampling localities into hypothesi-

sed populations (i.e., hypothesised flyways) as explained in the section *Sampling* (see above). Model group 1 (models 1A-1C) represents models in which each flyway is regarded as the population unit, and 1A is denoted the 'full migration model' (i.e., all pair-wise connections of gene flow are estimated). In model 1B, only geographically neighbouring flyways may exhibit gene flow, and in model 1C this is only possible in an eastwards direction due to a hypothesised influence of the Westerly winds on bird flight[102,103] (also see Chapter 2). Model 2 is the panmictic model defined by complete absence of global population structure: all sampled individuals belong to the same population. In model group 3 a panmictic Eurasian, North-American, and Greenland population are defined. Hence, model 3A is the full migration model, 3B forces migration only with the Westerlies, and 3C does not allow migration to and from Greenland, by which Greenland isolates Eurasia from North-America via the Atlantic route. This means there is no migration from North-America into Eurasia, but only the other way around, as possibly dictated by the Westerlies. Model group 4 depicts situations in which Greenland mallards are either part of a North-American population (models 4A1 and A2) or a Eurasian one (4B1 and B2). The difference between those two possibilities (A1/A2 and B1/B2) lies in the fact that models 4A1 and 4B1 are full migration models (migration may occur in both directions), and models 4A2 and 4B2 only allow gene flow from Eurasia into North-America, but not the other way around (due to a possible influence of the Westerlies). Finally, we also set up a model 5, which comprises population assignments as inferred by DAPC (see above) and all gene flow directions were permitted. A schematic of populations and partitioning of data is shown in Figure 4.1.

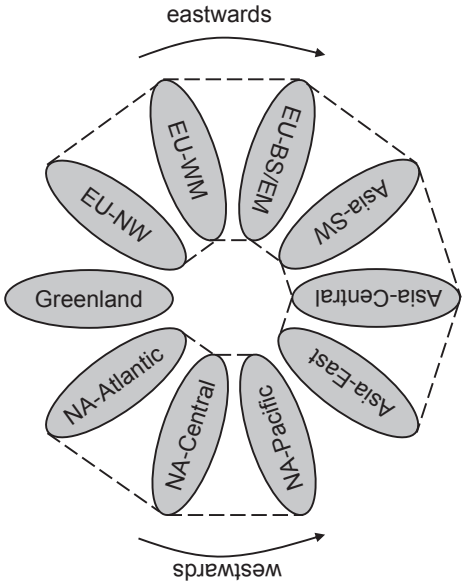


FIGURE 4.1: ***Hypothesised population structuring scheme, with flyways as basal units.***
As viewed from above the North Pole, dashed lines join several neighbouring flyways into 'land masses'. For abbreviations see Table 4.1.

TABLE 4.1: ***Sampling localities (also see main text for explanation and additional file "sample-details.xls" for more info and abbreviations).***

Sample sizes (N), pooling strategy of localities into hypothesized flyways and genetic marker performance.

Flyway	N	mono. Loci[#]	Locality	N	mono. Loci[#]	Lat[*]	long[*]
NA-Pacific	22	14	USMF	22	14	64.9	–148.9
NA-Central	22	17	CARM	20	18	50.628	–101.159
			CASL	2	—	49.542[*]	–112.056[*]
NA-Atlantic	16	13	CACO	4	—	45.58[*]	–63.845[*]
			CAEK	4	—	44.736	–75.969
			CAJC	1	—	42.324	–82.314
			CALM	7	—	43.962	–80.4
Greenland	29	20	GLIS	9	—	67.1	–50
			GLNU	20	75	64.19	–51.708
EU-NW	209	1	FIOU	19	10	65.057[*]	25.197[*]
			FOTO	24	8	62.02	–6.78
			GBAB	20	6	57.433	–2.393
			GBFE	11	20	55.901	–3.061
			GBNM	20	8	51.712[*]	–1.433[*]
			ISHV	4	—	63.748	–20.239
			NLFR	32	5	53.035	5.574
			NOBE	32	10	60.35[*]	5.323[*]
			NOSS	16	8	58.856[*]	7.332[*]
			SEAP	11	24	56.2	16.4
			SEOB	20	13	56.2	16.4
EU-WM	360	1	ATHO	25	9	48.615[*]	24.625[*]
			DEWU	27	4	50.042[*]	11.78[*]
			EETA	22	18	58.324[*]	27.178[*]
			FRAL	10	27	48.789	8.019
			FRMV	32	7	43.55	4.733
			LTVE	17	8	55.342	21.192
			PTDJ	32	8	40.664	–8.732
			RUIV	27	8	56.47[*]	41.37[*]
			RULE	31	14	59.65	28.35
			RUNO	8	—	58.167	31.517
			RUTV	19	15	57.81[*]	36.528[*]
			RUVL	29	9	55.884[*]	39.218[*]
			RUVO	31	4	59.498[*]	37.511[*]
			RUYA	25	16	56.319[*]	39.041[*]
			SILJ	19	8	46.17	14.69
			UADU	3	—	51.565	26.573
			UALV	3	—	49.825[*]	23.573[*]

(Table 4.1 continued)

Flyway	N	mono. Loci[#]	Locality	N	mono. Loci[#]	Lat[*]	long[*]
EU-BS/EM	21	13	CYLA	5	—	34.883	33.622
			GREV	16	18	40.86	25.89
Asia-SW	15	14	IRUK	15	14	36	51
Asia-Central	51	9	PKHA	17	—	34.001	72.934
			RUOM	12	41	55.961[*]	73.311[*]
			RUTO	32	13	56.526[*]	83.349[*]
Asia-East	12	28	CNLI	5	—	28.563	115.943
			RUKH	7	—	52.937[*]	138.941[*]

[#]genetic marker performance (localities with less than 10 sampled individuals not counted): amount of SNP loci that were monomorphic in all sampled individuals of the flyway/locality (e.g., only one allele appeared in the pool of individuals genotypes).

lat/long are decimal GPS co-ordinates of the sampling localities.

[*]some coordinates are averages of several near-by places, where ducks have been sampled and combined into one sampling locality.

For the MIGRATE-N analysis of each model we used Bayesian inference in version 3.2.14. The data type was specified as single nucleotide polymorphism. Starting values for Theta and M were calculated from Wright's F_{ST} as implemented in MIGRATE-N. The input data was defined as finite sites nucleotide data, and we calculated the transition/transversion ratio (2.76) as well as nucleotide frequencies (A: 0.445355, C: 0.104690, G: 0.416658, T: 0.033297) from the data and supplied them to MIGRATE-N as constants. Mutation rates were set to be constant among all loci. The prior for Theta was uniform between 0 and 0.1, and for M between 0 and 15,000. These settings were determined in several preliminary runs and performed best. Along the Markov chain the slice sampler option was used. After a burn-in of 2,000,000 steps we sampled 25,000 states from a single Markov chain, one every 20 steps. Four chains were run in parallel, with heating terms '1' (the main chain from which parameters were sampled), '1.5', '3', '10000'. The estimated mutation scaled migration parameter M was translated into the effective number of immigrants per generation (Nm) by multiplying with Theta and dividing by four (the SNPs are diploid and biparentally inherited): $Nm = Theta\_i \times M_{j->i}/4$.

Results

GENOTYPING AND BASIC STATISTICS

Virtually all genotyped SNPs are polymorphic within the flyways from which we obtained good sample sizes (North-West Europe, EU-NW, and Western-Mediterranean Europe, EU-WM). Even on those flyways with smaller sample sizes, however, only a few loci are monomorphic (a maximum of 28 in the east Asian flyway; Asia-East). As expected, this applied also within sampling localities: with larger sample sizes fewer loci remain monomorphic within a locality. Localities with less than ten sampled individuals were excluded as too small to be meaningful. Typically,

the number of monomorphic SNPs is low. Only in GLNU and RUOM did we find relatively high numbers. Table 4.1 lists all details in the columns 'mono. loci'.

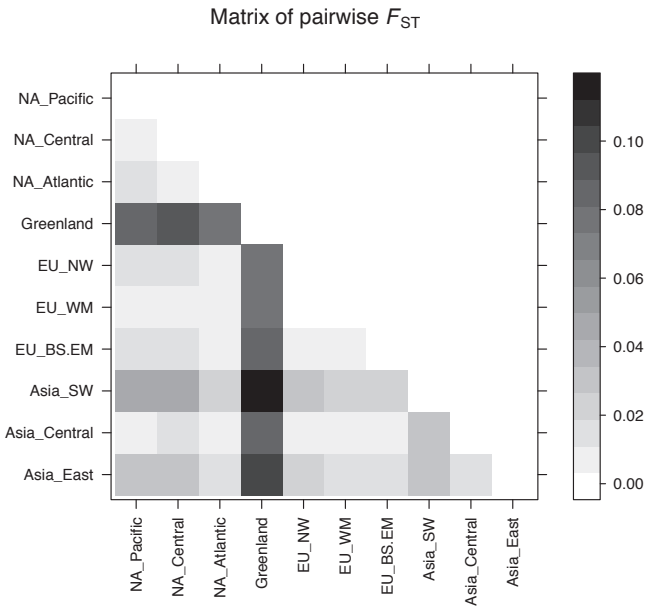


FIGURE 4.2: F_{ST} *values between flyways.*

TABLE 4.2: *Pairwise* F_{ST} *values for all flyways.*
Above the diagonal the statistical significance ($p < 0.05$) after Bonferroni correction is indicated, below the diagonal F_{ST} values are shown, with statically significant values printed bold-face for clarity.

	1	2	3	4	5	6	7	8	9	10
1) US-Pacific	-			*	*		*	*		*
2) US-Central	0.006	-		*	*			*		*
3) US-Atlantic	0.011	0.008	-	*						
4) Greenland	**0.086**	**0.091**	**0.080**	-	*	*	*	*	*	*
5) EU-NW	**0.011**	**0.012**	0.008	**0.076**	-	*	*	*	*	*
6) EU-WM	0.008	0.009	0.007	**0.073**	**0.003**	-				
7) EU-BS/EM	**0.013**	0.012	0.005	**0.085**	**0.009**	0.005	-			
8) Asia-SW	**0.043**	**0.046**	0.026	**0.112**	**0.031**	0.025	0.024	-		
9) Asia-Central	0.001	0.012	0.009	**0.082**	**0.006**	0.004	0.007	0.027	-	
10) Asia-East	**0.029**	**0.028**	0.016	**0.097**	**0.019**	0.017	0.015	0.030	0.016	-

Genetic differentiation on the continental scale is very low. F_{ST} between Eurasian and North-American samples is only 0.006, but significant ($p < 0.001$), whereas F_{ST} between Greenland and these former two geographical units is an order of magnitude greater ($F_{ST} = 0.073$ with Eurasia, $F_{ST} = 0.079$ with North-America; both statistically significant, $p < 0.001$). Also on a flyway level the Greenland population stands out as being most differentiated among all flyways with significant F_{ST} values around 0.1 with all other flyways. Most other flyway comparisons do not display significant differentiation, and those that do (mainly involving Asian flyways) are much lower in magnitude, as visualised in Figure 4.2. Table 4.2 gives details on values and statistical significance. Within land masses, only EU-NW is significantly differentiated from other flyways. Differentiation between all other flyways within Eurasia or North-America is insignificant.

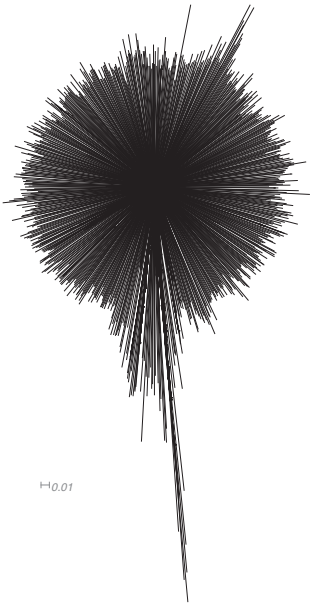


FIGURE 4.3: ***Phylogenetic network generated in SplitsTree.*** Node labels are omitted for clarity. No genetic groups can be detected (*c.f.* Figure 2 in Willing *et al.*[196] for an example of clear grouping).

PHYLOGENETIC NETWORK

If populations were differentiated from each other, the neighbor-net algorithm would display reticulate relationships more densely within less differentiated groups, and less densely in more differentiated groups. For example, using SNP data comparable in dimension to our data set, the neighbor-net method demonstrated clear genetic structuring among natural guppy populations[196]. In contrast, the network obtained from our data does not indicate any population genetic structure. The network resembles a bush rather than an unrooted tree (*cf.* Figure 2 of Willing *et al.*[196]), with complex reticulations remaining unresolved up towards the tips (Figure 4.3). No grouping can be distinguished. The sole irregularities that can be tentatively read from the result of this analysis are two discernable spikes: i) one at the bottom in Figure 4.3, and ii) a smaller

one in the top right corner. Both groups represent individuals from various geographic regions that do not resemble any recognisable pattern, e.g., Russia, Portugal, Ukraine, Norway, Alaska, Estonia, Iran or Canada for i), and the Faroe Islands, Russia, Slovenia, Iran or Great Britain for ii) (data not shown).

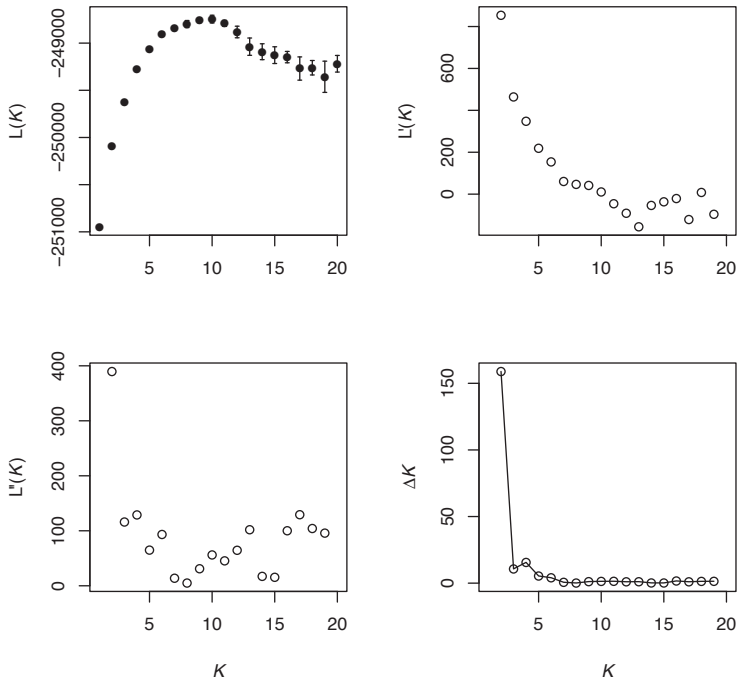


FIGURE 4.4: ***Posterior likelihood [L(K)] values from several STRUCTURE runs with different K.***
Top left panel shows L(K) means from 10 independent runs (error bars are SD). ΔK (bottom right panel) is based on the first and second order rates of change L'(K) and L''(K), based on the Evanno method[210]. The highest level hierarchical structure in the data suggest two genetic clusters (see bottom right panel). For details see Methods section.

POPULATION GENETIC CLUSTERING

As the 'traditional' means[210,217] of determining genetic clusters we employed STRUCTURE. We excluded 135 individuals from pairs under suspicion of close familial relatedness from the STRUCTURE analysis. We analysed models in which the number of genetic clusters was 1 to 20. ΔK, the statistic to detect the most likely value of K, clearly peaks at K=2. Hence, the best supported model according to the Evanno method[210] is a model with two genetic clusters (Figure 4.4). Unfortunately, the Evanno method does not allow evaluating a model of full panmixia in which K is 1 for inherent technical reasons of the ΔK statistic. The posterior probability of assignment of

individuals to the two inferred clusters was intermediate for the great majority of the individuals. No individual in the whole data set could be assigned to one of the two clusters with more than 86% posterior probability. Moreover, the bulk of individuals (521 out of 666) was assigned to their genetic cluster with 60% or less posterior probability. When K=3 (as inferred by DAPC, see below), the situation is essentially the same but individuals from Greenland form a separate cluster. The highest values for L(K) were observed for K=9 and K=10 (not significantly different from each other, p=0.5, t-test). In both these runs (and all other values of K) the Greenland individuals always form the only cluster in which individuals are not admixed by more or less equal proportions from all other clusters.

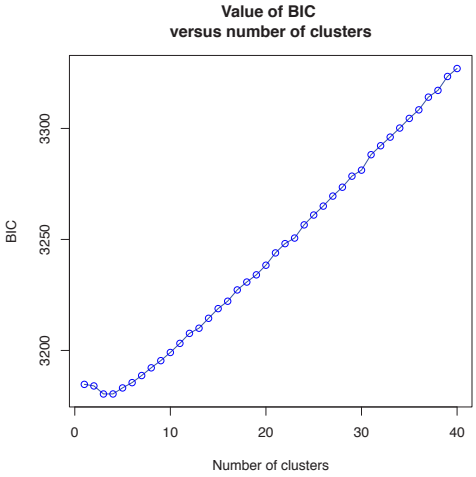


FIGURE 4.5: *Inference of the number of genetic clusters by DAPC.*
The Bayesian information criterion (BIC) as calculated during the find.clusters function of the DAPC package infers a most likely number of clusters when it is minimised. The lowest BIC values are found with three and four clusters. When two numbers of cluster have equal BICs, the smaller one (here, number of clusters = 3) is usually the correct one.

In DAPC, a number of three clusters had highest support from the data (Figure 4.5). For determining the posterior assignment probability of each individual to belong to one of these clusters 44 PCs were used, as determined being optimal according to the *optim.a.score* function of the DAPC package. Figure 4.6 shows a plot of the first two PCs calculated from the data. Clusters *One* and *Two* are relatively close to each other and cluster *Three* is more distant. Cluster *Three* is entirely composed of Greenland individuals. There is only a single Greenland individual (GLIS002) that is not assigned with high posterior probability to the Greenland cluster. It is assigned with a probability of almost 1 to cluster *One*. Most other individuals were assigned to their genetic cluster with high probabilities. Only 29 individuals had assignment probabilities of <0.9 to their respective cluster. The assignment to specific clusters is in part determined by geography. Of the 69 individuals assigned to cluster *One* (the smaller cluster of the continental clusters *One* and *Two*) with probability >0.5, 53 (76%) are from the flyway EU-NW and 13 (18.8%) from EU-WM (Table 4.3). The remaining three individuals are the one from Greenland (GLIS002, mentioned earlier) and two individuals from IRUK in the Asia-SW flyway. In total, a quarter (25.4 %) of the EU-NW flyway is assigned to cluster *One*, of which half the number of individuals stem from NLFR (26, that is, 81.3% of the total sample of 32 individuals). The remaining EU-NW samples in cluster

TABLE 4.3: Assignment of 69 mallards to cluster **One**, as determined by DAPC with > 0.5 posterior probability.

Flyway (N)[1]	count	fraction of flyway[2]	Locality (N)[1]	count	fraction of locality[2]
EU-NW (209)	53	25.4 %	FOTO (24)	6	25.0 %
			GBAB (20)	2	10.0 %
			GBFE (11)	3	27.3 %
			GBNM (20)	8	40.0 %
			NLFR (32)	26	81.3 %
			NOSS (16)	7	43.8 %
			SEOB (20)	1	5.0 %
EU-WM (360)	13	3.6 %	DEWU (27)	5	18.5 %
			LTVE (17)	1	5.9 %
			FRMV (32)	2	6.2 %
			RUIV (27)	2	7.4 %
			RUVO (31)	1	3.2 %
			SILJ (19)	2	10.5 %
Greenland (29)	1	3.4 %	GLIS (9)	1	11.1 %
Asia-SW (15)	2	13.3 %	IRUK (15)	2	13.3 %

[1]sample sizes for these flyways/localities are presented in brackets (cf. Table 4.1)

[2]percentages represent fraction of the total sample size for the respective flyways/localities

TABLE 4.4: *Details of model selection procedure in MIGRATE-N.*
Models are ranked by their marginal likelihoods as obtained by Bezier approximation.
Differences between each alternative model and model with highest rank (1A) are in column delta. Exponentiated model differences (column e^{delta}) are not presented with full precision because the values are so small that they are essentially zero (e.g., 2.2×10^{-2260} for the second best model, 1B).

model	marginal likelihood	delta	e^{delta}	probability*
1A	-205825.02	0	1	1
1B	-211028.08	-5203.06	0	0
1C	-211568.54	-5743.52	0	0
5	-231609.00	-25783.98	0	0
3A	-232656.77	-26831.75	0	0
3B	-232745.53	-26920.51	0	0
4A2	-233959.12	-28134.10	0	0
4A1	-234534.44	-28709.42	0	0
4B2	-236911.44	-31086.42	0	0
4B1	-237231.09	-31406.07	0	0
2	-242982.20	-37157.18	0	0
3C	-245189.81	-39364.79	0	0

*model probability calculated by dividing e^{delta} by the sum of all e^{delta}.

TABLE 4.5: *Migration matrix for model 1A.*
Immigration rates from "column" into "row" are given as effective numbers of immigrants (Nm) per generation as the mode of their posterior density function, and their low and high 95% posterior density bounds in brackets.

	NA-Pacific	NA-Central	NA-Atlantic	Greenland	EU-NW	EU-WM	EU-BS/EM	Asia-SW	Asia-Central	Asia-East
NA-Pacific	-	0.44 (0.31-0.57)	0.41 (0.27-0.55)	0.59 (0.44-0.73)	1.60 (1.44-1.75)	3.14 (2.92-3.30)	0.38 (0.26-0.50)	0.30 (0.18-0.42)	0.69 (0.58-0.81)	0.44 (0.31-0.56)
NA-Central	0.41 (0.28-0.54)	-	0.42 (0.27-0.56)	0.62 (0.49-0.75)	1.97 (1.83-2.11)	3.14 (2.96-3.32)	0.54 (0.41-0.67)	0.45 (0.30-0.59)	0.67 (0.54-0.80)	0.32 (0.17-0.47)
NA-Atlantic	0.45 (0.33-0.57)	0.49 (0.38-0.61)	-	0.52 (0.39-0.77)	1.71 (1.39-1.86)	2.92 (2.79-3.05)	0.53 (0.41-0.65)	0.53 (0.42-0.65)	0.83 (0.70-0.96)	0.48 (0.32-0.62)
Greenland	0.54 (0.38-0.70)	0.51 (0.34-0.67)	0.37 (0.20-0.55)	-	2.32 (2.11-2.53)	3.71 (3.54-3.89)	0.56 (0.40-0.71)	0.44 (0.24-0.63)	0.80 (0.63-0.98)	0.48 (0.33-0.63)
EU-NW	0.85 (0.10-1.59)	0.89 (0.14-1.63)	0.80 (0.06-1.52)	0.94 (0.18-1.69)	-	13.8 (13.0-14.6)	0.89 (0.14-1.63)	0.78 (0.05-1.49)	1.45 (0.70-2.21)	0.70 (0.00-1.39)
EU-WM	1.03 (0.00-2.20)	0.84 (0.00-2.03)	0.91 (0.00-2.09)	1.16 (0.00-2.33)	6.97 (5.67-8.26)	-	0.98 (0.00-2.17)	0.81 (0.00-2.00)	1.69 (0.39-2.98)	0.61 (0.00-1.81)
EU-BS.EM	0.48 (0.35-0.61)	0.49 (0.36-0.62)	0.48 (0.34-0.62)	0.61 (0.48-0.74)	2.03 (1.82-2.22)	3.25 (3.08-3.41)	-	0.42 (0.28-0.57)	0.70 (0.55-0.85)	0.40 (0.26-0.53)
Asia-SW	0.58 (0.45-0.71)	0.51 (0.40-0.63)	0.42 (0.29-0.60)	0.70 (0.53-0.84)	1.63 (1.34-1.76)	3.15 (3.01-3.28)	0.49 (0.36-0.62)	-	0.68 (0.55-0.81)	0.33 (0.21-0.44)
Asia-Central	0.52 (0.28-0.75)	0.50 (0.26-0.75)	0.42 (0.19-0.65)	0.54 (0.30-0.77)	3.10 (2.84-3.36)	5.83 (5.54-6.11)	0.45 (0.21-0.68)	0.38 (0.13-0.61)	-	0.39 (0.15-0.62)
Asia-East	0.58 (0.47-0.69)	0.59 (0.44-0.74)	0.61 (0.50-0.71)	0.69 (0.57-0.80)	1.84 (1.69-1.96)	2.72 (2.55-2.88)	0.47 (0.36-0.59)	0.49 (0.38-0.60)	0.70 (0.58-0.81)	-

One are from Great Britain, the Faroe Islands, or Scandinavia, but from these localities still more than half of the individuals are assigned to the main cluster *Two*. Individuals from outside Europe contribute to cluster *One* only in a negligible fashion (Table 4.3), while cluster *Two* is the main cluster with 659 individuals from all flyways but Greenland.

To verify, that the presence of Greenland individuals did not mask more subtle genetic structure in the remainder of the data, we repeated DAPC with Greenland excluded. This essentially lead to the same clustering as when Greenland was included in the data set (data not shown).

MIGRATION MODEL SELECTION

The evaluation of all tested models clearly ranks model 1A highest, i.e., each flyway constitutes a population and all of these flyways exchange migrants with every other one. The Bayes Factor of the second best model (1B, flyways are populations, but gene flow between neighbours only) is practically zero (2.2×10^{-2260}) and that of the other alternative models even smaller (for details see Table 4.4). Hence, model 1A has a probability of essentially 1 in comparison to the alternative models tested here and we summarise results for this model only (populations = flyways with migration between all possible pairs of flyways).

The long burn-in period we used in MIGRATE-N resulted in good convergence and narrow posterior density peaks for migration parameters (Table 4.5). Estimates of Theta were less accurate. 95% posterior densities span an order of magnitude in all flyways except EU-NW, EU-WM and Asia-Central (those flyways with largest sample sizes, N=209, 360 and 51 respectively). In the flyway with the smallest sample size, Asia-East (N=12), it even spans three orders of magnitude (see supplementary file "mallard-flyways-model-1A.pdf"). However, mode and

mean of all estimates of Theta are near identical and the density distribution symmetrical. For calculating Nm (Table 4.5, Figure 4.7) we thus use the modes of the Theta distributions. Migration rates among flyways are mostly even among pair-wise comparisons. Only emigration from EU-NW and EU-WM to all other flyways is higher, as well as immigration into these two flyways from Asia-Central.

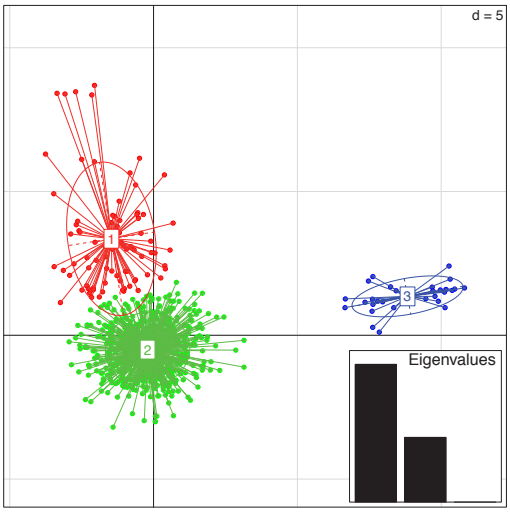


FIGURE 4.6: ***Principal component scatter plot.*** Samples are assigned to their genetic cluster by DAPC analysis. The bar graph inset displays the variance explained by the two discriminant eigenvalues used for plotting. 67% inertia ellipses are drawn for each cluster, representing the variance of both principal components. Cluster *Three* is composed of samples from Greenland, cluster *Two* of some samples from the flyways EU-NW and EU-WM, and cluster *One* contains samples from all flyways.

Discussion

ABSENCE OF POPULATION STRUCTURE

In this paper we employed basic population genetic techniques, individual-based genetic clustering algorithms and coalescent-based demographic modelling with subsequent model selection based on a data set of hundreds of SNP makers across the entire genome of the mallard duck. Samples from throughout the native range of the mallard were analysed in order to test the hypothesis of panmixia, which has been proposed previously[63] (also see Chapter 2), on several geographic scales in the mallard. F-statistics indicated hardly any genetic structure within continents, although the North Western European flyway seems to be an exception. F_{ST} between North-America and Eurasia is significant but so low in magnitude (F_{ST} = 0.006) that considerable gene flow needs to be assumed. Otherwise, only Greenland seems genetically differentiated from the remaining mallard population. From F-statistics, however, it is difficult to deduce gene flow patterns because calculating Nm from F_{ST} can only be done in a strict island model, which is usually too simplistic[218,219]. To answer the question, whether the global mallard population is an example of a large-scale panmictic population, we thus applied other techniques to solve this issue.

We used a phylogenetic network method that, in principle, has the advantage of being able to retrieve more information from, for example, DNA sequence data. In our study SplitsTree[205,206] was unable to resolve the complex reticulate structure of the world-wide mallard population structure.

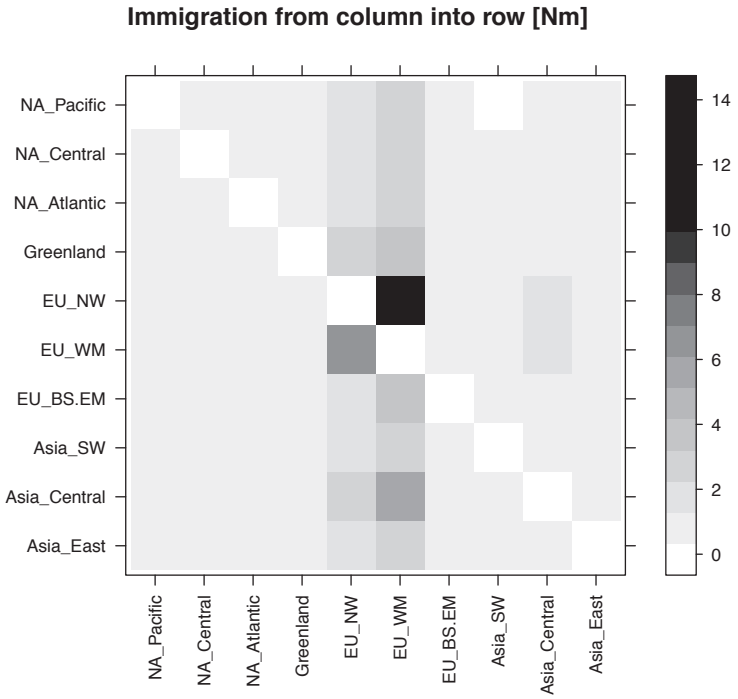


FIGURE 4.7: *Migration matrix between all flyways.*
Values are given as effective migrants per generation, and immigration is given from column into row.

A model selection procedure to infer the number of genetic clusters by using STRUCTURE[207] resulted in the best support for the model in which the number of genetic clusters was set to two. However, detailed examination of the posterior assignment probabilities indicates that STRUCTURE was not able to assign the individuals into these clusters with high probabilities. This pattern was also observed for models conditioned on the existence of larger numbers of genetic clusters. Hence, this method failed to detect significant substructuring of the global mallard gene pool, except that it consistently places individuals from Greenland as a separate group. This supports a panmictic model in mallards, with the exception of Greenland which seems significantly isolated. This is in line with the assumption that Greenland mallards only disperse within Greenland and may even constitute a subspecies *A. p. conboschas*[12].

Another method, DAPC[211], suggested subtle population structure. Similar to the STRUCTURE and F_{ST}-analysis it showed that the Greenland mallards are most differentiated from the rest of the mallards from the northern hemisphere. Furthermore, DAPC implies that some individuals from populations in northern Europe form a genetic cluster different from the main global population.

This finding confirms that some European mallards are genetically different from the remaining population (also see Chapter 2), due to, for instance, either resident lifestyle [12] or the impact of releases of farmed mallards for hunting purposes[71,72]. However, the inferred cluster *One* hold just a subset of European mallards. The majority of mallards are assigned to cluster *Two*, irrespective of their geographical origin (except for the Greenland mallards).

In contrast to the former analyses, the model selection approach with MIGRATE-N formally rejects global (model 2) and continental (model groups 3 and 4) panmixia in favour of the full flyway model (model 1). This is surprising since besides our own findings from phylogenetic and population genetic analyses, and individual-based clustering, also previous mtDNA studies[63] (also see Chapter 2) indicate panmixia at least on the continental scale. However, MIGRATE-N favours a model in which all flyways are connected in a full pair-wise fashion. The amount of effective migrants appears relatively low, being estimated at around 0.5-1 in most cases. Only emigration from the North Western Europe and Western Mediterranean Europe flyways was higher, as well as migration between these two (Figure 4.7, Table 4.5).

IS THE GLOBAL MALLARD POPULATION PANMICTIC?

The seemingly low migration rates inferred by MIGRATE-N seem to contradict panmixia. On the other hand there is migration not only between neighbouring but also between non-neighbouring flyways, e.g., between flyways on different continents, in fairly equal magnitudes. Geographical proximity has no relation to genetic proximity. Moreover, mallards from location x travel to location y with the same frequency as they travel to location z: the distance to y or z just does not matter. This uniformity of migration between distant flyways is an indication of more extensive gene pool connectivity than would be inferred from the numerical migration estimates at first sight. Important to note is that the SNP set used in this study was developed from European individuals only[110]. This will have introduced an ascertainment bias[65,220,221], possibly inflating estimates of differentiation between European samples and other regions. Although MIGRATE-N does correct for some ascertainment bias in the frequency spectrum of polymorphic nucleotide sites when the data type is set to 'SNP' (following[220], see MIGRATE-N manual), as we have done in all analyses, it cannot correct for the geographical component of ascertainment bias. Nevertheless, within all regions that were not part of the ascertainment process relative differences should be reflected accurately. Still, the genetic diversity per flyway, i.e., effective population size, will be under-estimated under such an ascertainment bias[222]. In that case the numbers of migrants between flyways from which the SNPs were not ascertained initially, are deflated because they are calculated as the product of effective population size and migration rate. After all, selecting a full flyway model over reduced flyway models in which migration would only be possible between neighbours indicates strong gene pool connectivity. This is not quite the same as panmixia because mating is not random (still locally biased), but gene flow between all locations is easily high enough to swamp all structuring that might emerge. The formal rejection of panmixia is likely due to ascertainment bias. The results of MIGRATE-N are thus compatible with a highly interconnected population structure on the global geographic scale.

Few previous studies have investigated the large-scale mallard migration system with molecu-

lar tools. A study on allozymes indicates that flyway structure resembles true population structure in North-America[223]. In contrast, two studies on mtDNA do not support currently delineated mallard flyways in Asia[63] or globally (Chapter 2). Even though results of the migration model analysis formally reject panmixia, individual-based genetic clustering does not resolve these flyways, nor delineates alternative migration routes. However, the diverse approaches followed in this study support each other in their basic conclusions: We consider the mallard population in its indigenous range, the Northern Hemisphere, essentially panmictic.

Panmixia within population units is one of the most important assumptions for basic population genetic analyses. A population can be defined in terms of being a panmictic unit and many algorithms to infer population units either use this measure directly, or via test statistics that assume panmixia themselves[207,224]. Detecting panmixia on continental or global scales is a challenge. Geographic structure is omnipresent in nature and forms the basis of the field of phylogeography[173]. Almost always larger-scale panmixia is rejected as the Null model against which data is tested. A few examples where this is not the case are known from microorganisms[225], but in higher organisms the literature is equivocal. For instance, in some marine species panmixia was proposed. Analysis of nuclear genetic markers of white shark pointed towards panmixia[114] but was questioned later[226]. Eel populations also were candidates[227,228]. But in these examples panmixia seems to be achieved by natal philopatry and aggregated mating in more restricted regions. Like marine species also birds usually have good dispersal abilities. Yet, panmixia is hardly ever observed (e.g. refs [201,229] and citations therein). To our knowledge this is the first study to report a pattern similar to panmixia for a globally distributed terrestrial vertebrate.

IMPLICATIONS FOR CONSERVATION AND MANAGEMENT

Mallards are abundant across the whole world, in some places – outside their native range – even considered an invasive pest species threatening the genetic integrity of indigenous ducks through introgressive hybridisation[230]. According to the IUCN list[231] the mallard is a species of least concern. However, recently strong population declines have been reported locally[232]. Ducks are important components of wetlands, and mallards are an abundant species in this community[70]. The preservation of wetlands for ecosystem services relies on the functionality of this community. Our finding that mallard populations are highly genetically connected implies that local declines in mallard numbers and genetic diversity can be buffered by the global population, but some alterations of local conditions (e.g., introduction of farmed mallards, massive hunting) might have cascading effects as well.

In our analysis the estimated effective number of immigrants into mallard flyways is around 0.5–1 individuals per generation. This measure is most likely under-estimated because of ascertainment bias. The number of migrants between the flyways from which the SNP set was ascertained (European flyways) was considerably higher, estimated to be ~7 into EU-WM and ~14 into EU-NW. Already 5-10 migrants per generation may be sufficient to prevent genetic differentiation[233]. If the unbiased number of migrants would be similarly high between other flyways, too, those mallard populations would not be independent of each other anymore. Such high migration rates, although not quantifiable here due to ascertainment bias, are indicated by our

analysis of population structure with individual based clustering techniques. There, population boundaries could not be delineated.

Management of wetland areas does not only have profound importance from a nature conservation point-of-view. In the last few years the spread of zoonotic diseases such as Avian Influenza, via wild birds, has gained considerable attention[2]. Aquatic birds are the natural reservoirs of this zoonotic virus[23], which is transmitted via the faecal-oral route, especially among birds that live and feed on water. Avian Influenza viruses have been shown to remain infectious in open water for several days, depending on environmental circumstances[234-236]. Our current study suggests that dispersion of AI in the wild could occur very rapidly even between distant flyways. Therefore, management and research on wetlands is necessary to increase our ability to anticipate routes of Avian Influenza outbreaks in humans[43].

POPULATION GENOMICS AND COALESCENCE MODELLING

In the current study we demonstrate the power of using a genome-wide SNP set. Over the past decade, advances in theoretical population genetics with special regard to fusing species level and population level phylogenetic and phylogeographic methods have been pursued with accelerated speed[179]. New genetic markers and genotyping technologies have enabled researchers to obtain genetic data at unprecedented throughput[62]. In our migration modelling we were able to infer migration rates with relatively narrow confidence intervals by using a genome-wide genetic marker set. However, using SNP data currently leaves the researcher with a possibly difficult challenge: ascertainment bias. The issue of ascertainment bias has long been recognised[220,237,238] and methods of bias correction been developed[239-241]. Unfortunately, these correction methods are in their infancy and applicable mainly to correct summary statistics. Working with individual-based methods likely requires adjustment within the respective programs, as is attempted by MIGRATE-N.

Ever-falling prices for genome scale studies[194] enable new projects that generate whole genome sequences of many individuals[242]. Datasets of this size can readily be created even for non-model organisms at moderate costs. Many thousands of nuclear sequences, not impaired by ascertainment bias because no pre-selection of polymorphic sites takes place, are then available to be evaluated with methods as those presented in this paper. An obstacle to using such a wealth of information will be computational demand. The analyses performed in this study, especially those with MIGRATE-N, took several months of CPU time and were only tractable through parallellisation of computational load. Future whole genomic data sets will be orders of magnitudes more complex. New approaches to optimise algorithms are crucial for bringing together the full power of molecular and theoretical population genomics. For instance, Approximate Bayesian Computation[243] is one new technology that is a candidate to overcome computational hurdles. Further, new algorithms are specifically designed to analyse whole genomic data[184,244,245]. Researches should neither be deterred by the loads of data that molecular population genomics offers, nor by technical obstacles in analysis. These will be overcome soon and allow the unprecedented analysis of models of biological complexity.

ACKNOWLEDGEMENTS

The following people contributed samples (in alphabetical order of the respective sample IDs): Ernst Niedermayer, Hans Jörg Damm (Stiftung Fürst Liechtenstein, Austria), Darren Hasson, Garnet Baker, Steven Evans, Thomas Kondratowicz, David Lamble, Garry Grigg, Aaron Everingham, Andrew Iwaniuk (Canada), Yan-Ling Son (Key Laboratory of Animal Ecology and Conservation Biology, Chinese Academy of Sciences), Nicolaos Kassinis (Game Fund Cyprus), Severin Wejbora (Lehr- und Forschungsrevier des Landesjagdverbandes Bayern, Germany), Urmas Võro (Estonia), Antti Paasivaara (Finnish Game and Fisheries Research Institute, Oulu, Finland), Jens Kjeld Jensen, Tróndur Leivsson (The environment agency, Faroe Islands), Mathieu Boos (NATURACONST@, Research Agency in Applied Ecology, Wilshausen, France), Matthieu Guillemain (Office National de la Chasse et de la Faune Sauvage, Arles, France), Anne Zeddeman (Laboratory for Infectious Diseases and Screening (LIS), National Institute for Public Health and the Environment (RIVM), The Netherlands), Andy Richardson (Safari in Scotland, Scotland), T. Cameron Manson (Scotland), Charles Bull (Northmore, Britain), Ruth Cromie (Wildfowl and Wetlands Trust, Britain), Apostolos Tsiompanoudis (Greece), Sasan Fereidouni (Friedrich-Loeffler-Institut, Germany), Bjorn Birgisson (The Icelandic Hunting Club), Ricardas Patapavicius, Julius Morkunas (Lithuania), Herman Postma (The Netherlands), Jan Bokdam (Nature Conservation and Plant Ecology, Wageningen University, The Netherlands), Alf Tore Mjøs (Museum Stavanger, Norway), Shah Nawaz Khan, Muhammad Hashim, Ahmed Khan (Pakistan Wetlands Programme, Islamabad), David Rodrigues (Escola Superior Agrária de Coimbra, Portugal), A. A. Samoilov, A. V. Karelov, A. Y. Volkov, M. V. Gavrilickev, G. V. Gonokhin, A. N. Orlov, N. D. Poyarkov, V.N. Stepanov, O. Tutenkov, V.I. Zalogin, V.N. Stepanov, Y. Konstantinov, V. S. Galtsov, Valery Buzun (Russia), Dmitry A. Sartakov (Ecological Watch Siberia, Omsk, Russia), Sergei A. Soloviev (Omsk State University, Russia), Sergey Fokin (State Informational-Analytical Centre of Game Animal and Environment of Hunting Department of Russia), Anna Palmé (Stockholm University, Sweden), David Schonberg Alm, Jonas Waldenström (Ottenby Bird Observatory, Sweden), Mitja Kersnik (Slovenia), O. V. Koshyn, I. V. Shydlovkyy, D. O. Klymysmyn, Ihor V. Shydlovkyy, Oksana Zakala (Ivan Franko National University of Lviv, Ukraine), Brandt Meixell, Danielle Mondloch, Jonathan Runstadler (University of Alaska Fairbanks, USA). Alyn Walsh and Dominic Berridge (Wexford Wildfowl Reserve, Ireland) helped in sampling set-up and coordination in Great Britain. Holly Middleton helped in coordinating sampling efforts in Canada. We would like to thank the staff of the Greenland Institute of Natural Resources for their support during our Greenland expedition, especially Aili Lage Labansen for organising our stay, and Carsten Egevang for hosting us in his laboratory. Hans Geisler supported our trapping activities at the sampling site in Nuuk. The Animal Breeding and Genomics Group (Wageningen, The Netherlands) generously hosted us in their molecular laboratory. Sylvia Kinders, Tineke Veenendaal and Bert Dibbits are thanked for helping with lab work. Peter Beerli provided outstanding support for MIGRATE-N, and Rudy Jonker helped in exploring DAPC and MIGRATE-N. This work was financially supported by the KNJV (Royal Netherlands Hunters Association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik trusts (both in The Netherlands).

CHAPTER 5

Widespread horizontal genomic exchange does not erode species barriers among duck species

Please note: This chapter has the form of a concise research letter.
More extensive information is found in the supplementary material at the end of this chapter.

Robert HS Kraus, Hindrik HD Kerstens, Pim van Hooft, Hendrik-Jan Megens, Johan Elmberg, Arseny Tsvey, Dmitry Sartakov, Sergej A Soloviev, Richard PMA Crooijmans, Martien AM Groenen, Ronald C Ydenberg, Herbert HT Prins

Article submitted for publication

The study of speciation and maintenance of species barriers is at the core of evolutionary biology. During speciation the genome of one population becomes separated from others of the same species, which may lead to genomic incompatibility over time[246]. This separation is complete when no fertile offspring is produced from inter-population matings, which forms the basis of the biological species concept. Birds, in particular ducks, are recognised as a good group of higher vertebrates for speciation studies[247]. Fertile hybrids between many duck species occur relatively frequently in nature, yet, duck species remain distinct[248]. Here we show that the degree of shared single nucleotide polymorphisms (SNPs) between duck species is an order of magnitude higher than that found previously between any pair of species with comparable evolutionary distances. Evidently, hybridisation leads to sustained exchange of genetic material between duck species on an evolutionary time scale without disintegrating species boundaries. Even though behavioural, genetic and ecological factors uphold species boundaries in ducks, there are opposing forces allowing for viable interspecific hybrids, with long-term evolutionary impact. Based on the superspecies concept[249] we here introduce the term "supra-population" to explain the persistence of SNPs identical by descent within the studied ducks despite their histories as distinct species dating back millions of years[6]. By reviewing evidence from speciation theory, palaeogeography and palaeontology we propose a fundamentally new model of duck speciation to accommodate our genetic findings. Application of the analyses presented in this paper may also shed light on longstanding unresolved general speciation and hybridisation patterns in higher organisms, e.g. in other bird groups with unusually high hybridisation rates[247]. Parallels to horizontal gene transfer in bacteria challenge our ideas why ducks have been such an evolutionary successful group of animals.

Biology has seen the proposition of several species concepts. Very influential is the biological species concept[250], often extended by the morphological species concept in cases where natural matings do not occur due to, e.g., geographical isolation: individuals are assigned to the same species if they produce fertile offspring in nature. Species boundaries are strengthened by accumulation of genomic incompatibilities preventing formation of zygotes, so called Dobzhansky-Muller incompatibilities[246,251,252]. Once evolved, post-zygotic isolation is irreversible, in contrast to pre-zygotic barriers such as mate recognition. There is much evidence that post-zygotic barriers evolve slowly in birds[246,253], potentially contributing to the high rates of hybridisation observed in this group[247] and explaining why genetic distances can be low in spite of large morphological differences[254]. For example, ducks (family Anatidae) show much hybridisation in the wild, with viable and fertile offspring[247-249]. In spite of this, duck species remain morphologically distinct. Males especially display species-specific plumage, ornamentation, and courtship behaviour (Figure 5.1).

We screened 364 single nucleotide polymorphisms (SNPs; developed for mallard[110]) in the genomes of six duck species, five of genus *Anas* and one of *Aythya*, the latter mainly for outgroup comparison: mallard (*Anas platyrhynchos*, N=197), northern pintail (*Anas acuta*, N=7), common teal (*Anas crecca*, N=9), Eurasian wigeon (*Anas penelope*, N=14), gadwall (*Anas strepera*, N=10) and tufted duck (*Aythya fuligula*, N=17). The SNPs were evaluated for minor allele frequency (MAF) spectrum, Hardy-Weinberg equilibrium and linkage disequilibrium in mallards from nine localities on three

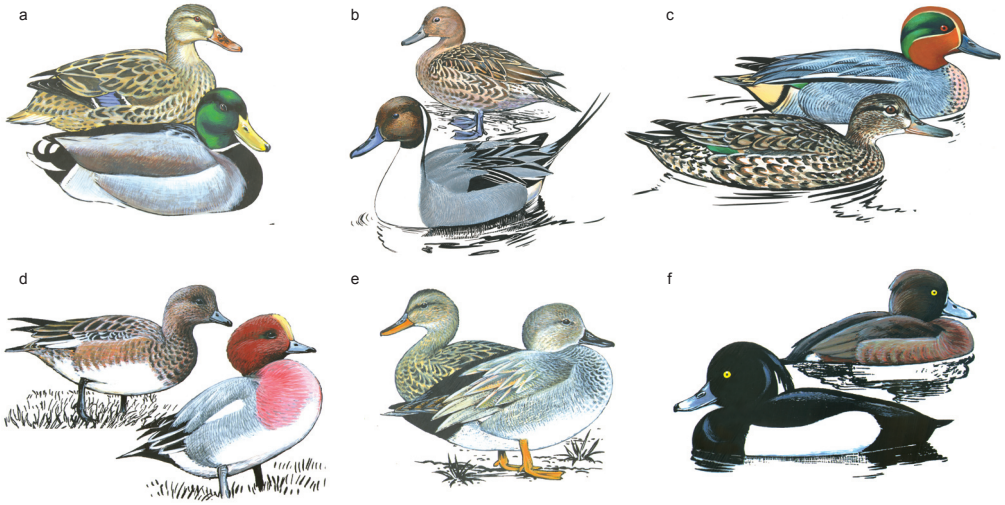


FIGURE 5.1: ***The studied duck species.***
Male and female of each of the studied duck species: a) *Anas platyrhynchos*, b) *Anas acuta*, c) *Anas crecca*, d) *Anas penelope*, e) *Anas strepera*, f) *Aythya fuligula*. Except for c) the more colourful male is in the front. Drawings are from the artwork stocks of the WWT, Slimbridge, UK, and used with permission.

continents (Supplementary Text, section 1.1). The great majority of SNPs does not significantly deviate from neutrality and are unlinked. Genotyping was successful in the non-mallard duck species with only 14-24% missing genotypes (mallard: 4%). Of 364 mallard SNPs, 86 (24%) were polymorphic in *Anas acuta*, 102 (28%) in *Anas crecca*, 60 (16%) in *Anas penelope*, 41 (11%) in *Anas strepera*, and 11 (3%) in *Aythya fuligula*. The frequencies in *Anas* species are high compared with those reported in other species comparisons. Bovines (cattle, bison and yak), for instance, have a relatively recent, Pleistocene radiation 2.5 million years ago (Mya), yet SNP sharing does not exceed 5%[255]. In sheep, also with putative Pleistocene speciation, it is estimated at only 2%[256]. The same low levels of SNP sharing also occur in invertebrate and plant species. *Drosophila pseudoobscura* and *D. miranda* show 2.9% SNP sharing[257] (divergence time 3.7 Mya[168]) while the pairs *Arabidopsis halleri/A. lyrata petraea* and *A. lyrata lyrata/A. l. petraea* share 4.7% and 1.6%, respectively[258] (divergence times <5 Mya). Given the divergence time of mallard from, e.g., *Anas acuta* and *Anas crecca* of at least 6.4 Mya[6] (Figure 5.2) they share up to an order of magnitude more SNPs than has been reported before.

Generally, the rate of SNP sharing in closely related species, as reported thus far, appears to be in the order of a few percent, at maximum. Random genetic drift usually purges polymorphisms as a function of time (generations), effective population size (N_e) and initial MAF, allowing an approximation of the time to fixation of allele frequencies under genetically neutral conditions[259]. For mallard we estimate the mean persistence time (i.e., how long the polymorphisms segregate) for alleles with the highest possible MAF to be 5.3 million years, assuming a generation time of

one year and N_e being constant at the present-day number. In the other duck species studied here it ranges between 0.8 and 2 million years. Rare alleles, e.g. MAF<0.1, are lost more quickly (Table 5.1). The probability distribution for this loss has a long tail towards longer persistence times, with 5% of the shared polymorphisms with a MAF=0.5 expected to be retained after a calculated threshold of $3.8N_e$ generations[260]. For mallard this would equate to 7.2 million years (divergence from *Anas crecca*/*Anas acuta* 6.4 Mya[6]). Thus, mallards could have retained some of their ancestral shared polymorphisms. However, *Anas acuta* and *Anas crecca* currently have much smaller N_e, and are unlikely to have retained more than 5% of their ancestral polymorphisms for periods longer than 2 and 2.6 million years (on the basis of $3.8N_e$ generations), if these species were reproductively fully isolated. Even with three times higher N_e or generation time, the number of shared SNPs between the studied duck species is higher than expected (see Supplementary Text, section 2.3).

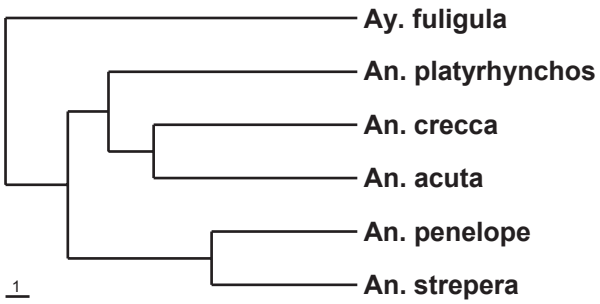


FIGURE 5.2: ***Phylogram of the studied duck species.***
Branch lengths are scaled to Mya (scale bar is 1 Mya). *Aythya fuligula* is added as outgroup (branch length shortened at the split of the genus). Redrawn from ref [6] and Javier Gonzales (pers. comm.). *An.* codes for the genus *Anas*, and *Ay.* for *Aythya*.

What can then explain the high level of shared polymorphisms? We argue that these (and other closely related) duck species are part of a superspecies complex[249] – here defined as a group of distinct species that frequently hybridise with fertile offspring (Supplementary Text, section 3.3). There is longstanding anecdotal and experimental evidence for high hybridisation rates in ducks[247-249], but molecular proof has been limited thus far. Two studies using mitochondrial DNA in the *Anas rubripes*/mallard[79] and *Anas zonorhyncha*/mallard[82] complexes confirm hybridisation between these species. These findings were corroborated by studies investigating one or two nuclear markers[100,133]. Our study, using shared polymorphisms at hundreds of independent loci across the entire genome is a more powerful means of analysing gene pool connectivity between closely related species and its results are consistent with a high level of genetic transfer between species via hybrid production and backcrossing. Such a superspecies complex, for instance mallard with *Anas acuta* and *Anas crecca*, would have a joint census population size of 31 million individuals and hence an N_e of 3.1 million (see methods for assumptions), although sub-division of this possible

superspecies due to assortative mating makes this an over-estimate. However, an N_e of 3.1 million results in a mean persistence time of almost 9 million years (for initial MAF=0.5). With an estimated most recent common ancestor at 6.4 Mya, these species could have on average retained even SNPs of lower MAF=0.2. We refer to this analysis as 'persistence time analysis' (Supplementary Text, section 4).

TABLE 5.1: *Population sizes N_c and N_e and mean persistence times in generations of the most balanced ($p = 0.5$; maximum frequency) and a rare ($p = 0.1$) SNP in each duck species.*

	Census size N_c	Effective size N_e	Persistence time	
			p = 0.5	p = 0.1
Anas platyrhynchos	19,000,000	1,900,000	5,267,919	2,470,631
Anas acuta	5,400,000	540,000	1,497,198	702,179
Anas crecca	6,900,000	690,000	1,913,086	897,229
Anas penelope	3,300,000	330,000	914,954	429,110
Anas strepera	3,800,000	380,000	1,053,584	702,179
Aythya fuligula	2,900,000	290,000	804,051	377,096

A ratio of 0.1 for N_e/N_c was assumed (see methods for more info).

Nevertheless, the ducks studied here have not only remained morphologically distinct, their genetic cluster species designation[250] is strongly supported by principal component analysis of SNP genotypes: we find clear genetic differentiation between mallard and the other duck species, as well as between the non-mallard species (Figure 5.3). We therefore propose that duck species in the genus *Anas* form a "supra-population": individuals of a sympatric superspecies complex of genetically-connected hybridising species, in which species barriers are maintained by pre-zygotic barriers. Genomic incompatibilities usually lead to irreversible post-zygotic isolation of populations, but other, reversible, barriers can also be strong drivers of speciation. Visual cues have been identified as drivers of speciation in sexually dimorphic bird species[254,261] while sexual imprinting alone can explain assortative mating in modeling studies[262].

A theory for speciation in ducks must be able to explain the observed pattern of genetic and morphological differentiation in spite of the high degree of shared polymorphisms. Paleogeographic and paleoclimatic evidence suggest that ecological conditions have been favourable for a duck radiation 6-12 Mya. This late Miocene period was warm and humid[263], but in transition towards a colder climate. Precipitation remained relatively high[264], making wetlands abundant and turning large inland salt water bodies brackish or even freshwater (e.g., Lake Pannon in Eurasia[265]). Globally, during this transition towards a colder, wet climate tropical forests were largely replaced by open grasslands[266], a habitat well suited for ducks. Fossil studies suggest that morphological change in duck species has been limited since the Pliocene[267], to which the first mallard-like fossil is dated (5 Mya)[268]. This is approximately at the suggested lower bound of divergence times of some *Anas* species[6]. We propose an *Anas*-like duck split into multiple sister morphs sympatrically and simultaneously at that time, subsequently diverging by assortative mating. Our results indicate that the resulting species still exchange portions of their genomes. We argue that since branching

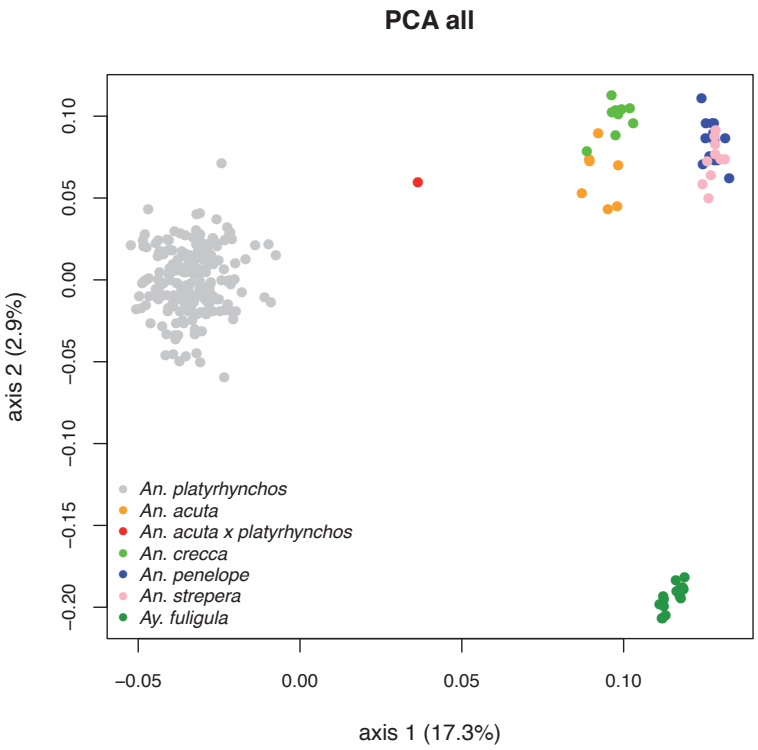


FIGURE 5.3: ***PCA analysis of genotypes of all mallard and non-mallard individuals.***
A joint calculation of PCA axes including all ducks analysed in this study is shown. Additionally, a hybrid between *Anas acuta* and *Anas platyrhynchos* was included (ANACPLA). The percent variation explained by PCA axes 1 and 2 is given in brackets. A legend that explains the colour coding is presented in the bottom left part of the graph. *An.* codes for the genus *Anas*, and *Ay.* for *Aythya*.

off of the *Anas* clade at least 6 Mya these mostly sympatric species remain separate by isolating mechanisms other than genetic incompatibilities, mostly by assortative mating.

Theoretical studies suggest that sexual imprinting can drive speciation even in sympatry[269]. Moreover, experimental manipulations clearly demonstrate that mallards can be imprinted on nearly any species of waterfowl but when raised in isolation they recognise conspecifics as mates[270]. This suggests that imprinting is important but incomplete in ducks; genetic factors also contribute to mate recognition. The presence of assortative mating and recognition mechanisms are prerequisites for sympatric speciation in the mallard superspecies complex.

The amount of shared polymorphism between the studied duck species cannot be explained by large population size only. We suggest extraordinary hybridisation rates as drivers of ongoing gene pool mixing. Gene flow continues and will allow the transfer of genetic material among duck species. Besides conservation implications (Supplementary Text, section 4), this creates large evolutionary potential, comparable to bacteria, which are able to exchange genes among different

species by horizontal gene transfer. SNP-based analysis at hundreds of independent loci across the entire genome can aid the re-examination of long-standing puzzling patterns of speciation and hybridisation in several bird groups, such as other waterfowl, galliforms, hummingbirds and woodpeckers[247], as well as in many other organisms where species pairs exhibit unusually high levels of hybridisation. It is entirely conceivable that ducks' special mode of evolution make them as adaptable and successful as they are.

Methods

SAMPLES

Mallard individuals from nine localities representing Eurasian and North American populations were sampled and their blood stored on FTA cards[67] at room temperature until DNA isolation. Numbers of samples with abbreviation codes in brackets: Eurasian samples were from Austria (25, ATHO), Estonia (22, EETA), Portugal (32, PTDJ), and three Russian localities: Yaroslavl (25, RUYA), Omsk (12, RUOM) and Tomsk (32, RUTO). North America was represented by Ontario (7, CALM), Manitoba (20, CARM) and Alaska (22, USMF). We are thankful to the following persons and institutions for providing these samples: Ernst Niedermayer, Hans Jörg Damm (Stiftung Fürst Liechtenstein, Austria), David Rodrigues (Escola Superior Agrária de Coimbra, Portugal), Brandt Meixell, Danielle Mondloch, Jonathan Runstadler (University of Alaska Fairbanks, USA), V.N. Stepanov, O. Tutenkov, V.I. Zalogin, Sergey Gashkov, Sergey Fokin (State Informational-Analytical Centre of Game Animal and Environment of Hunting Department of Russia), Urmas Võro, David Lamble, Garry Grigg, Aaron Everingham. Details are available in Supplementary_File_5.xls and further explanation is found in Supplementary Text, section 1.1. Samples from other duck species were obtained world wide from various sources (hunting bags, live-trapped, zoos) and localities. Most often blood on FTA cards was used, sometimes other tissues stored in ethanol, and also previously isolated DNA from collections. The cross species testing was applied to ducks of the following *Anas* and *Aythya* species (numbers of samples and abbreviation code in brackets): *Anas acuta* (7; ANAC), *Anas crecca* (9; ANCR), *Anas penelope* (14; ANPE), *Anas strepera* (10; ANST), *Aythya fuligula* (17; AYFU) and one F_1 hybrid between *Anas acuta* and *Anas platyrhynchos* (ANACPLA). Rolik Grzegorz (Zoo Opole, Poland), Magnus Hellstöm (Ottenby Bird Observatory, Sweden), Michael Wink, Javier Gonzales (University of Heidelberg, Germany), Dirk Ullrich (Alpenzoo Innsbruck, Austria), Kamil ihák (Zoo Dvur Kralove, Czech Republic), Marina Euler (Tierpark Lange Erlen, Switzerland), Sascha Knauf (Opel Zoo, Germany), Yang Liu (University of Bern, Switzerland), Mathieu Boos (CNRS Strasbourg, France), Crystal Matthews (Virginia Aquarium, USA), Timm Spretke (Zoologischer Garten Halle, Germany) and Valery Buzun provided these samples. Details are available in Supplementary_File_6.xls.

DNA ISOLATION

DNA extraction was done using the Gentra Systems Puregene DNA purification Kit according to the manufacturer's instructions, with modifications when handling of FTA cards. Appropriate amounts of tissue or blood on FTA cards were digested with 9 µg Proteinase K (Sigma) in Cell Lysis Solution (Gentra Systems) at 65ºC over night, or longer in case of some tissues. Proteins were

subsequently precipitated with Protein Precipitation Solution (Gentra Systems) and spun down together with the FTA card material. DNA from the supernatant was precipitated with isopropanol and washed twice with 70% ethanol. DNA quantity and purity were measured using the Nanodrop ND1000. Samples with 260/280nm absorption ratios less than 1.8 were purified again.

SNP GENOTYPING

We used Illumina's GoldenGate Genotyping assay, on the Illumina BeadXpress. The SNP set consisted of 384 SNPs from Kraus *et al.*[110] ("mallard 384 SNP set"). SNPs are numbered according to their dbSNP accession numbers from ss263068950 (SNP 0) to ss263069333 (SNP 383). Raw genotyping results were analysed in GenomeStudio (Illumina), and SNP clusters adjusted by hand. The respective OPA (oligo pooled assay) and cluster files can be found online with this paper (files GS0011809-OPA.opa and mallard_cluster_file.EGT).

SNP SET EVALUATION

We assessed technical and biological properties of the SNP set in mallards:

I) *Minor Allele Frequencies and Heterozygosity*

For each of the nine localities we counted the occurrences of each of the two alleles. The count of the allele occurring less frequently (minor allele) was divided by the total number of alleles, giving the population wide frequency of minor alleles per locus (minor allele frequency, MAF). Additionally, we counted heterozygote individuals as fraction of all individuals (observed heterozygosity, H_{obs}).

II) Hardy-Weinberg Equilibrium

Each locus was tested for deviation from Hardy-Weinberg equilibrium in each locality with the software Arlequin 3.5.1.2[128] using the analog to Fisher's exact test for arbitrary table size[129] (1,000,000 Markov chain steps, 100,000 dememorisation steps).

III) Linkage Disequilibrium

Per locality, pairs of SNP loci were tested for presence of linkage disequilibrium (LD) in Arlequin. The implemented likelihood-ratio test[271] employs the EM algorithm[272] to infer haplotypes from unphased genotypic data to test for statistical significance of LD. Repeated use of a SNP in multiple statistical tests requires a correction of the significance level α. In our 364 SNP data set each SNP is involved in 66066 pairwise tests, significance levels for LD are thus Bonferroni corrected.

IV) Physical SNP Locations Inferred from Chicken Genome

We searched the mallard SNP positions on the chicken genome (from Kraus *et al.*[110]) in Ensemble[273] (accessed 5.7.2010) for chicken gene information using bioconductor[274] with biomaRt in R[172]. We recorded if a SNP was situated in a gene, or even intron.

CROSS-SPECIES COMPARISONS

Genotypes from non-mallard species were assessed for their MAF and H_{obs} as described for the mallard samples. We expected technical genotyping success to decrease with increasing evolutionary distance due to possible variation in the flanking regions of the SNPs. Thus we compared the amounts of missing data (no genotype called at the SNP loci) of each species with that of the mallard (all nine mallard localities pooled) using the prop.test function in R[172] which applies a χ^2 test. We used Yates' continuity correction[275] because counts of missing data were most often <10.

PERSISTENCE TIMES OF SNPS

The equation for mean persistence time t(p) is a combination of the time to loss and to fixation[276,277]. It can be written as $-4N_e[(1-p)\ln(1-p) + p\ln(p)]$ where p denotes the initial MAF and N_e the effective population size[259]. Estimates of current census population sizes (N_c) of the investigated duck species were taken from the BirdLife species fact sheets[278]. Only upper estimates were used when a range was given. A ratio of 0.1 for N_c to N_e was used[279,280] (for further details see Supplementary Text, section 2.3).

MULTIVARIATE GENETIC CLUSTERING OF GENOTYPES

We tested for genetic similarity of individuals using principal component analysis (PCA) on their genotypes with the program smartpca from the Eigenstrat package[281] with default settings, but outlier removal switched off. The analysis was repeated for every new subset of the data.

Acknowledgements

Persons and institutions who provided duck samples used in this study are listed in Supplementary_File_5.xls and Supplementary_File_6.xls. Technical assistance with genotyping was provided by Bert Dibbits. Daniël Goedbloed helped with the software package Eigenstrat. We thank Michael Turelli and Carlo Dietl for discussions. Javier Gonzales provided unpublished data on divergence times of duck species, and Brian Cade helped with statistics. The WWT, Slimbridge, UK, provided drawings for Figure 5.1. This work was financially supported by the KNJV (Royal Netherlands Hunters Association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik trusts (both in The Netherlands) and the Swedish Environmental Protection Agency, grants V-220-08 and V-205-09.

Supplementary text

TABLE OF CONTENTS

1 Supplementary Methods

In this section we describe methodology not detailed in the main paper or in the online methods section. We elaborate on specific details and explain the reasoning behind certain assumptions.

1.1 SAMPLES

We analysed 212 mallard samples obtained from nine localities representing Eurasian and North American populations (see main text for information). Preliminary multivariate clustering of single nucleotide polymorphism (SNP) genotypes (see online methods section) positioned 15 of these individuals far outside the mallard species cluster, sometimes well within the clusters of other duck species (Supplementary_File_1.pdf). We discarded these 15 individuals as mislabelled because they showed obvious deviation from their putative genotypic species cluster. We analysed 67 samples of five other duck species obtained from various sources (from hunting bags, live-trapped, zoos) and localities all over the world. Using the same procedure as with the mallard set, we identified nine of these samples as apparently mislabelled. These samples were excluded from all subsequent analyses (Supplementary_File_2.pdf).

1.2 PERSISTENCE TIMES OF SNPS

To calculate the persistence time t(p) of a SNP, an estimate of the effective population size (N_e) from the census population size (N_c) is required (see main text). The ratio between N_e and N_c for species of dabbling ducks has to our knowledge not been studied, but this ratio is probably rather low as most census estimates are based on winter counts made several months before the breeding season starts and most mortality may occur before breeding. Further, dabbling ducks are generally r-selected and their population sizes fluctuate greatly by swift responses to benign and detrimental conditions, respectively[282,283], with N_e being dominated by the smallest values[259]. Estimated N_e:N_c ratios from white-winged wood ducks range between 0.052 (genetic measurements) and 0.094 (demographic measurements)[280]. Thus, we use a ratio of 0.1 as a conservative estimate (on the high side) for the ratio of N_e to N_c.

The onset of the current Ice Age about 2.5 million years ago (Mya) brought multiple cycles of glaciation, often lasting more than 100,000 years. This must have dramatically affected duck habitat distribution for thousands of generations. Large-scale range shifts, changes in migratory pathways, and local bottlenecks in population size are thus very likely to have occurred repeatedly during this time, all of which acted to reduce the long term effective population sizes by fragmentation of populations[284], even if the total number of ducks happened to remain stable.

1.3 GENETIC ADMIXTURE ANALYSIS

A Bayesian genetic clustering algorithm as implemented in the software STRUCTURE[207] (version 2.3.3) was used to test for genetic admixture, i.e., the incorporation of genes from one discrete population into another. Two datasets were analysed: i) all mallards and other duck species (the same individuals as analysed by PCA, see Figure 5.3 in main text); ii) the non-mallard species (*cf.* Figure 5.S1) plus the putative hybrid between mallard and *Anas acuta*. A value of K=6 simulated clusters (as many clusters as species) was chosen in the analysis of all ducks (i), and consequent-

ly K=5 when mallard was excluded (ii). Default settings were used with the admixture model of STRUCTRE, run for 300,000 steps (the first 100,000 discarded as burn-in). Additionally, we compared the results of the STRUCTURE analysis with those of the program InStruct[285]. The same datasets and settings were used, including the default settings, with the same values for K. Mode 1 – "infer population structure only with admixture" – in InStruct was chosen because it is most comparable to the program STRUCTURE as explained in its manual. The dataset containing only non-mallards (K=5) was also run for the same amount of iteration steps. The larger dataset, mallards and non-mallards combined (K=6), was run substantially longer because the Markov chain converged very slowly (2,000,000 steps, of which 1,000,000 were discarded as burn-in).

2 Supplementary Results

2.1 TECHNICAL AND BIOLOGICAL PROPERTIES OF THE "384 MALLARD SNP SET"

2.1.1 *Descriptive Statistics*

In Kraus *et al.*[110] the general properties of allele frequencies and heterozygosity have already been presented for a large set of mixed wild and domestic mallards. Of the 384 SNPs in that set 20 failed technically. Here, we studied the detailed patterns of minor allele frequencies (MAF) and observed heterozygosity (H_{obs}) of the remaining 364 SNPs in each of the nine localities, and among all these specimens pooled. In our total mallard data set, only one locus was monomorphic (SNP 5). Within the respective localities, we found between 9 loci (ATHO) and 47 loci (CALM) to be monomorphic. Table 5.S1 gives values for individual localities.

TABLE 5.S1: ***Minor allele frequency (MAF) and observed heterozygosity (H_{obs}) in nine wild mallard populations.***

	ATHO	EETA	PTDJ	CALM	CARM	USMF	RUOM	RUTO	RUYA	total
Monom.	9	18	8	47	18	14	41	13	16	1
MAF										
< 0.2	129	131	122	135	129	136	146	128	125	120
< 0.2 [%]	35.44	35.99	33.52	37.09	35.44	37.36	40.11	35.16	34.34	32.97
H_{obs}										
< 0.2	69	77	68	71	64	80	96	75	61	
< 0.2 [%]	18.96	21.15	18.68	19.51	17.58	21.98	26.37	20.6	16.76	
>0.8	0	0	0	13	0	0	9	0	0	
>0.8 [%]	0	0	0	3,57	0	0	2,47	0	0	

For locality codes see methods section of the main text. The numbers represent counts of SNP loci out of 364 wild mallard SNPs, or percentages were indicated. "monom." is "monomorphic".

During development of the "384 mallard SNP" set it became apparent that it was biased towards higher MAFs[110]. This ascertainment bias is expected and very hard to prevent in natural populations[164]. However, when investigating the SNP loci for MAF, we found that when all localities were pooled, 120 of our 364 loci (33%) had low MAFs (<0.2; taken as an arbitrary threshold

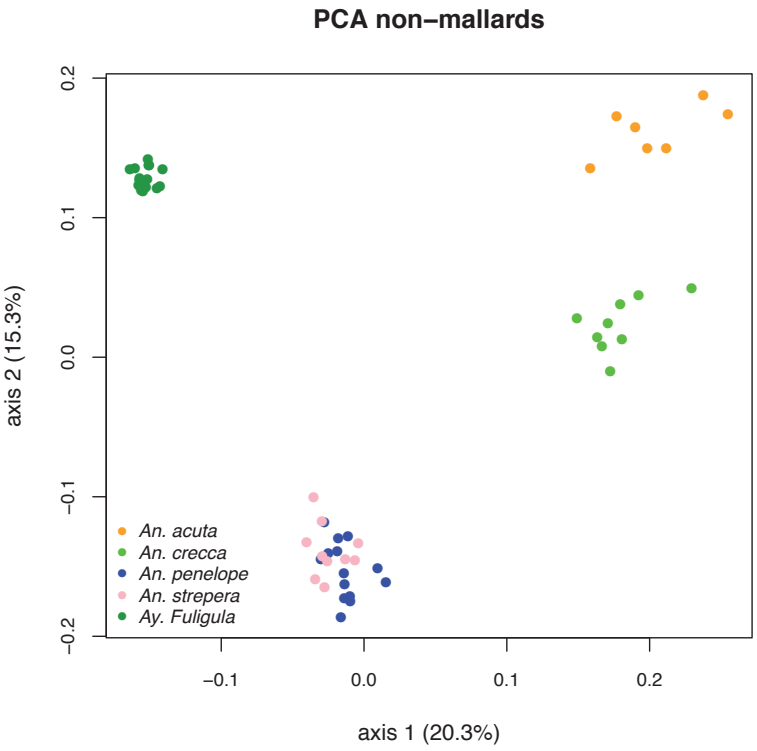


FIGURE 5.S1: ***PCA analysis of genotypes of all non-mallard individuals.***
The program smartpca from the Eigenstrat package was used to calculate multivariate eigenvectors of the mallard genotypes. The first two eigenvectors for each individual are plotted and colour coded by species: *An. acuta* (*Anas acuta*, ANAC), *An. crecca* (*Anas crecca*, ANCR), *An. penelope* (*Anas penelope*, ANPE), *An. strepera* (*Anas strepera*, ANST), *Ay. fuligula* (*Aythya fuligula*, AYFU).

between intermediate and low MAF). The SNPs with low MAF were seen in the Portugal sample (PTDJ; 122 loci, 34%), while most were present in the sample from Omsk (RUOM; 146 loci, 40%). Details about each locality can be found in Table 5.S1. Histograms of the MAFs are available in Figure 5.S2.

Patterns of H_{obs} can be indicative of several shortcomings in usability of SNP loci, such as – among many possibilities – false positive SNPs due to nucleotide variation in orthologous sequences in repetitive regions on the same chromosome or selection against heterozygotes. The former would be indicated by comparably high H_{obs}, the latter by low H_{obs}. We set arbitrary thresholds for high and low H_{obs} to 0.2 and 0.8 respectively. High H_{obs} values were present only in the Ontario sample (CALM; 13 loci, 4%) and Omsk (RUOM; 9 loci, 2%). Several localities had low H_{obs} values, ranging between 61 loci (17%) at Yaroslavl (RUYA) and 96 loci (26%) in Omsk (RUOM). Full details are given in Figure 5.S3 and Table 5.S1.

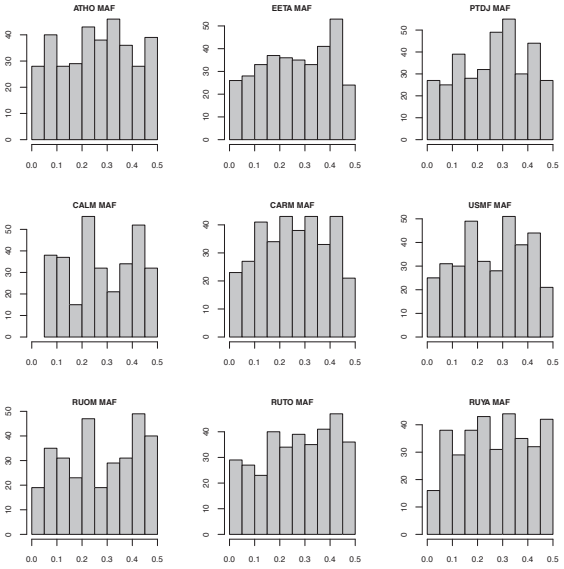


FIGURE 5.S2: ***Minor allele frequencies of SNPs of mallards from nine locations*** (see text for locality abbreviations).

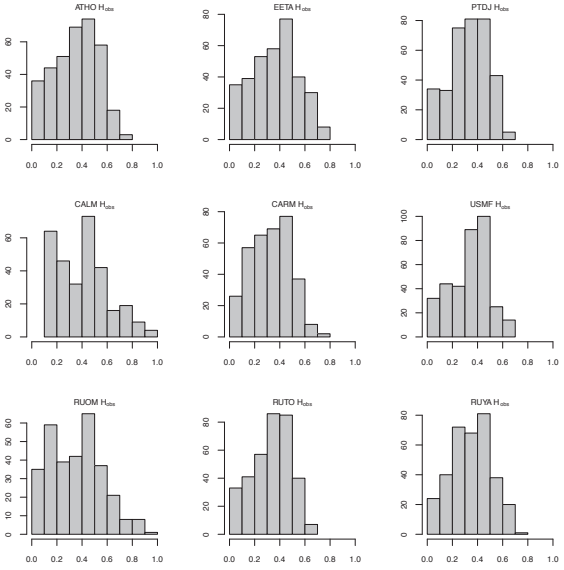


FIGURE 5.S3: ***Observed heterozygosities of SNPs of mallards from nine locations*** (see text for locality abbreviations).

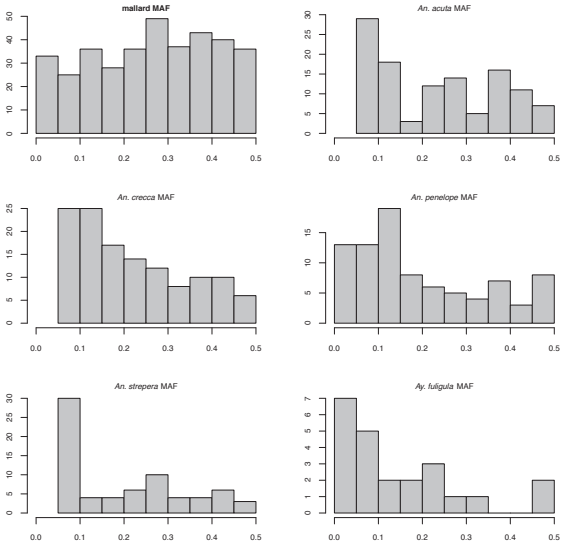


FIGURE 5.S4: ***Minor allele frequencies of SNPs in mallards (nine study localities pooled) and the other five study species.***
Abbreviations as in Figure 5.S1. The mallard histogram serves as a comparison.

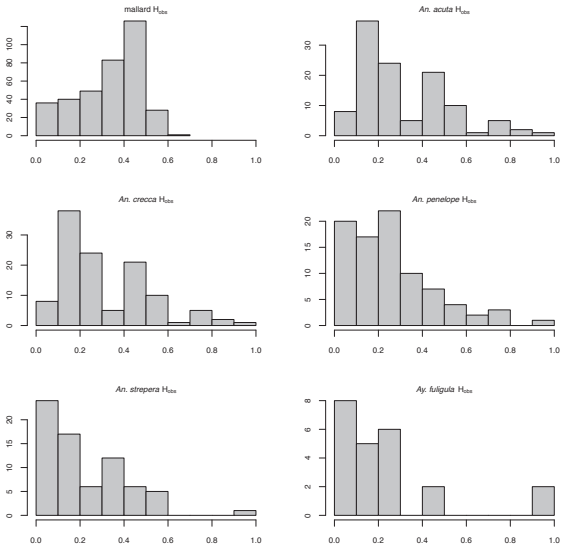


FIGURE 5.S5: ***Observed heterozygosities of SNPs in mallards (nine study localities pooled) and the other five study species.***
Abbreviations as in Figure 5.S1. The mallard histogram serves as a comparison.

2.1.2 *Hardy-Weinberg Equilibrium*

One of the important assumptions of the Hardy-Weinberg equilibrium (HWE) is that a population is panmictic. We tested for deviations from HWE for each locality individually to account for possible substructure in our data, indicating a violation of this condition. We found significant ($p<0.05$) deviation from HWE at some loci in all localities, numbering between 3 (Ontario; CALM) and 21 (Estonia; EETA) loci, about the level expected by chance ($364 \times 0.05 = 18$). Very few SNPs deviated significantly from HWE at more than half of the localities. SNPs 35, 129 and 306 were significantly out of HWE at four localities (with p-values just above 0.05; or monomorphic at other localities). SNPs 103, 235 and 280 were out of HWE in six localities, and SNP 83 in seven localities. Overall, the great majority of SNPs fulfilled HWE conditions.

2.1.3 *Linkage Disequilibrium*

We examined all of the possible 66,066 pairs of SNPs for linkage disequilibrium (LD). We arbitrarily set the criterions that if a SNP locus was in LD in at least five localities with any other of the SNPs in the test, it was concluded to be in global LD. This does not preclude the possibility that certain ecological or demographic conditions at some localities may lead to LD, but this would be locality specific and not a general property of the SNP locus itself in the species.

We found that all of our SNPs were in Linkage Equilibrium (LE) in at least five localities, and 87.7% were in LE in all of the nine localities, after Bonferroni correction of the p-values. Six SNP pairs involving 11 SNPs were in LD at three localities (SNP pairs 25/118, 42/330, 64/322, 124/127, 124/333 and 286/293). Based on the mapping positions of our mallard SNPs in the chicken genome (data of Kraus *et al.*[110]), and assuming no major re-arrangements, we find that only one (the SNP pair 124/127 on chromosome 3) of the six SNP pairs has its putative linkage partner on the same chromosome. The loci of SNPs 124 and 127 are about 8.6 million bp apart from each other, which practically precludes physical linkage. Further lowering the threshold for global LD to LD in at least two localities gives 430 candidate pairs. Given the possible 66,066 pairs in the 364 SNPs this is still a tiny fraction of 0.65%. Further, based on a query using Ensemble[273], we also know that none of the 384 SNP loci lie in a gene.

2.2 CROSS-SPECIES APPLICATION

MAFs of the SNPs in species other than mallard were mainly biased towards lower values, but in almost all species we also found high MAFs (>0.3) in substantial numbers (Figure 5.S4). A similar picture emerges for observed heterozygosity (Figure 5.S5), where each duck species had SNP loci representing a wide range of heterozygosity values. Expected heterozygosities based on HWE proportions were not calculated because the samples were from different parts of the world, violating the important random mating assumption. Many candidate loci for SNPs showed low MAFs, and some of those may have been caused by genotyping errors. In order to avoid false positives, we conservatively accepted in non-mallard species only those SNPs with MAF > 0.1. We found 86 such SNPs in *Anas acuta*, 102 in *Anas crecca*, 60 in *Anas penelope*, 41 in *Anas strepera*, and 11 in *Aythya fuligula*. Figure 5.S6 summarises the sharing pattern of the four Anas species with the mallard.

A comparison of missing data between mallard and the other species shows that few loci have significantly elevated levels of missing data: *Anas acuta* 19 loci (5.2%); *Anas crecca* 23 loci (6.3%); *Anas penelope* 28 loci (7.7%); *Anas strepera* 26 loci (7.1%). In *Aythya fuligula*, however, we found more than twice as many loci with elevated levels of missing data: 61 (16.8%). Overall, the extent of the increase of missing data varied substantially, from 1.8 fold (SNP 235 in *Anas crecca*, *Anas penelope*, *Anas strepera* and *Aythya fuligula*) to 98.5 fold (SNP 9 in *Anas acuta*, *Anas penelope*, *Anas strepera* and *Aythya fuligula*).

2.3 ALTERNATIVE SCENARIOS FOR PARAMETERS IN SNP PERSISTENCE TIMES

We made some assumptions to calculate SNP persistence times (see above and the main paper). We set generation time to one year, because all studied species can breed after their first winter. Further, we set N_e to 10% of the current estimated maximal world wide census population size (see above). Moreover this may be an under estimate if today's duck populations are much smaller than they used to be in the distant past (e.g., due to human hunting and habitat destruction). We recapitulate from the main text that the probability distribution for fixation of a SNP has a long tail towards longer persistence times, and 5% of the shared polymorphisms of highest initial MAF of 0.5 are expected to be retained after $3.8N_e$ generations[260]. In a scenario where generation time is three years instead of one year, or where N_e is three times larger than our assumed values, the persistence times of this 5% fraction of SNPs with MAF=0.5 for *Anas acuta* and *Anas crecca* (6.2 and 7.9 million years) just exceeds their divergence time from mallard (6.4 Mya[6]). On the other hand,

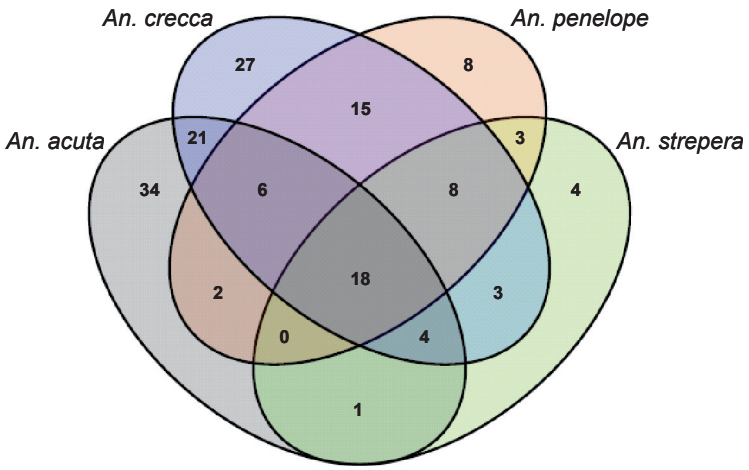


FIGURE 5.S6: *Venn diagram of shared SNPs with mallard by the four other Anas species.*
A core of 18 SNPs was polymorphic in all four *Anas* species. The closer phylogenetic relationship of *Anas acuta* and *Anas crecca* to mallard is reflected in their polymorphism sharing pattern. Abbreviations as in Figure 5.S1.

under these scenarios *Anas penelope* and *Anas strepera* would not have retained more than 5% of SNPs with MAF=0.5 after 3.8 and 4.3 million years, respectively, at a divergence time from mallard of minimally 8 Mya[6]. In conclusion, it seems the number of shared SNPs between the studied duck species exceeds what is likely under the neutral theory even when conservatively high estimates of N_e (from the upper bounds of the official counts) and conservatively low divergence times (mean times minus standard deviation of the values presented in Gonzalez *et al.*[6]) are assumed.

2.4 GENOTYPIC DIFFERENTIATION BETWEEN MALLARD AND OTHER DUCK SPECIES

We plotted the results of a series of principal component analyses (PCAs) for several combinations of individual mallards and other species genotypes. All plots are based on the first and second PCA axes. Other axes were investigated visually but did not provide further insight (data not shown). No clear genetic clusters among mallards were discernable in this analysis when analysed separately (Figure 5.S7). Geography had no influence on genetic similarity. A few individuals seem to lie

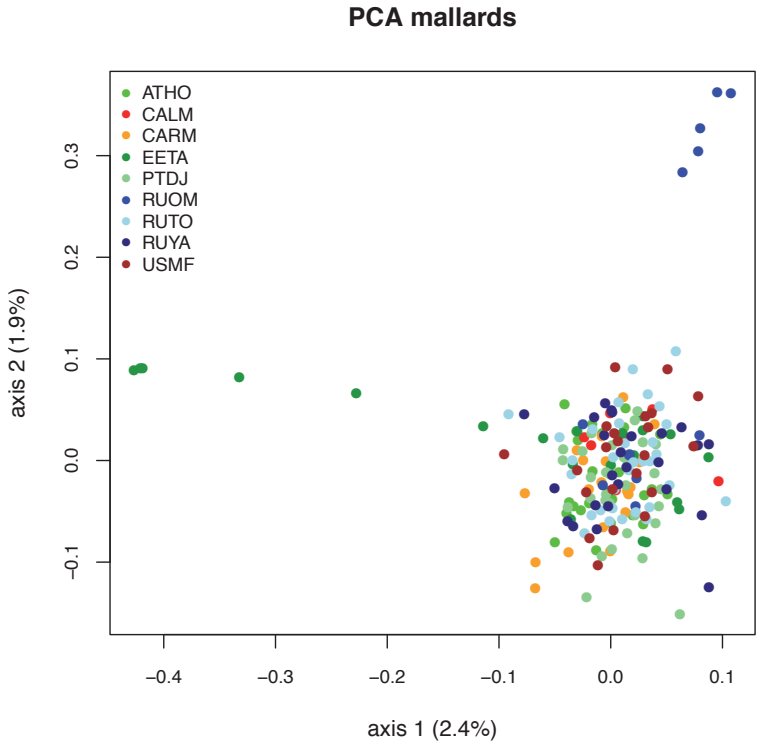


FIGURE 5.S7: ***PCA analysis of genotypes of all mallard individuals.***
This plot was created analogous to Figure 5.S1. The first two eigenvectors for each individual are plotted and colour coded by locality (for abbreviations see main text).

a bit outside the main cluster, but note that the scaling of differences between mallard individuals in this PCA is different from the scaling in analyses involving other duck species (cf. Supplementary_File_1.pdf and Figure 5.3 from main paper). The non-mallard species form distinct clusters if analysed together (Figure 5.S1): *Anas penelope* and *Anas strepera* form one cluster and are hard to distinguish. *Anas acuta* and *Anas crecca* each form their own specific clusters. *Aythya fuligula* clearly lies outside these clusters. When mallard individuals and non-mallard individuals are analysed jointly in this way (Figure 5.3 from main paper), mallards are clearly distinct from the other species. A putative hybrid between *Anas acuta* and *Anas platyrhynchos* is placed exactly in between its assumed parental species, thereby confirming its supposed hybrid status.

2.5 GENETIC ADMIXTURE

The genetic clustering software STRUCTURE is able to estimate the degree of genomic admixture of individuals, i.e., how large a fraction of the genome of each individual stems from which genetic cluster. When all six duck species were analysed jointly, all non-mallard individuals were assigned to the same cluster (Supplementary_File_3.pdf). *Anas acuta* individuals in particular showed partial mallard genome admixture, and many mallard individuals displayed some admixture from non-mallards, most notably, e.g., CARM009, RUTO017 or USMF017. Undetermined genetic sources contribute to the genomes of many analysed mallards (see all non-purple and non-light blue bars in Supplementary_File_3.pdf). Within the mallard samples, several undetermined clusters were created by STRUCTURE (see blue, red, green and orange bars in Supplementary_File_3.pdf). As is obvious in the PCA analysis, the mallard-*Anas acuta* hybrid displays a genome mix close to 50% mallard, 50% non-mallard.

With mallards excluded, STRUCTURE assigned *Anas penelope*, *Anas strepera* and *Aythya fuligula* individuals to their species clusters, although one *Anas strepera* individual (ANST001) was almost fully assigned to *Anas penelope*. *Anas acuta* and *Anas crecca* were lumped into one cluster, and the hybrid was correctly assigned to that cluster by only 50% of its genome (Supplementary_File_4.pdf). Excluding the hybrid from analysis did not alter the assignment of these two species to the same cluster (data not shown). Mixed ancestry of several individuals was found in a similar manner to that found when analysed with the mallard data. Besides the cases within *Anas acuta* and *Anas crecca*, ANPE013, ANST008 and ANST010 were proposed to contain genomic portions of other genetic clusters.

The same data sets were analysed with comparable settings in the software InStruct. This software does not assume HWE in the inferred populations, and yields qualitatively similar results as the STRUCTURE analysis, however (data not shown).

3 Supplementary Discussion

3.1 THE "384 MALLARD SNP SET" IS BIOLOGICALLY MEANINGFUL

The evaluation of the genotyping results from 197 mallards from three continents shows that the SNP set performs very well in this species. Even though minor allele frequencies (MAFs) were biased towards higher values, the "384 mallard SNP set" includes a good amount of lower MAF loci. In all studied localities about one third of the SNP loci were of MAF <0.2. Observed hetero-

zygosities (H_{obs}) were hardly ever very high. If there is a bias at all, the values are on the lower side. Comparing observed heterozygosities with those expected under HWE yielded low numbers of loci that significantly deviated. Further, only a handful of loci deviated from HWE in more than half of the localities under study. This may indicate that the forces acting in each of the different localities create a different set of markers to deviate from HWE in the respective localities. This can also be due to a lack of statistical power in the other localities, but out of the nine localities only three had a sample size of less than 20 individuals. Hence, this is less likely to account for the overall pattern. We also detected significant LD between only very few pairs of loci. This is not unexpected because based on their homologous positions in the chicken genome (cf. Kraus *et al.*[110]) each SNP should be physically far enough apart from its neighbours for recombination to break any incipient linkage very quickly. In conclusion, the SNPs in the "mallard 384 SNP set" are behaving neutrally and mostly unlinked.

3.2 MALLARD SNPS IN OTHER DUCK SPECIES

Significant increases of loci displaying missing data were between five and ten percent, except for the more distantly related *Aythya fuligula* where it was still less than a 20% increase. Additionally, many of these increases were rather moderate in their extent. This shows that despite long divergence times the flanking sequences are sufficiently conserved to allow the genotyping assay to perform well. This is an expected result since the "mallard 384 SNP set" was designed from SNPs that had a bias towards conserved flanking sequences. All SNPs with their flanking sequences can be mapped to the genome of the chicken, which is in a different order (Galliformes) than ducks. However, technical success and conserved flanking sequences do not yet constitute a good SNP set. The number of monomorphic (or nearly so, see above) SNP loci in each of the non-mallard duck species was substantial. With increasing phylogenetic distance (see below) fewer loci retained their polymorphisms.

3.3 DUCK SPECIATION: SUPERSPECIES AND THE SUPRA-POPULATION

Further elaboration of the species concepts used by us is necessary at this point. Species concepts have been dealt with in their own right, being listed and compared in most undergraduate evolutionary textbooks. Traditionally, a widespread one is the biological species concept[286], in which all individuals are deemed to belong to the same species if they produce viable and fertile offspring in nature, i.e., they share a common gene pool. To account for inherent difficulties to test this concept in practise, especially in populations that are geographically isolated and never encounter each other, biologists tend to supplement it by elements of the morphospecies concept. With the advance of molecular genetic data over the past decades many researchers define species by genetic characteristics rather than morphological ones because they provide a means of actually measuring recent or ongoing genetic connections between species. Since hybridisation leading to fertile offspring is rather common in ducks, even under natural conditions, we tested the species membership of our study objects with the genotypic cluster definition of Mallet[250] (Figure 5.S1 and Figure 5.3 from main text). Indeed, our PCAs showed clear genetic differentiation between mallard and other species. This indicates that despite tentatively high levels of hybridisation duck

species remain morphologically, ecologically and reproductively distinct. We also found genetic differentiation between the non-mallard species, although less pronounced. This is due to the fact that in the other species there were fewer informative SNPs. For instance, with about 100 polymorphic SNPs *Anas acuta* and *Anas crecca* formed non-overlapping clusters. The clusters of *Anas penelope* and *Anas strepera* were overlapping, probably because only around 50 polymorphic SNPs in these two species did not have enough power to delineate them as species, although for the outgroup *Aythya fuligula* only 11 polymorphic SNPs were enough to separate it in ducks of a different genus. Here, the larger evolutionary distance to *Anas spp.* plays a significant role.

The STRUCTURE[207] analysis identifies several cases where genetic admixture from other species seems supported by their genotypes. This may be direct evidence of partial genome sharing between species. When mallards and non-mallard species were analysed together, it was not possible to delineate species clusters. Nor was it possible to assign all mallards to the same genetic cluster, as indicated by some individuals forming their own distinct clusters. The STRUCTURE analysis is only suggestive because it relies on maximising HWE within clusters. Since we include mallards from several locations into the analysis, and the non-mallard species samples are composed of individuals from very different places, we cannot expect all individuals within one species to be in HWE. However, our parallel approach to do the same analysis with the program InStruct, which does not work with HWE, yields qualitatively similar results and thus strengthens our interpretation.

The species status of the currently recognised dabbling duck species is in line with morphological, ecological and reproductive observations. Additionally, we add genetic evidence from PCA clustering of genotypes to corroborate this taxonomic treatment. However, we also show that gene pools are likely connected between several of these species.

To solve the latter issue, we propose that the dabbling duck species studied here (and likely many others) should be regarded as a superspecies complex. The superspecies concept was put forward by Mayr in 1931[287], as a translation of *Artenkreis*, based on the work of Rensch[288]. Initially, it was used to assign species status to allopatric "races" which were too distinct to be lumped into the same species[289-291] (superspecies *sensu stricto*). Later, the definition was widened by Kiriakoff[292] and Mayr and Short[293] to be no longer exclusive to allopatric populations. For the mallard complex it has previously been proposed by Scherer[249]. Alternatives to the term superspecies (*sensu lato*) have been presented for sympatric cases, e.g., semispecies[294], megasubspecies[295] or supraspecies[296]. We dismiss the term 'semispecies' for our system for it semantically implies a negative flavour. We deem 'megasubspecies' a technically possible but confusing term. It indicates a group of taxa in transition to reaching species status. In our duck system, though, we believe that equilibrium has been reached already. Further, the prefix 'mega' designates 'quantity' instead of 'level' and should be avoided in nomenclatural discussions[296]. Lastly, the term 'supra-species' has indeed been proposed before the *International Code of Zoological Nomenclature*[296] but not received official status or frequent usage yet. Thus we choose to use the term superspecies (*sensu lato*) to embrace the sympatric distribution of duck species. This usage of the term is not uncommon in the literature. After all, that discussion is a semantic one. We do not attempt to scrutinise nomenclatural classification schemes, nor do we propose to change current nomenclature. The term superspecies is

clearly "an evolutionary taxonomy category but not nomenclatural rank"[296], thus to be preferred when studying biological systems rather than nomenclature.

The most important biological consequences of a shared gene pool among several species for the persistence times of SNPs is the enlarged population size. Even though all these duck species live in sympatry, such a combined population is highly structured by assortative mating. While geographical substructure would be indicated by the term "meta-population", the situation in ducks leads us to define a new term that is not loaded with geographical connotation. Thus, we think it is necessary to coin the term "supra-population". A supra-population is the group of individuals which are part of the same superspecies complex if it exists in sympatry and hybridisation occurs in nature.

4 Conclusions

Paleogeographic and paleoclimatic evidence over the last few million years strongly suggests that ecological conditions have been favourable for a duck speciation model in which an ancestral duck species radiated extremely successfully. The fossil record of ducks is still very poor[297] but the few studies on the subject suggest that morphological change in respective duck species has been very limited over the last few million years[267,298], after a larger waterfowl species turn-over 15-23 million years ago[299]. The first fossil that resembles a mallard is thought to be from the late Pliocene, about five million years ago[268]. This is close to the suggested lower bound of divergence times of some *Anas* species in the latest phylogeny of Anatidae[6]. An *Anas*-like duck must have split into multiple sister morphs within a short time and diverged by assortative mating, but still share large portions of their nuclear gene pools. The degree of SNP variation present in this ancestral *Anas* duck can well have been very similar to that observed today. Increased effective population size due to a supra-population within the superspecies around the mallard can preserve polymorphisms for many million years.

At present, extensive hybridisation still occurs. The genetic compatibility of different duck species, combined with mixed effects of genetically determined and imprinted mate choice leads to speciation reversals[300] despite genotypically and morphologically defined species boundaries. Present-day mallards may even drive some of their close relatives to extinction by hybridisation[115]. This is a major concern in many parts of the world, especially where mallards are not indigenous[230]. Many duck species of the genus *Anas* are hard to fit into the biological species concept under these circumstances. Their evolution has rather led to a superspecies complex with discernable lineages.

We also propose that the discussed analytical framework "*persistence time analysis*" is useful to study the evolutionary history of sister species pairs in general. Increased N_e due to supra-population formation within a superspecies complex can preserve polymorphisms much longer than would be anticipated on basis of the demography of each individual species.

Additional files

upon request from robert.kraus@senckenberg.de

Supplementary_File_1.pdf
A vector graph of a PCA analysis of genotypes of all mallard and non-mallard individuals. First and second axes are plotted against each other (explained variation in brackets). Grey dots represent individuals designated as mallard (*Anas platyrhynchos*) at sampling. Other colours indicate other duck species. A tentative hybrid between Anas acuta and *Anas platyrhynchos* is in red colour. Labels next to the dots represent individual study IDs. This file is scalable in order to retrieve details if needed. (Adobe Acrobat, 42 KB)

Supplementary_File_2.pdf
PCA analysis of genotypes of all non-mallard individuals. Details as in Supplementary_File_1, but without mallards. (Adobe Acrobat, 13 KB)

Supplementary_File_3.pdf
Bar graph of the genetic admixture analysis of individual ducks: all mallards and other duck species (see Supplementary Text, section 2.5). Each bar represents one individual and colours indicate membership to a certain cluster as identified with STRUCTURE without using prior information. Individual IDs are explained in the text and Supplementary_File_5.xls and Supplementary_File_6.xls. On the y-axis the percentage of membership to a certain cluster is given. For instance, individual ATHO001 (individual 1 from the mallard locality in Austria) is almost 100% assigned to the light blue cluster, while individual CARM009 (from a Canadian locality) is mainly assigned to the light blue, but also with about 15% to the purple cluster (an effect of genetic admixture between these two otherwise discrete clusters). This file is scalable in order to retrieve details if needed. (Adobe Acrobat, 24 KB)

Supplementary_File_4.pdf
Bar graph of the genetic admixture analysis of individuals: mallards excluded. See Supplementary_File_3.pdf for details. (Adobe Acrobat, 17 KB)

Supplementary_File_5.xls
A list of all mallard samples analysed in this study, accompanied by information on specific ID, collection date, country of origin, names of collectors, sampling locations, and further additional info. (MS Excel spreadsheet, 46 KB)

Supplementary_File_6.xls
A list of all other duck species samples analysed in this study, with details similar those given for mallards in Supplementary_File_5.xls. (MS Excel spreadsheet, 21 KB)

GS0011809-OPA.opa

'Oligo pooled assay (OPA)' summary file containing all necessary information for genotyping the presented SNPs on an Illumina BeadXpress system. (plain ASCII text, 230 KB)

mallard_cluster_file.EGT

'Cluster file' for use with GenomeStudio software. These configuration settings are used by the SNP genotyping software to convert raw signal into genotypes. (plain ASCII text, 170 KB)

CHAPTER 6

Avian Influenza surveillance: on the usability of FTA® cards to solve biosafety and transport issues

Robert HS Kraus, Pim van Hooft, Jonas Waldenström,
Neus Latorre-Margalef, Ronald C Ydenberg, Herbert HT Prins

Article published: Wildfowl. 2009. Special Issue 2: 215-223

Abstract

Many zoonotic diseases have birds as their natural hosts. Waterfowl are the natural hosts of avian influenza viruses (AIVs) and most avian influenza infections of wild birds appear mild, with infected inviduals displaying no or few symptoms. It is clear that the epidemiology of avian influenza cannot be fully understood without taking the ecology of its hosts into account. However, large scale studies and surveillance are still hampered by issues about preservation, transport and storage of AIVs, including bio-safety regulations and maintaining samples. This complicates the possibilities of the many small projects across the world if they are not done within the framework of one of the few big projects. Here, evidence is provided of the potential for using Whatman FTA® cards as a new preservation method to solve the above mentioned issues. Its efficiency is comparable to that of a standard method in virology, and saves time and money. In both large scale AIV sampling and small scale independent projects this method might be the means by which the field of the AIV ecology will be lifted beyond the constraints of difficult and expensive sampling, storage and laboratory facilities.

Introduction

Many zoonotic diseases have birds as their natural hosts[301,302]. For example, waterfowl are the natural hosts of avian influenza viruses (AIVs[23]) and there are a few reports of virulent epizootics in populations of wild birds and other wild animals[3]. AIVs are known to infect other hosts such as poultry, domestic livestock and humans[1,2] and may cause significant economic losses[303]. Highly virulent variants of AIVs have been recorded in many non-native hosts. The role of ducks and other wildfowl in the origin and spread of low and high pathogenic strains of avian influenza is debated[2,4].

Influenza viruses are a "genus" within the Orthomyxoviridae family of viruses. They have a segmented negative sense single-stranded RNA-genome. The influenza A virus can infect a wide variety of host species including birds, pigs, horses, seals, minks, whales and humans. The AIV genome consists of eight RNA segments. Viral subtypes are classified according to two of the encoded genes: the hemagglutinin (HA) gene and the neuraminidase (NA) gene[23]. These genes code for surface proteins that play a key role in host recognition and initial infection. Sixteen HA and nine NA "subtypes" are recognised, amounting to 144 ($= 9 \times 16$) possible subtype combinations[24]. These are described as, for instance, H5N1 (subtype '5' of HA, and subtype '1' of NA). Until recently, the classification used to rely on immunoassays using standard procedures[25-27]. Nowadays it is also possible to determine the nucleotide sequence of the virus genome using reverse transcriptase polymerase chain reaction (RT-PCR) with a set of universal primers for all genes and all subtypes[28]. The cDNA (complementary DNA) sequence obtained by this process can be identified with databases like GenBank[304].

Due to their tendency to feed in shallow waters and to congregate in large numbers, dabbling ducks are considered as one of the main vectors in avian influenza dispersal[23]. Moreover, some ducks may show no clinical signs when infected with AIVs[305], though recent studies report subtle influences of infection on the migration and feeding behaviour of swans[306] and mallard[307]. Therefore, as main vectors that survive most avian influenza infections, wild duck and many other

wildfowl would be a prime target for managers to monitor the potential spread of strains of highly pathogenic AIVs.

The importance of the ecological aspects of host biology, such as migration, and its consequences for the dispersal of AIV have led to the fusion of virology and ecology into many highly interesting projects (cf. refs [74,76,308-312]). Still, the ecological research to aid the understanding of the host-pathogen system "AIV and wild birds" has not been utilised to its full potential. The field of research is hampered by the fact that working with AIV may require biosafety precautions. Standard sampling and storage during avian influenza surveillance is bound to the availability of nearby deep freezers and transport of samples is subjected to strict regulations. Analysis can only take place in specialised laboratories. These facts make avian influenza research almost impossible if not conducted within the infrastructure of one of the few big collaborative projects. Hence, important contributions from the many smaller ecological projects may be missed[44,313].

Here, a possible solution for this problem is examined: a method to sample, store and analyse potential AIV containing samples. This method does not require immediate deep freezing. The issue of preserving RNA viruses for later analysis[314] has been addressed several times already in similar fields[315-323]. The so-called FTA cards® (Whatman®) are used to preserve AIV RNA on a dry storage basis. The chemicals in the FTA (Flinders Technology Associates) card render pathogens inactive upon contact[324] and transport can be arranged safely with only few further biosafety measures to be taken. FTA cards would therefore also be suitable for working with highly pathogenic strains of AIV. Proof of the potential of this principle is given in this short communication. The basis of this method is the isolation of the RNA followed by a one-step RT-PCR. The establishment of these protocols will be possible in any molecular laboratory, without the need for further biosafety measures. Samples can be mailed by normal postal services. Both sampling and analysis will be available to any molecular ecologist, thereby facilitating further scientific progress. This holds new possibilities for innovative studies in the fields of, for instance, molecular ecology, host-pathogen interactions or ecological immunology.

Methods

Wild mallards were caught in a duck trap at Ottenby Bird Observatory, Sweden (56°12'N 16°24'E), and cloacal samples were taken for AIV detection. Detailed information about trapping, sampling techniques and methodology are described by[325]. Of these, one avian influenza isolate subtype H5N2 from 2004 was tested for the usability of Whatman FTA® cards. A volume of 125 µl of the allantoic fluid of an infected embryonated chicken egg (equalling 48 HA units as measured by standard titration) were applied to an FTA card[326]. The dried sample on the FTA card was shipped at ambient temperatures for five days. Three 2 mm punches from this FTA card were incubated with RNA rapid extraction solution (Ambion) for 20 min at room temperature. RNA isolation was carried out with the MagMAX Viral RNA Isolation Kit (Ambion) according to the manufacturer's protocols. In short, RNA is captured by paramagnetic beads and washed in several steps to assure maximal purity, since biological samples from bird faeces would likely contain different PCR inhibitors.

The RNA was eluted into 50 µl elution buffer as provided by the kit. Three 2 mm punches from an untreated FTA card were carried along as negative extraction control; that is, to determine any contamination of the laboratory's tools or devices with AIV material. For RT-PCR detection we used the one-step Access RT-PCR System (Promega) – i.e., where reverse transcription into cDNA and PCR amplification is carried out in one tube – following a protocol adjusted from Fouchier *et al.*[327]. Stock solutions of 0.5 µl with 100 mM of the primers M52C and M253R[327] were used in reactions containing 10 µl AMV/*Tfl* 5x buffer, 1 µl dNTPs, 7 mM $MgSO_4$, 5U AMV reverse transcriptase and *Tfl* Polymerase each. A volume of 5 µl of isolated template RNA was added and the reaction volume adjusted to 50 µl with nuclease-free water. To exclude the possibility of AIV contamination and carry over in the RT-PCR kit chemicals, a negative control an additional sample with nuclease-free water as template was included. To test if the RT-PCR reaction works as expected, a positive control reaction is provided by the kit with its own primers.

RT-PCR commenced with an initial reverse transcription of 45 min at 45°C, followed by 2 min initial denaturation at 94°C and 40 cycles of: 94°C for 1 min, 45°C for 1 min, and 68°C for 2 min. An additional 7 min elongation at 68°C concluded the amplification. Amplicons (the amplified targeted fragments of the PCR reaction) were visualised on agarose gel stained with ethidium bromide and purified from it using the Millipore Montage DNA Gel Extraction Kit (Range 100–10,000 bp, Millipore Montage) as described in the kit manual. Cycle sequencing of the amplified Matrix gene fragment was carried out with ABI Big Dye 3.1 chemistry in 10 µl reactions containing 10–20 ng gel-purified template cDNA, 1.75 µl 5x dilution buffer, 0.5 µl Big Dye V3.1 premix, 1 µl forward primer (M52C, 10 mM), and ddH_2O. Cycling conditions were 1 min initial denaturation, followed by 25 cycles of: 10 s at 96°C, 5 s at 45°C and 4 min at 60°C. Samples were analysed on an ABI 3730 capillary sequencer. Wherever possible, preparation of reactions and handling of reagents was carried out under a fume hood and using RNAse-free barrier pipette tips. Care was taken to ensure that pre-PCR steps were carried out in a different room to the one in which the PCR and gel steps (post-PCR) were carried out, to avoid aerosol contamination of the laboratory.

In Wallensten *et al.*[325], 125 µl of the same isolate was used for direct RNA extraction and for determining avian influenza subtypes, using standard protocols. Extraction was carried out using the MagAttract Virus Minikit (Qiagen) on an M48 extraction robot (Qiagen). Virus detection was performed by RRT-PCR (real-time reverse transcription polymerase chain reaction) for the presence of the matrix gene[328], and the test proved to be positive[325].

Results

RT-PCR was successful for the positive sample extraction as well as for the positive RT-PCR kit control (Fig. 6.1), and amplicons were of the expected size (244 bp). Both negative extraction and negative RT-PCR control displayed no band on agarose gel, indicating that there were no contamination issues during the working procedures. A slight shadow below 100 bp in size indicates the possibility that primer dimer – an artefact of PCR wherein primers act as their own templates to make a small PCR product and appear faintly on an electrophoresis gel – might have been formed. The occurrence of multiple bands in the positive reaction control is described by the

user manual of the RT-PRC kit and is a normal sign of good amplification. The sequencing of the amplicon yielded a 95 bp good quality cDNA sequence (TCT TTA GCC ATT CCA TGA GAG CCT CGA GAT CTG TGT TTT TCC CTG CAA AGA CAT CTT CAA GTC TCT GCG CGA TCT CGG CTT TGA GGG GGC CTG AC). The chromatogram of the sequencer covered the whole fragment, however. A BLAST search against the National Centre for Biotechnology Information (NCBI) nucleotide database (blastn, Zhang *et al.* 2000; http://blast.ncbi.nlm.nih.gov/Blast.cgi) identified this fragment as being an AIV matrix gene fragment with 100% sequence identity and with 100% sequence coverage on comparison with several AIV isolates.

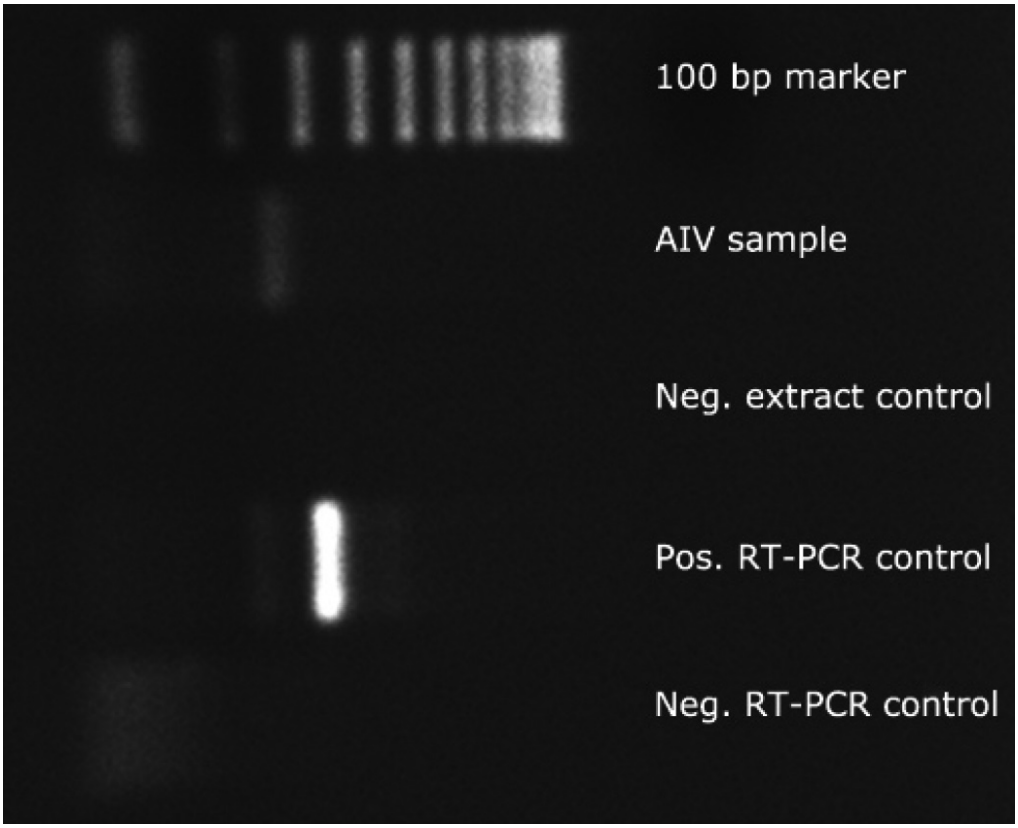


FIGURE 6.1: ***Agarose gel with RT-PCR products of the avian influenza virus (AIV) sample along with negative extraction control, and a positive and negative RT-PCR control (see main text).***
The AIV matrix gene (244 bp) amplicon appears between the 200–300 bp markers as expected.

Discussion

The study provides evidence that a new technology of viral RNA detection can be used to process samples containing AIV. The detection was successful without maintaining a cold chain for preserving the samples or involving unduly complicated biosafety measures regulations, and was

comparable to a standard and fully validated RRT-PCR[328]. In our RT-PCR experiment the slight shadow below 100 bp in size could indicate primer dimer. This might be due to the one-step nature of the RT-PCR protocol, where polymerase and primer are present in the reaction mix at lower temperatures already for some time during reverse transcription. This possible issue is easily solved by purification from agarose gel by excising only the relevant amplicon for subsequent sequencing. Sequencing of the 244 bp amplicon yielded 95 bp of high quality cDNA sequence. The chromatogram of the remaining fragment was too weak for analysis in this trial. Since only one test run has been conducted so far there is potential to develop this work further. The amplified fragment is also larger than one previously reported where alternative preservation methods were used (< 200 bp in ethanol[329]).

Since it has recently been shown that RNA fragments of > 700 bp in size can be amplified successfully in other systems[323] we assume that storage of avian influenza samples on FTA cards have the potential to be superior to the ethanol fixation method if primers for larger fragments are used. Some studies tested the sensitivity (*e.g.*, RNA quantity) required for detection. They reported the detection of a positive signal only after many-fold dilutions[318] or for only 0.1 fg of RNA template[330], and after storage at ambient temperatures for > 2 weeks. Others claim that RNA on FTA cards is stable even after six months of storage under ambient conditions[330]. Whether these methods are applicable under fieldwork conditions remains to be tested. In natural samples like faeces or oral/cloacal swabs there is also the chance that AIV is present in lower concentrations than tested here. This poses the risk of not detecting an avian influenza infection when there actually is one (*i.e.*, a false negative). In particular the effects of storage time and temperature, as well as sensitivity at lower concentrations and contamination through faecal material, would need attention in such a systematic test. Recent studies have however shown that PCR is more sensitive than traditional methods, even when AIVs are only present as unviable particles[331]. This also makes detection possible when infection is almost cleared by immune response. To this point, cloacal samples were not tested directly in the present study but it seems that the use of FTA cards in large scale AIV sampling may be the means by which the field of AIV ecology can be lifted beyond the constraints of difficult sampling, storage and laboratory facilities.

Acknowledgements

Technical assistance was provided by Bert Dibbits and Haisheng Nie, and we thank Jan van der Poel for helpful discussions on preparing for the experiment. Further we would like to thank the Animal Breeding and Genomics Group, Wageningen University, Wageningen, The Netherlands, for hosting us in their laboratory and the Ottenby Bird Observatory, Sweden, for hosting collaborators and providing samples. Financial support was given by the KNJV (Dutch hunters association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik Trusts (both in The Netherlands), the Swedish Research Council (grant no. 2007-20774) and the EC-funded Newflubird project. This is contribution No. 230 from the Ottenby Bird Observatory.

CHAPTER 7

Avian Influenza surveillance with FTA® cards: Field methods, biosafety, and transportation issues solved

NB: This is a video article, for full experience watch:
http://www.jove.com/details.php?id=2832

Robert HS Kraus, Pim van Hooft, Jonas Waldenström,
Neus Latorre-Margalef, Ronald C Ydenberg, Herbert HT Prins

Article published: Journal Of Visualized Experiments. 2011. 54:e2832

Abstract

Avian Influenza Viruses (AIVs) infect many mammals, including humans[332]. These AIVs are diverse in their natural hosts, harboring almost all possible viral subtypes[44]. Human pandemics of flu originally stem from AIVs[333]. Many fatal human cases during the H5N1 outbreaks in recent years were reported. Lately, a new AIV related strain swept through the human population, causing the 'swine flu epidemic'[32]. Although human trading and transportation activity seems to be responsible for the spread of highly pathogenic strains[42], dispersal can also partly be attributed to wild birds[43,139]. However, the actual reservoir of all AIV strains is wild birds.

In reaction to this and in face of severe commercial losses in the poultry industry, large surveillance programs have been implemented globally to collect information on the ecology of AIVs, and to install early warning systems to detect certain highly pathogenic strains[325,334-337]. Traditional virological methods require viruses to be intact and cultivated before analysis. This necessitates strict cold chains with deep freezers and heavy biosafety procedures to be in place during transport. Long-term surveillance is therefore usually restricted to a few field stations close to well equipped laboratories. Remote areas cannot be sampled unless logistically cumbersome procedures are implemented. These problems have been recognised[307,314] and the use of alternative storage and transport strategies investigated (alcohols or guanidine)[331,338,339]. Recently, Kraus *et al.*[77] introduced a method to collect, store and transport AIV samples, based on a special filter paper. FTA cards[67] preserve RNA on a dry storage basis[330] and render pathogens inactive upon contact[324]. This study showed that FTA cards can be used to detect AIV RNA in reverse-transcription PCR and that the resulting cDNA could be sequenced and virus genes and determined.

In the study of Kraus *et al.*[77] a laboratory isolate of AIV was used, and samples were handled individually. In the extension presented here, faecal samples from wild birds from the duck trap at the Ottenby Bird Observatory (SE Sweden) were tested directly to illustrate the usefulness of the methods under field conditions. Catching of ducks and sample collection by cloacal swabs is demonstrated. The current protocol includes up-scaling of the work flow from single tube handling to a 96-well design. Although less sensitive than the traditional methods, the method of FTA cards provides an excellent supplement to large surveillance schemes. It allows collection and analysis of samples from anywhere in the world, without the need to maintaining a cool chain or safety regulations with respect to shipping of hazardous reagents, such as alcohol or guanidine.

Protocol

I. DUCK TRAPPING AND CLOACAL SWABBING

1. Trap dabbling ducks of the genus *Anas* (or other birds) in a cage by attracting them with lure ducks or food, and put them into individual cardboard boxes. Transportation time in the box should be kept to a minimum, for instance, by installing a field station within a short distance to the trap. Logistic circumstances may vary in each study. The advice and approval of the relevant animal ethics committee needs to be sought for each new set-up. Alternatively, also hunter shot birds can be sampled.
2. Take the duck out of the box by holding its wings tight to its body. Sample a fresh dropping, which will be present in most of the cases, directly from the bottom of the box. Pick

them up with a sterile swab and apply the fluid to a Whatman FTA card. If not, perform cloacal swabbing.

3. Turn the duck on its back. This permits access to the cloaca and facilitates sampling. The cloaca is a protruding structure situated below the abdomen, close to the base of the tail. An experienced person could hold and sample the bird at the same time, or else a second person can assist.
4. Carefully insert the sterile plastic rayon swab (Copan, Italy) approximately 1 cm and make a gentle swirl of the cloaca. Roll the swab with the fluids from the cloaca over the surface of a Whatman FTA card. After sampling has been performed immediate release of the animal is desirable.
5. Dry the FTA cards at room temperature for at least 1 hour, then store each sample individually in paper bags (e.g. envelopes). Storage is possible at room temperature, possibly with silica beads in humid climates. The sending and receiving institutions' biosafety officers can permit to send FTA cards by regular mail, since the pathogens become inactive upon contact with the FTA card surface.

II. VIRAL RNA ISOLATION

1. Extract the FTA card material including faeces/cloacal fluid. Punch three discs with a Harris 2 mm puncher and place all of those into a well of an RNase free 96well plate. Between each new sample, clean the puncher carefully with alcohol and a Kim precision wipe (Kimtech). Always use positive and negative controls. A positive control can consist of a laboratory strain freshly applied to an FTA card, or a natural sample which is know to yield a positive result. As negative control, punch out discs from an empty FTA card. Additionally, leave another well free for a PCR control later in your experiment (with water as template).
2. Add 70µl RNA rapid extraction solution (Ambion) and heat-seal the plate (AbGene Thermo-Sealer). Incubate 5 minutes on a plate shaker at room temperature with the determined speed that does not cause spill-over (this depends on the used plate shaker model; test in advance).
3. Carry viral RNA isolation according to the manufacturer's protocols with the MagMAX-96 viral RNA isolation kit (Ambion). In brief: Add 130µl prepared lysis/binding solution (from the kit) to each well. Transfer 50µl extracted FTA card material to each well. Shake plate for 1 minute.
4. Add 20µl prepared bead mix (from the kit) to each sample and mix by pipetting up and down. Shake on determined speed for 5 minutes. RNA molecules will bind to the magnetic beads.
5. Move the plate to a magnetic 96-well stand to capture the beads. Depending on the model of magnetic stand used, this can take more than 5 minutes. When the solution has turned completely clear, remove and discard the supernatant. Then remove the plate from the magnetic stand.
6. Wash the beads twice with wash solution 1 and twice with wash solution 2 (from the

kit). For each of the 4 wash steps add 150µl prepared wash solution to each sample, shake for 1 minute on determined speed, move the plate to the magnetic stand, capture beads for about 5 minutes (or until the solution is completely clear), remove and discard the supernatant. Then remove the plate from the magnetic stand, add the next wash solution, and carry out the washing steps as previously. After the last washing step, remove as much wash solution as possible and air dry the beads at room temperature in the plate shaker for 2 minutes.

7. Add 50µl of elution buffer (from the kit) to release RNA from the beads and shake for 4 minutes on the plate shaker. Capture beads as previously described. The supernatant now contains the isolated RNA which is ready for downstream applications.

III. RT-PCR OF THE AIV MATRIX GENE

Carry out reverse transcription PCR (RT-PCR) with the One-Step Access RT-PCR system (Promega) in 25 µl reactions (adjusted from Kraus *et al.*[77] and Fouchier *et al.*[327]):

Nuclease free water	1.5µl
AMV/Tfl5 buffer	5.0µl
dNTPs	0.5µl
primer M52C[327] (10 µM)	2.5µl
primer M253R[327] (10 µM)	2.5µl
MgSO4 (25 mM)	7.0µl
AMV reverse transcriptase (5u/µl)	0.5µl
Tfl polymerase (5u/µl)	0.5µl
RNA sample	5.0µl

PCR conditions in a Biometra T1 thermocycler are: initial reverse transcription of 45 minutes at 45°C, followed by 2 minutes initial denaturation at 94°C and 40 cycles of: 94°C for 1 minute, 56°C for 1 minute, and 68°C for 2 minutes. An additional 7 minutes elongation at 68°C concludes the amplification.

IV. SCREENING FOR AI-POSITIVE SAMPLES AND PURIFICATION OF TARGETED FRAGMENTS FROM GEL

1. Load 2µl of the PCR product mixed with 2µl 5x loading dye (BioRad) and 6µl water on a 1% agarose gel (Roche) stained with 1% ethidium bromide (2.5µl per 100 ml gel) for a pre-screening.
2. Run the gel for 1 hour at 120V, along with a DNA size standard (BioRad EZ load 100 bp ladder), visualise with a gel documentation system. See an example in Figure 7.1.
3. Select samples with amplified fragments in the expected size range (between 200 bp and 300 bp (target fragment is 244 bp). Load the whole PCR reaction volume (of which ~23µl are left) of these candidates with 6µl 5x loading dye on a 2% ethidium bromide stained agarose gel and run for 2 hours.

4. Place the gel on a UV-transilluminator (Bioblock Scientific) and inspected visually. See an example in Figure 7.2. Excise bands of the correct size from gel with a scalpel and placed into individual 1.5 ml reaction tubes. Purify fragments from gel, for instance with the Zymoclean Gel DNA Recovery Kit (Zymo Research).

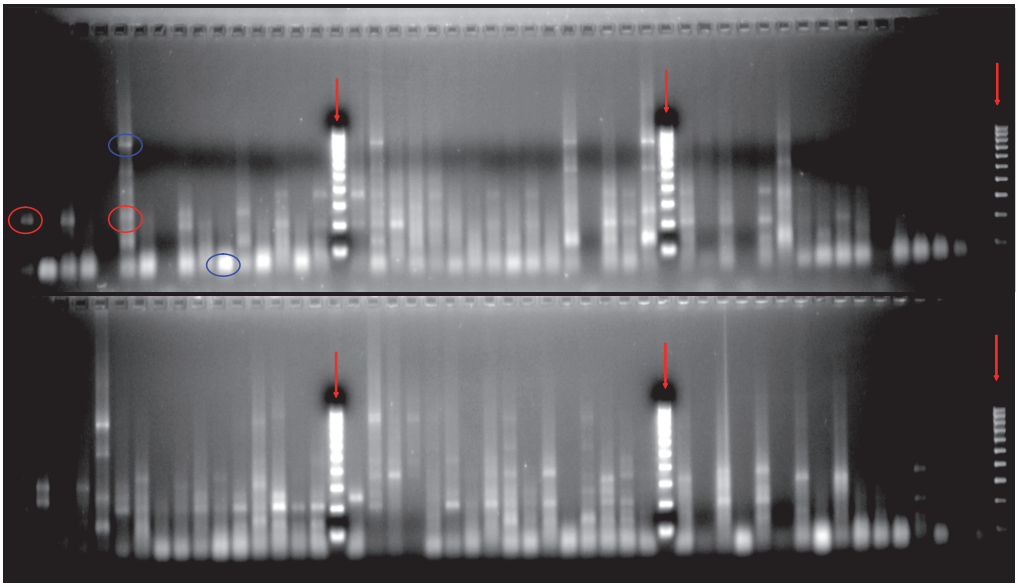


FIGURE 7.1. ***Gel picture of a preliminary screening of the PCR products.***
2 µl PCR product of positive control, serial dilutions of the positive control, two negative controls and 84 cloacal samples were loaded. 48 samples are shown in the top gel panel, and 48 samples in the bottom panel. Red arrows indicate gel lanes used for 100 bp DNA size standard. For illustration, red circles show bands in the expected size range. Blue circles indicate an unspecific amplicon (top left) or a primer dimer artefact below 100 bp in size (bottom right).

V. SEQUENCING AND IDENTIFICATION OF PCR PRODUCTS

1. Carry out Sanger sequencing of the target fragments, for instance on an ABI 3730 capillary sequencer with ABI Big Dye 3.1 chemistry (Applied Biosystems). Prepare sequencing reactions in 10µl volumes containing 10-20 ng gel-purified template cDNA, 1.75µl 5x dilution buffer, 0.5µl Big Dye V3.1 premix, 1 µl forward primer (M52C20, 10 mM), and ddH2O. Cycling conditions are: 1 min initial denaturation, followed by 25 cycles of: 10 s at 96°C, 5 s at 45°C and 4 min at 60°C. Purify and prepare the sequencing reaction according to your internal protocols.
2. Identify resulting cDNA sequences against nucleotide databases such as GenBank[304], e.g. by web based tools such as BLAST at the National Centre for Biotechnology Information (NCBI, http://blast.ncbi.nlm.nih.gov/Blast.cgi).

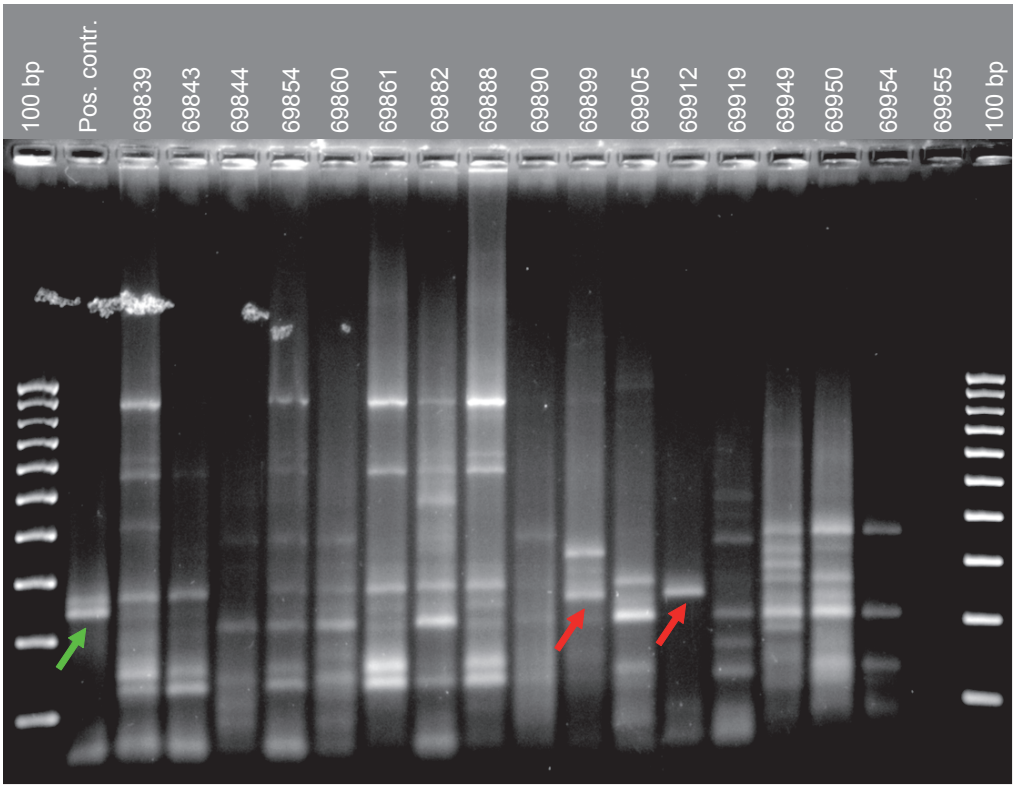


FIGURE 7.2. ***Selection of samples with fragments in the correct size range.***
Samples with bands between 200 bp and 300 bp (target fragment 244 bp) were chosen. The green arrow indicates the positive control, the two red arrows indicate samples which were confirmed to be AIV positive by comparing their cDNA sequences to the NCBI nucleotide database.

Representative Results

Mallards (*Anas platyrhynchos*) were sampled at Ottenby Bird Observatory in December 2007. From each mallard a sample on FTA card was taken as described in this protocol. After shipping, the FTA cards were kept in a freezer at -20°C for two years. The same FTA card sample of the laboratory isolate tested in Kraus *et al.*[77] was included as positive control, as well as nine tenfold serial dilutions of it. Two negative controls were i) extraction from an empty FTA card, to test if there was carry-over from the puncher, and ii) RT-PCR reaction in which nuclease free water was used as template, to test if contamination occurred during, or in preparation of the PCR reaction.

84 samples were analysed. A gel picture of the PCR products from these 84 samples can be found in Figure 7.1. From natural samples a multitude of unspecific bands can be observed due to the presence of various microbial contaminations in the faeces. However, the target fragment

of the primer pair is 244 bp long. The whole PCR-reaction volume of a subset of the samples which produced fragments in approximately the correct size range (between 200 bp and 300 bp) was loaded on gel (Figure 7.1). An illustration of which of the bands were cut from the gel can be found in Supplementary Figure 7.1. In addition to the positive control, two of these samples (69899 and 69912) were positive by the new protocol. A BLAST search against the NCBI nucleotide database revealed their identity as AI matrix gene (Figure 7.3), while all the other bands resembled bacterial sequences most closely, or did not yield a readable sequence at all.

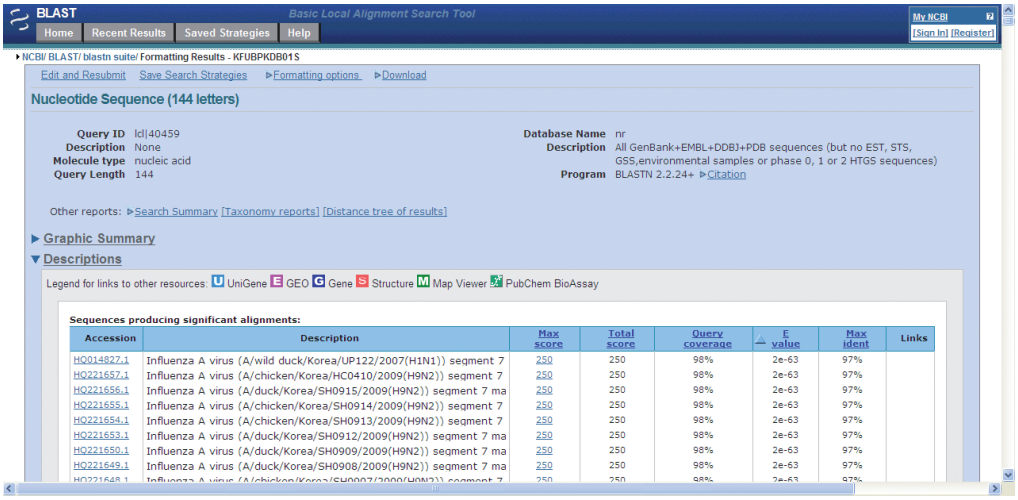


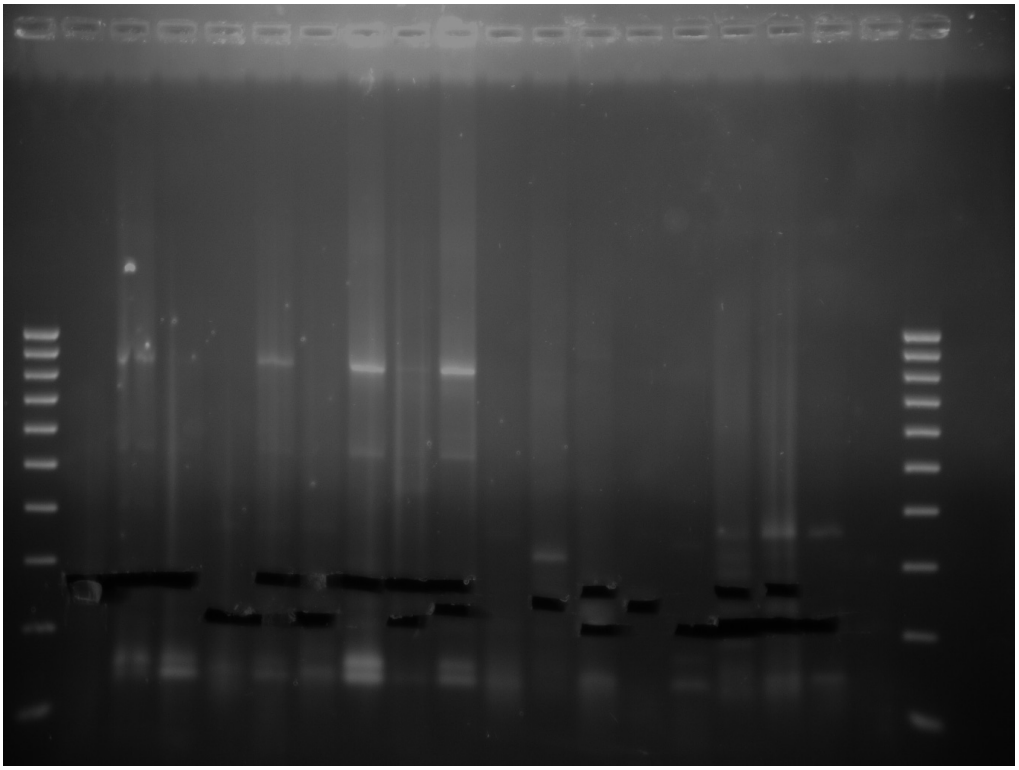
FIGURE 7.3. *Screen capture of a representative BLAST search at NCBI.*
One of the cDNA sequences obtained from the excised fragments was queried against the nucleotide database at the NCBI website. The sample is correctly identified as AI Matrix gene fragment.

Discussion

The protocol described here provides a supplementary method to screen faecal or similar samples for the presence of AIV. It was especially designed to make sample collection quick and easy. This makes it possible for less trained persons, such as hunters or wildlife managers, to contribute to AI surveillance. No cool chains need to be applied, although freezing the samples is recommended where possible. A few days of room temperature, for instance during transport to the laboratory, were not a problem for RNA molecules on FTA cards, as long as the cards remained dry. Other non-cooled storage systems that are currently evaluated by the research community, such as alcohol[331] or guanidine[338], require special shipping arrangements because of their hazardous nature. In contrast, the FTA card method does not require shipping of hazardous materials. However, another interesting storage medium in this respect is RNAlater™ that is not hazardous, either[339]. Sample analysis can be carried out in any standard molecular laboratory. No special equipment other than for usual PCR reactions and capillary sequencing was needed. All steps could be carried

out without a biosafety level because already at sample collection the potential pathogenic agents were inactivated by the antibacterial and antiviral activity of the FTA card.

Cross-contamination and subsequent false positive samples were not observed in our trials with wild bird samples. However, when working with RNA and PCR it is always advisable to pay special attention to clean working places and separate rooms for pre- and post-PCR steps. Working in a fume hood decreases the risk of aerosol contamination in the laboratory. Pipetting needs to be carried out with filter tips.



SUPPLEMENTARY FIGURE 7.S1. ***Bands of selected samples excised from gel.***
This picture was taken from the gel depicted in Figure 7.1 after candidate bands between 200 bp and 300 bp were cut out.

From a previous study on samples taken simultaneously from the same ducks we know that six of the 84 samples were positive by the traditional RealTime RT-PCR method[328] and the Ct values from RealTime RT-PCR are known. The two positive samples detected by our method stem from ducks which had Ct values <30 (indicating a high concentration of viral RNA). The other four samples which were positive with the traditional protocol had Ct values >30 (less concentrated) and could not be detected by our protocol.

Only samples with relatively high virus titers were positive in our assay and it is likely that sensitivity of the method was the source of failure to detect all positive samples. Further, the sample size in the current study was very low and the method can be completely developed and assessed if more controlled and rigorous experiments are carried out. However, if these samples would have been collected from a remote area, traditional analysis would not have been possible at all. Additionally, these first results stem from pilot experiments which need further optimisation. A period of two years storage in a regular -20°C freezer after sample collection probably also affected the quality of the viral RNA. This possible RNA degradation is an important issue when dealing with room temperature storage of inactivated viruses. Samples stored in alternative liquids as mentioned above suffer from significant degradation which impacts analysis of longer stretches of the viral genome[338]. Although not tested in our study, FTA cards have proven to be well suited to preserve intact RNA molecules in other RNA systems that are very similar to Avian Influenza viruses[318,323].

Acknowledgements

We thank Bert Dibbits for technical assistance. The Animal Breeding and Genomics Group, Wageningen University, The Netherlands, generously hosted us in their laboratory. The personnel of the Ottenby Bird Observatory, Sweden, is thanked for trapping and sampling the mallards, in particular Magnus Hellström, Marcus Danielsson, Christopher Magnusson and Stina Andersson. We thank Sanne Svensson, Jonatan Qvist and Per-Axel Gjöres for filming at Ottenby, and Mano Camon for filming in the lab. Daniel Bengtson provided beautiful duck photographs for the video portion of this publication. Further free material from the CDC Public Health Image Library (PHIL; http://phil.cdc.gov/phil/) was used: The electron micrograph (no. 280; by Dr. Erskine Palmer) and illustration (no. 11823; by Douglas Jordan) of the influenza virus. Financial support was given by the KNJV (Royal Netherlands Hunters Association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik Trusts (both in The Netherlands), the Swedish Research Council (grant no. 2007-20774) and the EC-founded Newflubird project. RNA isolation chemicals were a generous gift of Ambion, Inc, the RNA company. This is contribution No. 245 from the Ottenby Bird Observatory.

CHAPTER 8

Synthesis

The Genetic Tool Box for Ducks

Ducks play a significant role in science and are recognised study objects for research on evolution[340], speciation[341] and hybridisation[248], sexual selection[9], immunology[112,342,343], especially with respect to Avian Influenza[342,344], and genomics[157,345]. Further, after chicken and turkey, duck meat and eggs are the third most important food on the world's poultry market[144]. However, after some initial studies on their molecular ecology and conservation genetics[79,92,223,346-349], little attention has been allocated to molecular studies during the last decade. A phylogenetic study has been carried out recently[6], refining earlier such studies in waterfowl[5,350]. In North-America just a few molecular ecological or evolutionary genetic studies have employed relatively recent methods, such as demographic modelling and the joint analysis of nuclear and mitochondrial DNA (mtDNA) sequence data[88,94,100,116,133,134]. Similar studies were carried out in Asia[63,82,351,352]. Multi-locus methods based on microsatellites, for instance, were mainly focused on smaller spatial scales[98,353].

In my thesis I specifically expand much-needed mtDNA sequences of mallards of geographically diverse origin, by more than doubling the number of control region sequences from the mitochondrial genome (Chapter 2). With the sequence data produced during this thesis I was for the first time able to analyse the population genetic structure of the mallard throughout its entire native range. These new sequences cover Europe and also large parts of North-America, from which mtDNA sequences were previously rarely sampled. Further, I present the first genetic data from Greenland mallards, which were thought to be an isolated mallard population, maybe even constituting a subspecies[12] (I do not attempt to argue about a subspecies status but in Chapter 4 I confirm the Greenland mallards to be isolated). This backbone of sequence information is publicly available through GenBank[29]. Future studies can also benefit from my experience that the primer sequences for PCR reactions can be used in any mallard population without the need for redesign. Adding further individuals to increase local sample sizes or including regions that are currently not sampled for mtDNA sequences should be straightforward based on the methods presented and evaluated in Chapter 2.

Mitochondria are cell organelles situated outside the nucleus. As such, they are usually only passed to offspring by the mother because the egg cell harbours all the cytoplasm. Although rare instances of paternal mtDNA transmission have been observed in invertebrates[354,355] and vertebrates[356,357] mtDNA is generally assumed to be a genetic marker that reflects the maternal lineage exclusively because the ratio of maternal and paternal mitochondria in the zygote is heavily skewed towards the maternal lineage, and hence the paternal mitochondria are usually diluted out[358]. As there are other issues, too, with inferences based on mtDNA (e.g., it is only a single locus; it has comparatively low effective population sizes; the assumption of selective neutrality is not necessarily valid; and other issues reviewed in[359]), I set out to develop more nuclear genotyping resources. In previous molecular ecological or evolutionary studies of duck species, one or just a small number of nuclear sequences were sequenced to overcome the limitations of mtDNA[100,133,134]. This is tedious to scale up and offers little improvement in terms of coverage of a representative genomic history over multiple loci. A moderate increase of genome coverage is possible by genotyping microsatellite DNA. Some microsatellites have been described for mallards initially for molecular ecological purposes[360,361], and more recently this set has been extended for performing

genomic linkage mapping of Chinese domestic duck breeds in order to facilitate food production for human consumption[345,362]. However, as described more extensively in the General Introduction and in the introduction to Chapter 4, microsatellites suffer from a number of shortcomings (see ref [186] for a good overview).

Recent advances in DNA sequencing technology, the so-called 'next generation sequencing technologies'[194], made it possible to obtain nuclear sequences from the whole genome of multiple individuals. Technologies to reduce the complexity, i.e., the amount of information produced during a next generation sequencing experiment[149,155], to an affordable and manageable subset, greatly advanced the access of non-model organism studies to such genome-wide analyses[150,151]. As one of the early studies employing next generation sequencing to non-model organisms, I sequenced ~5% of the mallard duck genome from a diverse pool of European mallards in Chapter 3. I further became a member of the Duck Genome Consortium (China Agricultural University and Beijing Genomics Institute), and used the duck genome draft sequence (Huang *et al.*, unpublished data) to detect SNPs within the European mallard samples. More than 100,000 SNPs, highly likely to represent true segregating polymorphisms in wild mallards, were found in just ~5% of the duck genome. A set of 384 of these SNPs were subjected to extensive evaluation in Chapters 3 and 5. The advantage of this SNP set over existing microsatellites is not only in its statistical power, which is believed to be in the order of 100 microsatellites, varying with the specific application of interest[62,142]. It moreover provides common grounds for future projects to join their genotyping results with those obtained in my study. SNPs, in contrast to microsatellites, can be consistently genotyped across several subsequent studies in the same laboratory, across different laboratories, and also across different genotyping technologies. I envisage that future studies employing molecular markers for any type of mallard project perform genotyping with the SNP set presented in this thesis and incorporate the genotypes that I generated into their own studies. This will ultimately unleash the potential of multiple cooperating research groups independent of time, space and genotyping technology used.

Mallards and the spread of Avian Influenza

Recently, seasonal migration of birds has come into the focus of the public due to outbreaks of Avian Influenza (AI)[40]. Dense network structures[363] are thought to enhance the spread of diseases, especially those in multi-host systems[364]. It is clear that one needs to understand population structuring, and hence biological host networks, as such aspects of a host's ecology are key to understanding the spread of diseases such as AI[2].

Avian Influenza viruses (AIVs) are transmitted via the faecal-oral route, especially among birds that live and feed on water. AIVs have been shown to remain infectious in open water for several days, depending on environmental circumstances[234-236]. Therefore, if water birds congregate into single or multispecies flocks on water bodies, there is always potential for the spreading of AI. When using genetic methods to infer population structure it is obviously not possible to detect the mere meeting of ducks on a lake, however, the patterns of mating will be recognisable. While meeting is not necessarily mating, the opposite is definitely true: mating implies meeting. Therefore, if I detect strong genetic connectivity due to matings between populations I can also strongly

argue for a possible route of AI through this path. Detecting real time migrants by genetic clustering and subsequent assignment tests tells the tale. Moreover, the genetic traces left by mating can be studied backwards in time, helping to understand changes in this network over time.

Gene Pool Connectivity

The study of the genetic connectivity of populations has a long tradition in molecular ecology[173]. These patterns have important bearings on, for instance, conservation and management. The amount of gene flow between an endangered population and a larger source population may determine the strength of the genetic effect of a possible bottleneck, and hence the risk of loss of genetic diversity, or even the detrimental impact of inbreeding in small populations. Nature conservation efforts world-wide benefit from including a genetic agenda in order to retrieve more detailed answers to key questions, and spend time and money wisely.

An important outcome of the study of genetic connectivity between populations is that researchers will have direct evidence of movement of individuals between populations, which usually means between geographic regions. Inferring gene flow means inferring migration of individuals between gene pools (at least in mobile animals; in sessile organisms such as corals or plants, only zygotes or seeds are passively dispersed by ocean or wind currents and hence adult individuals do not migrate). With migration I refer to the movement of individuals between populations, in contrast to the phenomenon of seasonal migration, in which the population as a whole moves between regions.

Building on my molecular toolbox (see above) I conducted two studies to investigate gene pool connectivity among the world's mallard populations. For the mallard, clear geographical boundaries between populations have been difficult to delineate based, for instance, on bird ringing activities[12], and the largest study about genetic population structure published independent of this thesis suggested very little differentiation among mallards dispersed across many thousands of kilometres in Siberia[63]. Based on these observations I hypothesised that for mtDNA I would expect large scale panmixia within which little structure is present. Indeed, I found mallards within landmasses to be almost unstructured in their mtDNA (Chapter 2). However, between Eurasia and North-America I observed considerable genetic distance. This indicates that oceans form effective barriers to movements, at least for the female portion of the mallard population. However, when investigating the pattern for both sexes based on the SNP set developed in Chapter 3, it became apparent that males probably do cross the oceans often enough to obscure emerging population structure due to the lower mobility of females (Chapter 4). Dispersal of genes in a mallard population network is therefore expected to occur very fast. Close contact, as is necessary for mating and dispersing one's genes, results in structures that greatly facilitate the spread of AI in a small-world networks[365].

Interestingly, in both studies I observed signs of subtle genetic structure within Europe. The peculiar signal in the data for mtDNA (Chapter 2) was a strong unevenness of the haplotype frequency spectrum as identified by Tajima's D[83] and Fu's FS[84] (most pronounced in Europe, although present on other continents as well). This signal either points towards natural selection against deleterious mutations, or towards recent population growth. I also found evidence that

some mallards in Europe form a separate genetic cluster from the rest of the world (Chapter 4). These two results from two independent genetic marker sets distinguish sub-populations of European mallards from other populations of the species worldwide. A scenario of extensive population growth is unrealistic, as I would not expect such exceptional growth in Europe alone. Data from demographic modelling of the coalescent process of mtDNA in Chapter 2 indicates large population growth for both Eurasia and North-America, not only for Europe. Natural selection and subsequent local adaptation might be more likely to contribute to the pattern found in Europe. There, especially, resident populations of mallards exist, staying in a region throughout the year[12]. This is facilitated by comparatively mild winters and a dense human population, leaving sufficient food in its surroundings. Adaptation to a resident life style may have lead to the observed genetic patterns. A way to study this hypothesis would be to specifically sample and genotype a mallard population that is known to be resident and a population that is known to be migratory in an area where these intermix during winter. Large samples per population (>50 individuals), larger than taken in this thesis, would allow genetic differentiation between these populations to be evaluated.

Another possibility to explain the observed distortion of the mtDNA frequency spectrum of the mtDNA sequences as well as the detected population structure in SNP genotypes in Europe might be the extensive release of farmed mallards for hunting purposes (probably not all released mallards are always shot). During the breeding process severe changes in allele frequencies would act on the farm population due to genetic drift in small populations coupled with artificial selection for farming conditions. For instance, it was found that just 30 years of mallard farming have changed bill morphology substantially[96]. It would be interesting to genotype mallards from such farms to potentially identify a source of the genetic alteration in the European mallards. Additionally, museum specimens that were collected before mallard farming for hunting purposes became wide-spread in Europe, would be a valuable source of 'uncontaminated' mallard DNA.

Duck Evolution and Reduced Susceptibility to Avian Influenza

Influenza viruses have caused severe pandemics in humans, with records dating back to the 18th century[366]. However, it seems that currently recognised strains of AIVs that cause disease in humans all date back to mutational alterations of AIV subtypes ('Spanish Flu', 1918), followed by early reassortment events between indigenous avian-adapted Influenza viruses with human-adapted Influenza A viruses (epidemics in 1957 and 1968)[31,367]. Besides the potentially deadly events of AI outbreaks in the human population, AIVs are a part of the ecosystem for at least centuries. Waterfowl, the major natural reservoir of a vast diversity of subtypes of AIVs, and AIVs themselves are co-adapted. Antigenic drift in AIVs, i.e., mutational processes leading to new antigenic subtypes, is in equilibrium with the immune system of their hosts[23]. Outbreaks of highly pathogenic AI in poultry, for instance, can be seen as the outcome of natural selection on the virus for high virulence, adapting to conditions in dense networks such as commercial poultry farms[44].

The long evolutionary association of water birds and AIVs[44,368] is manifested in the diversity of AIV subtypes continuously circulating in these birds[311]. Ducks in particular harbour the most diverse pool of AIVs[369], although there are differences of subtype frequencies between waterfowl

and shore birds/gulls[370], as some subtypes occur in gulls much more frequently. Studies on mallard antibodies show that some parts of the innate immune system of this waterfowl species is less efficient during AI infection, when compared to other animals, which might facilitate continuous shedding of the virus for several days or weeks[343], during which mallards remain largely physically unaffected[305,371], even when challenged with viral strains highly pathogenic in poultry or humans[34-36] (but see [37-39] for examples of strains that are fatal for mallards, too).

In addition to their unusual innate immune response to AIVs, also the mallard's demography is special. According to the BirdLife species fact sheets, the world-wide mallard population is estimated to be 19 million individuals[278]. According to the rule of thump found by Frankham[279] the effective genetic population size is about 10% of the census size, hence 1.9 million mallards would be estimated to be the genetically effective population currently inhabiting the world (ignoring the issue of subdivided populations[284] because I find that mallards largely form a global population; Chapter 4). The large census size of living mallards will facilitate the spreading of AIVs due to high host densities. However, more important on evolutionary time scales, at which the host-pathogen arms race takes place, is the ability of large populations to maintain genetic variation. The larger the effective population size is, the slower will be loss of neutral genetic variation due to stochastic processes, i.e., genetic drift. However, the loss of non-neutral variation, for instance, due to purifying selection, is also slowed down by large population sizes[372]. Hence, slightly deleterious mutations, such as those benefiting AIVs instead of the mallard, remain in the population and facilitate the survival of the virus. From demographic modelling in Chapter 2 I know that the rule-of-thumb estimate of roughly 2 million effective mallards is conservative, as my calculated value is between 2.8 and 7.1 million (Old World and New World estimate combined, see Chapter 2). This is a high number, and to my knowledge exceptional among terrestrial birds and mammals. Some mouse species which would be candidates for very high populations sizes, reach up to a few hundred thousand effective individuals[373]. Some groups of small passerine birds, such as sparrow species or the red-billed quelea[374], which is the most numerous bird in the world[375], have census population sizes in the hundred millions, or even >1 billion. However, to my knowledge, the effective population size has not been studied in these species, and could be much lower due to geographical genetic substructure. At least for the quelea, high connectivity has been shown on a somewhat smaller scale[376], though.

Another important aspect is the finding of inter-species gene pool connectivity among dabbling ducks (Chapter 5). In this thesis I propose that extensive ongoing hybridisation between several duck species contributes to a further increased effective population size. This potentially retains even more 'pre-adaptations' in the joint gene pool of a 'supra-population' of mallards and allies. Ducks are therefore well-equipped for the host-pathogen arms race with a virus. High potential on the side of the virus for co-adaptation seems to lead to the relatively stable relationship between ducks and the many subtypes of AIVs[377].

Molecular Ecology and Virology

Molecular ecology is a booming discipline. It benefits from accelerating technological advances in molecular genetics (e.g., advances in molecular markers and sequencing technology[62,194,378,379]),

theoretical population genetics (e.g., the coalescent[380]) and, recently, bioinformatics[381,382] and the associated hardware of computers[383,384]. Molecular ecologists are integrated into a large variety of ecological research agendas, however, the paucity and maintenance costs of high-level biosafety laboratories is a hurdle for independent molecular ecology research labs to take part in the study of the molecular ecology of zoonotic diseases, such as Avian Influenza. As a consequence, the research on the ecology of Avian Influenza is largely dominated and monopolised by virological laboratories. This is not a problem *per se*. Extensive collaborations of virologists and molecular ecologists have lead to fruitful results (for instance, see [2,306,307,325,377,385]). However, the complete dependence of smaller molecular ecology laboratories on a few well-equipped virology groups is sub-optimal in a scientific context. Although collaborative efforts often lead to great achievements it is paramount for scientists to maintain independence.

One way of maintaining scientific independence is to be able to collect samples without being affiliated to one of the very few large surveillance programs that usually get installed per continent. The development of techniques to sample, store and process faecal samples suspect of being AI-positive samples in Chapters 6 and 7 is a first step towards more independence. Ultimately, these techniques have the potential to motivate smaller molecular ecological laboratories to invest their creativity and imagination into projects that involve sampling of animals for AI. Additionally, also traditional multi-national surveillance programs can benefit from methods that help leveraging sampling under arduous conditions. As I have shown in the example of the mallard in Chapters 2 and 4 bird populations can be highly connected. When monitoring only specific corridors, authorities may miss an approaching infection wave in the wild, if environmental conditions favour another route. Current surveillance methods depend on the installation of well-equipped field stations, the maintenance of cold chains, and the implementation of biosafety measures during transport. The use of FTA cards or similar methods (reviewed in Chapter 7), also in more traditional projects, can aid in filling gaps in regions where advanced AI sampling infrastructure is not yet built up, or cannot be installed at all.

Outlook

Understanding the ecology of the mallard, probably the most important host species of Avian Influenza (AI) in the wild[10,66], is paramount for understanding the ecology of AI. Since AIVs have caused or contributed to several global pandemics in humans during the past century, causing millions of deaths[31], the study of AI is perceived as very important by the public. Further, since the natural reservoir of AI encompasses so many bird species it is difficult to start ecological research on all these multiple hosts. Hence, a lot more effort is allocated to the study of the single disease AI than to the multiple host system of water birds, especially ducks.

This thesis provides powerful modern genetic tools for the study of mallard molecular ecology (Chapter 3), and provides first insights into the global population structure of this waterfowl species (Chapters 2 and 4), hence facilitating the research of a host of AI on the appropriate scale: the world. I believe that the work presented here can be seen as an exemplary project in this respect. Further efforts to study global scale population structure with respect to AIV dispersal are essential, and some first steps are currently being undertaken. Such a research program, aiming

at *Aythya* species (diving ducks), is carried out at the University of Bern, Switzerland. This, as yet, unpublished data also aims to characterise global population structure. A further project on mallards, using the SNP markers described in my thesis (Chapter 3), is starting in Kristianstad, Sweden, setting out to study the aspects of the European mallard population with respect to mallard farming, which I highlight in my thesis (Chapters 2 and 4). The use of the same SNP set in these two projects will aid the joint analysis of genotype data collected here and that collected in the new project.

Hopefully, there will be many more studies on various water bird species on a global scale, using genetic makers that are sufficiently powerful for global projects, and representative for the whole genome. I have proven that developing such marker systems is possible for any non-model species and allows for data sharing. In Chapter 5 I show that the study of other duck species is not only important because they are also a part of the AI reservoir, but because their gene pools are directly linked to each other. Thereby, I offer a hypothesis why ducks and other waterfowl are such a successful group of animals. Extensive ongoing inter-specific hybridisation between many species within waterfowl, producing viable and fertile hybrids[247], could be seen as a mechanism similar to horizontal gene transfer in bacteria. On the one hand, a greatly increased genetically linked population can maintain more polymorphisms, thereby allowing a more complex and finer tuned co-evolution of AI and the birds. On the other hand, favourable adaptations to living with viral infections can spread from one species to another.

Aside from the various scientifically relevant and novel findings in the presented chapters, touching upon diverse facets such as genetic markers (Chapters 2 and 3), population genomics (Chapter 4), evolutionary genetics (Chapter 5) and sample collection technology (Chapters 6 and 7), my thesis will help answer key questions of societal relevance about AI research, such as which hosts potentially contribute most to the spread of this disease, and through which channels this spread occurs. As I conclude in my thesis, the mallard is perfectly suited to not only constitute the major natural reservoir of Avian Influenza, but also to spread the virus efficiently and quickly through a highly interconnected network. It is important to organise future surveillance into many small sampling stations, since an Influenza outbreak can spread anywhere due to the high connectivity of the mallard populations.

References

1. Cavanagh, D. Coronaviruses in poultry and other birds. Avian Pathol. 34, 439-448 (2005).
2. Olsen, B. *et al.* Global patterns of influenza A virus in wild birds. Science 312, 384-388 (2006).
3. Li, K. S. *et al.* Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430, 209-213 (2004).
4. Chen, H. *et al.* Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. Proc. Natl. Acad. Sci. U. S. A. 103, 2845-2850 (2006).
5. Johnson, K. P. & Sorenson, M. D. Phylogeny and biogeography of dabbling ducks (genus: *Anas*): A comparison of molecular and morphological evidence. Auk 116, 792-805 (1999).
6. Gonzalez, J., Düttmann, H. & Wink, M. Phylogenetic relationships based on two mitochondrial genes and hybridization patterns in Anatidae. J. Zool. 279, 310-318 (2009).
7. Livezey, B. C. A phylogenetic analysis and classification of recent dabbling ducks (Tribe Anatini) based on comparative morphology. Auk 108, 471-507 (1991).
8. Cramp, S. & Simmons, K. E. L. *Handbook of the Birds of Europe, the Middle East, and North Africa: The Birds of the Western Palearctic. Volume 1 - Ostrich to Ducks.* (Oxford University Press, Oxford, UK, 1977).
9. Figuerola, J. & Green, A. J. The evolution of sexual dimorphism in relation to mating patterns, cavity nesting, insularity and sympatry in the Anseriformes. Funct. Ecol. 14, 701-710 (2000).
10. Atkinson, P. W. *et al. Urgent preliminary assessment of ornithological data relevant to the spread of Avian Influenza in Europe* (eds. Delany, S., Veen, J. & Clark, A.) (Wetlands International, Wageningen, The Netherlands, 2006).
11. Kulikova, I. V. *et al.* Phylogeography of the mallard (*Anas platyrhynchos*): Hybridization, dispersal, and lineage sorting contribute to complex geographic structure (vol 122, pg 949, 2005). Auk 122, 1309-1309 (2005).
12. Scott, D. A. & Rose, P. M. *Atlas of Anatidae populations in Africa and Western Eurasia.* (Wetlands International Publication No. 41, Wetlands International, Wageningen, The Netherlands, 1996).
13. Alerstam, T. *Bird migration* (Cambridge University Press, Cambridge, 1990).
14. Wink, M. Use of DNA markers to study bird migration. J. Ornith. 147, 234-244 (2006).
15. Boere, G. C. & Stroud, D. A. The flyway concept: what it is and what it isn't. In *Waterbirds around the world* (eds. Boere, G. C., Galbraith, C. A. & Stroud, D. A.) p. 40-47 (The Stationary Office, Edinburgh, UK, 2006).
16. Miyabayashi, Y. & Mundkur, T. *Atlas of Key Sites for Anatidae in the East Asian Flyway* (Wetlands International - Japan, Tokyo, and Wetlands International - Asia Pacific, Kuala Lumpur, 1999).
17. Anon. U.S. Fish and Wildlife Service. ([http://www.flyways.us/flyways/info#flyways-bio]).

18. Bowlin, M. S. *et al.* Grand challenges in migration biology. Integr. Comp. Biol. 50, 261-279 (2010).
19. Jonker, R. M., Eichhorn, G., van Langevelde, F. & Bauer, S. Predation danger can explain changes in timing of migration: The case of the Barnacle goose. PLoS ONE 5, e11369 (2010).
20. Jonker, R. M. *et al.* Genetic consequences of breaking migratory traditions in barnacle geese. In *Revolutionary non-migratory migrants, Ph.D. Thesis, Wageningen University* (ed. Jonker, R. M.) p. (2011).
21. Kaleta, E. F., Hergarten, G. & Yilmaz, A. Avian influenza A viruses in birds - an ecological, ornithological and virological view. Dtsch. Tierärztl. Wochenschr. 112, 448-456 (2005).
22. Hanson, B. A., Stallknecht, D. E., Swayne, D. E., Lewis, L. A. & Senne, D. A. Avian influenza viruses in Minnesota ducks during 1998-2000. Avian Dis. 47, 867-871 (2003).
23. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56, 152-179 (1992).
24. Fouchier, R. A. M. *et al.* Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J. Virol. 79, 2814-2822 (2005).
25. Aymard-Henry, M. *et al.* Influenzavirus neuraminidase and neuraminidase-inhibition test procedures. Bull. W.H.O. 48, 199-202 (1973).
26. Salk, J. E. Simplified procedure for titrating hemagglutinating capacity of influenza virus and the corresponding antibody. J. Immunol. 49, 87-98 (1944).
27. Rimmelzwaan, G. F., Baars, M., Claas, E. C. J. & Osterhaus, A. D. M. E. Comparison of RNA hybridization, hemagglutination assay, titration of infectious virus and immunofluorescence as methods for monitoring influenza virus replication in vitro. J. Virol. Methods 74, 57-66 (1998).
28. Hoffmann, E., Stech, J., Guan, Y., Webster, R. G. & Perez, D. R. Universal primer set for the full-length amplification of all influenza A viruses. Arch. Virol. 146, 2275-2289 (2001).
29. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. Nucleic Acids Res. 39, D32-D37 (2011).
30. Cox, N. J. & Subbarao, K. Influenza. Lancet 354, 1277-1282 (1999).
31. Taubenberger, J. K. & Morens, D. M. 1918 Influenza: The mother of all pandemics. Emerg. Infect. Dis. 12, 15-22 (2006).
32. Butler, D. Swine flu goes global. Nature 458, 1082-1083 (2009).
33. Hulse-Post, D. J. *et al.* Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia. Proc. Natl. Acad. Sci. U. S. A. 102, 10682-10687 (2005).
34. Keawcharoen, J. *et al.* Wild ducks as long-distance vectors of highly pathogenic avian influenza virus (H5N1). Emerg. Infect. Dis. 14, 600-607 (2008).
35. Leigh Perkins, L. E. & Swayne, D. E. Pathogenicity of a Hong Kong-origin H5N1 highly pathogenic avian influenza virus for emus, geese, ducks, and pigeons. Avian Dis. 46, 53-63 (2002).

36. Jourdain, E. *et al*. Influenza virus in a natural host, the mallard: Experimental infection data. PLoS ONE 5, e8935 (2010).

37. Sturm-Ramirez, K. M. *et al*. Reemerging H5N1 Influenza viruses in Hong Kong in 2002 are highly pathogenic to ducks. J. Virol. 78, 4892-4901 (2004).

38. Pantin-Jackwood, M. J. & Swayne, D. E. Pathobiology of Asian highly pathogenic avian influenza H5N1 virus infections in ducks. Avian Dis. 51, 250-259 (2007).

39. Swayne, D. E. & Pantin-Jackwood, M. Pathogenicity of avian influenza viruses in poultry. Dev. Biol. 124, 61-67 (2006).

40. Normile, D. Avian influenza. Are wild birds to blame? Science 310, 426-428 (2005).

41. Normile, D. Avian influenza. Evidence points to migratory birds in H5N1 spread. Science 311, 1225 (2006).

42. Normile, D. Avian influenza. Wild birds only partly to blame in spreading H5N1. Science 312, 1451 (2006).

43. Si, Y. *et al*. Spatio-temporal dynamics of global H5N1 outbreaks match bird migration patterns. Geospat. Health 4, 65-78 (2009).

44. Bin Muzaffar, S., Ydenberg, R. C. & Jones, I. L. Avian influenza: An ecological and evolutionary perspective for waterbird scientists. Waterbirds 29, 243-257 (2006).

45. Keller, L. Adaptation and the genetics of social behaviour. Phil. Trans. R. Soc. B 364, 3209-3216 (2009).

46. Klepsatel, P. & Flatt, T. The genomic and physiological basis of life history variation in a butterfly metapopulation. Mol. Ecol. 20, 1795-1798 (2011).

47. Chan, L. M., Brown, J. L. & Yoder, A. D. Integrating statistical genetic and geospatial methods brings new power to phylogeography. Mol. Phylogenet. Evol. 59, 523-537 (2011).

48. Holderegger, R. & Wagner, H. H. Landscape genetics. BioScience 58, 199-207 (2008).

49. Cordellier, M. & Pfenninger, M. Inferring the past to predict the future: Climate modelling predictions and phylogeography for the freshwater gastropod *Radix balthica* (Pulmonata, Basommatophora). Mol. Ecol. 18, 534-544 (2009).

50. Bonin, A., Taberlet, P., Miaud, C. & Pompanon, F. Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Mol. Biol. Evol. 23, 773-783 (2006).

51. Welch, J. J., Eyre-Walker, A. & Waxman, D. Divergence and polymorphism under the nearly neutral theory of molecular evolution. J. Mol. Evol. 67, 418-426 (2008).

52. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating Inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467 (1977).

53. Saiki, R. K. *et al*. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487-491 (1988).

54. Cooke, F. & Cooch, F. G. The genetics of polymorphism in the goose *Anser caerulescens*. Evolution 22, 289-300 (1968).

55. Mendel, J. G. Versuche über Pflanzen-Hybriden. Verhandlungen Des Naturforschenden Vereines In Brünn 4, 3-47 (1866).

56. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum. Genet. 122, 565-581 (2008).

57. Levene, P. The structure of yeast nucleic acid. J. Biol. Chem. 40, 415-424 (1919).

58. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. Nature 171, 737-738 (1953).

59. Avery, O., MacLeod, C. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. 79, 137-158 (1944).

60. Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. 36, 39-56 (1952).

61. Shaw, C. R. Electrophoretic variation in enzymes. Science 149, 936-943 (1965).

62. Schlötterer, C. The evolution of molecular markers - Just a matter of fashion? Nat. Rev. Genet. 5, 63-69 (2004).

63. Kulikova, I. V. *et al.* Phylogeography of the Mallard (*Anas platyrhynchos*): Hybridization, dispersal, and lineage sorting contribute to complex geographic structure. Auk 122, 949-965 (2005).

64. Brookes, A. J. The essence of SNPs. Gene 234, 177-186 (1999).

65. Morin, P. A., Luikart, G. & Wayne, R. K. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19, 208-216 (2004).

66. Nishiura, H., Hoye, B., Klaassen, M., Bauer, S. & Heesterbeek, H. How to find natural reservoir hosts from endemic prevalence in a multi-host population: A case study of influenza in waterfowl. Epidemics 1, 118-128 (2009).

67. Smith, L. M. & Burgoyne, L. A. Collecting, archiving and processing DNA from wildlife samples using FTA® databasing paper. BMC Ecol. 4, 4 (2004).

68. Thompson, A. L. On 'abmigration' among the ducks, an anomaly shown by the results of bird-marking. Proc. Int. Ornithol. Congr. 7, 382-388 (1931).

69. Guillemain, M., Sadoul, N. & Simon, G. European flyway permeability and abmigration in Teal *Anas crecca*, an analysis based on ringing recoveries. Ibis 147, 688-696 (2005).

70. Elmberg, J. Are dabbling ducks major players or merely noise in freshwater ecosystems? A European perspective, with references to population limitation and density dependence. Wildfowl, 9-23 (2009).

71. Laikre, L., Palmé, A., Josefsson, M., Utter, F. & Ryman, N. Release of alien populations in Sweden. Ambio 35, 255-261 (2006).

72. Champagnon, J., Guillemain, M., Gauthier-Clerc, M., Lebreton, J. D. & Elmberg, J. Consequences of massive bird releases for hunting purposes: Mallard *Anas platyrhynchos* in the Camargue, southern France. Wildfowl, 184-191 (2009).

73. Edgell, M. C. R. Trans-hemispheric movements of Holarctic Anatidae - The Eurasian wigeon (*Anas penelope* L.) in North-America. J. Biogeogr. 11, 27-39 (1984).

74. Koehler, A. V., Pearce, J. M., Flint, P. L., Franson, J. C. & Ip, H. S. Genetic evidence of intercontinental movement of avian influenza in a migratory bird: The northern pintail (*Anas acuta*). Mol. Ecol. 17, 4754-4762 (2008).

75. Ramey, A. M. *et al.* Transmission and reassortment of avian influenza viruses at the Asian-North American interface. Virology 406, 352-359 (2011).

76. Winker, K. *et al.* Movements of birds and avian influenza from Asia into Alaska. Emerg. Infect. Dis. 13, 547-552 (2007).

77. Kraus, R. H. S. *et al.* Avian influenza surveillance: On the usability of FTA® cards to solve biosafety and transport issues. Wildfowl, 215-223 (2009).

78. Kraus, R. H. S. *et al.* Avian influenza surveillance with FTA Cards: field methods, biosafety, and transportation issues solved. J. Vis. Exp. 54, e2832 (2011).

79. Avise, J. C., Ankney, C. D. & Nelson, W. S. Mitochondrial gene trees and the evolutionary relationship of Mallard and Black Ducks. Evolution 44, 1109-1119 (1990).

80. Hey, J. & Nielsen, R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167, 747-760 (2004).

81. Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. U. S. A. 104, 2785-2790 (2007).

82. Kulikova, I. V., Zhuravlev, Y. N. & McCracken, K. G. Asymmetric hybridization and sex-biased gene flow between Eastern Spot-billed Ducks (*Anas zonorhyncha*) and Mallards (*A. platyrhynchos*) in the Russian Far East. Auk 121, 930-949 (2004).

83. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595 (1989).

84. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147, 915-925 (1997).

85. Saitou, N. & Nei, M. The neighbor-joining method - a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425 (1987).

86. Wright, S. Evolution in Mendelian populations. Genetics 16, 97-159 (1931).

87. Pearce, J. M. Philopatry: A return to origins. Auk 124, 1085-1087 (2007).

88. Peters, J. L. & Omland, K. E. Population structure and mitochondrial polyphyly in North American Gadwalls (*Anas strepera*). Auk 124, 444-462 (2007).

89. Lokemoen, J. T., Duebbert, H. F. & Sharp, D. E. Homing and reproductive habits of mallards, gadwalls, and blue-winged teal. Wildl. Monogr. 106, 3-28 (1990).

90. Robertson, G. J. & Cooke, F. Winter philopatry in migratory waterfowl. Auk 116, 20-34 (1999).

91. Bishop, R. A. & Andrews, R. D. Survival and homing of female Mallards. J. Wildl. Manage. 42, 192-196 (1978).

92. Cronin, M. A., Grand, J. B., Esler, D., Derksen, D. V. & Scribner, K. T. Breeding populations of northern pintails have similar mitochondrial DNA. Can. J. Zool. 74, 992-999 (1996).

93. Pearce, J. M. *et al.* Lack of spatial genetic structure among nesting and wintering King Eiders. Condor 106, 229-240 (2004).

94. Peters, J. L., Gretes, W. & Omland, K. E. Late Pleistocene divergence between eastern and western populations of wood ducks (*Aix sponsa*) inferred by the 'isolation with migration' coalescent method. Mol. Ecol. 14, 3407-3418 (2005).

95. Gay, L., Defos Du Rau, P., Mondain-Monval, J.-Y. & Crochet, P.-A. Phylogeography of a game species: The red-crested pochard (*Netta rufina*) and consequences for its management. Mol. Ecol. 13, 1035-1045 (2004).

96. Champagnon, J., Guillemain, M., Elmberg, J., Folkesson, K. & Gauthier-Clerc, M. Changes in Mallard *Anas platyrhynchos* bill morphology after 30 years of supplemental stocking. Bird Study 57, 344-351 (2010).

97. Laikre, L., Schwartz, M. K., Waples, R. S. & Ryman, N. Compromising genetic diversity in the wild: Unmonitored large-scale release of plants and animals. Trends Ecol. Evol. 25, 520-529 (2010).

98. Baratti, M., Cordaro, M., Dessì-Fulgheri, F., Vannini, M. & Fratini, S. Molecular and ecological characterization of urban populations of the mallard (*Anas platyrhynchos* L.) in Italy. Ital. J. Zool. 76, 330-339 (2009).

99. Hofreiter, M. & Stewart, J. Ecological change, range fluctuations and population dynamics during the Pleistocene. 19, R584–R594 (2009).

100. Peters, J. L. *et al.* Phylogenetics of wigeons and allies (Anatidae: *Anas*): The importance of sampling multiple loci and multiple individuals. Mol. Phylogenet. Evol. 35, 209-224 (2005).

101. Li, J. & Wang, J. X. L. A modified zonal index and its physical sense. Geophys. Res. Lett. 30, 34 (2003).

102. Lamb, H. H. Our understanding of global wind circulation and climatic variations. Bird Study 22, 121-141 (1975).

103. Liechti, F. Birds: Blowin' by the wind? J. Ornithol. 147, 202-211 (2006).

104. Richardson, W. J. Timing and amount of bird migration in relation to weather: a review. Oikos 30, 224-272 (1978).

105. Lyngs, P. Migration and winter ranges of birds in Greenland. An analysis of ringing recoveries. Dan. Ornitol. Foren. Tidsskr. 97, 1-167 (2003).

106. Zeddeman, A., van Hooft, P., Prins, H. H. T. & Kraus, R. H. S. Mallard (*Anas platyrhynchos*) gene pool connectivity between Greenland and Eastern Canada, Great Britain and The Netherlands. In *Nuuk Ecological Research Operations, 2nd Annual Report*, 2008. (eds. Jensen, L. M. & Rasch, M.) p. 67 (National Environmental Research Institute©, Aarhus University – Denmark, Aarhus, 2009).

107. Evrard, J. O. Male philopatry in Mallards. Condor 92, 247-248 (1990).

108. Rohwer, F. C. & Anderson, M. G. Female-biased philopatry, monogamy, and the timing of pair formation in migratory waterfowl. In *Current Ornithology* (ed. Johnston, R. F.) p. 187-221 (Plenum Press, New Work, 1988).

109. Huang, Y. *et al.* A genetic and cytogenetic map for the duck (*Anas platyrhynchos*). Genetics 173, 287-296 (2006).

110. Kraus, R. H. S. *et al.* Genome wide SNP discovery, analysis and evaluation in mallard (*Anas platyrhynchos*). BMC Genomics 12, 150 (2011).

111. Mesa, C. M., Thulien, K. J., Moon, D. A., Veniamin, S. M. & Magor, K. E. The dominant MHC class I gene is adjacent to the polymorphic TAP2 gene in the duck, *Anas platyrhynchos*. Immunogenetics 56, 192-203 (2004).

112. Moon, D. A., Veniamin, S. M., Parks-Dely, J. A. & Magor, K. E. The MHC of the duck (*Anas platyrhynchos*) contains five differentially expressed class I genes. J. Immunol. 175, 6702-6712 (2005).

113. Greenwood, P. J. Mating systems, philopatry and dispersal in birds and mammals. Anim. Behav. 28, 1140-1162 (1980).

114. Pardini, A. T. *et al.* Sex-biased dispersal of great white sharks. Nature 412, 139-140 (2001).

115. Mank, J. E., Carlson, J. E. & Brittingham, M. C. A century of hybridization: Decreasing genetic distance between American black ducks and mallards. Conserv. Genet. 5, 395-403 (2004).

116. McCracken, K. G., Johnson, W. P. & Sheldon, F. H. Molecular population genetics, phylogeography, and conservation biology of the mottled duck (*Anas fulvigula*). Cons. Gen. 2, 87-102 (2001).

117. Funk, D. J. & Omland, K. E. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Ann. Rev. Ecol. Evol. Syst. 34, 397-423 (2003).

118. Maddison, W. & Knowles, L. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21-30 (2006).

119. Omland, K. E. Examining two standard assumptions of ancestral reconstructions: Repeated loss of dichromatism in dabbling ducks (Anatini). Evolution 51, 1636-1646 (1997).

120. Desjardins, P. & Morais, R. Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. J. Mol. Biol. 212, 599-634 (1990).

121. Sorenson, M. D. & Fleischer, R. C. Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus. Proc. Natl. Acad. Sci. U. S. A. 93, 15239-15243 (1996).

122. Sorenson, M. D., Ast, J. C., Dimcheff, D. E., Yuri, T. & Mindell, D. P. Primers for a PCR-based approach to mitochondrial genome sequencing in birds and other vertebrates. Mol. Phylogenet. Evol. 12, 105-114 (1999).

123. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596-1599 (2007).

124. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680 (1994).

125. Tajima, F. & Nei, M. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1, 269-285 (1984).

126. Clement, M., Posada, D. & Crandall, K. A. TCS: A computer program to estimate gene genealogies. Mol. Ecol. 9, 1657-1659 (2000).

127. Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496-2497 (2003).

128. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Res. 10, 564-567 (2010).

129. Guo, S. W. & Thompson, E. A. Performing the exact Test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48, 361-372 (1992).

130. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358-1370 (1984).

131. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular Variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial-DNA restriction data. Genetics 131, 479-491 (1992).

132. Weir, B. S. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data* (Sinauer Assoc., Inc., Sunderland, MA, USA, 1996).

133. Peters, J. L., Zhuravlev, Y., Fefelov, I., Logie, A. & Omland, K. E. Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas spp.*). Evolution 61, 1992-2006 (2007).

134. Peters, J. L., Zhuravlev, Y. N., Fefelov, I., Humphries, E. M. & Omland, K. E. Multilocus phylogeography of a Holarctic duck: Colonization of North America from Eurasia by gadwall (*Anas strepera*). Evolution 62, 1469-1483 (2008).

135. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160-147 (1985).

136. Palsbøll, P. J., Bérubé, M., Aguilar, A., Notarbartolo-Di-Sciara, G. & Nielsen, R. Discerning between recurrent gene flow and recent divergence under a finite-site mutation model applied to North Atlantic and Mediterranean Sea fin whale (*Balaenoptera physalus*) populations. Evolution 58, 670-675 (2004).

137. Gilbert, M. *et al.* Free-grazing ducks and highly pathogenic avian influenza, Thailand. Emerg. Infect. Dis. 12, 227-234 (2006).

138. Munster, V. J. *et al.* Towards improved influenza A virus surveillance in migrating birds. Vaccine 24, 6729-6733 (2006).

139. Si, Y. *et al.* Environmental factors influencing the spread of the highly pathogenic avian influenza H5N1 virus in wild birds in Europe. Ecol. Soc. 15, 26 (2010).

140. Paul, M. *et al.* Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: prospects from a spatial-based model. Vet. Res. 41, 28 (2010).

141. Bauer, H., Bezzel, E. & Fiedler, W. *Kompendium der Vögel Mitteleuropas* (Aula-Verlag, Wiebelsheim, Germany, 2005).

142. Morin, P. A., Martien, K. K. & Taylor, B. L. Assessing statistical power of SNPs for population structure and conservation studies. Mol. Ecol. Res. 9, 66-73 (2009).

143. Ryman, N. *et al.* Power for detecting genetic divergence: Differences between statistical methods and marker loci. Mol. Ecol. 15, 2031-2045 (2006).

144. Anon. Food and Agriculture Organisation of the United Nations. ([http://faostat.fao.org/]).

145. Huang, C.-W. *et al.* Duck (*Anas platyrhynchos*) linkage mapping by AFLP fingerprinting. Genet Sel Evol 41, 28 (2009).

146. Bennett, S. Solexa Ltd. Pharmacogenomics 5, 433-438 (2004).

147. Bentley, D. R. Whole-genome re-sequencing. Curr Opin Genet Dev 16, 545-552 (2006).

148. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34, e22 (2006).

149. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513-516 (2000).

150. Kerstens, H. H. D. *et al.* Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: Applied to Turkey. BMC Genomics 10, 479 (2009).

151. van Bers, N. E. *et al.* Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. Mol. Ecol. 19, 89-99 (2010).

152. Sánchez, C. C. *et al.* Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics 10, 559 (2009).

153. Ramos, A. M. *et al.* Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE 4, e6524 (2009).

154. Wiedmann, R. T., Smith, T. P. L. & Nonneman, D. J. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genetics 9, 81 (2008).

155. Van Tassell, C. P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 247-252 (2008).

156. Fillon, V. *et al.* FISH mapping of 57 BAC clones reveals strong conservation of synteny between Galliformes and Anseriformes. Anim Genet 38, 303-307 (2007).

157. Skinner, B. M. *et al.* Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. BMC Genomics 10, 357 (2009).

158. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-194 (1998).

159. Amaral, A. J. *et al.* Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics 10, 374 (2009).

160. Gregory, T. R. *et al.* Eukaryotic genome size databases. Nucleic Acids Res. 35, D332–D338 (2007).

161. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-1858 (2008).

162. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrates genomes. Hum. Genet. 83, 181-188 (1989).

163. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum. Genom. 4, 406-410 (2010).

164. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34, 275-305 (2002).

165. Scarano, E., Iaccarino, M., Grippo, P. & Parisi, E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. Proc. Natl. Acad. Sci. U.S.A. 57, 1394-1400 (1967).

166. Sherry, S. T. *et al.* DbSNP: The NCBI database of genetic variation. Nucleic Acids Res. 29, 308-311 (2001).

167. Kao, W. C., Stevens, K. & Song, Y. S. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res 19, 1884-1895 (2009).

168. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: A public knowledge-base of divergence times among organisms. Bioinformatics 22, 2971-2972 (2006).

169. Illumina. Protocol for Whole Genome Sequencing using Solexa Technology. BioTechniques Protocol Guide 29 (2006).

170. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859-1875 (2005).

171. Kent, W. J. BLAT - The BLAST-like alignment tool. Genome Res 12, 656-664 (2002).

172. R Development Core Team. *R: A language and environment for statistical computing, http://www.R-project.org* (R Foundation for Statistical Computing, Vienna, Austria, 2009).

173. Avise, J. C. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Evol. Syst. 18, 489-522 (1987).

174. Knowles, L. L. The burgeoning field of statistical phylogeography. J. Evol. Biol. 17, 1-10 (2004).

175. Carstens, B. C. & Richards, C. L. Integrating coalescent and ecological niche modeling in comparative phylogeography. Evolution 61, 1439-1454 (2007).

176. Nielsen, R. & Beaumont, M. A. Statistical inferences in phylogeography. Mol. Ecol. 18, 1034-1047 (2009).

177. Beaumont, M. A. *et al.* In defence of model-based inference in phylogeography. Mol. Ecol. 19, 436-446 (2010).

178. Templeton, A. R. Coalescent-based, maximum likelihood inference in phylogeography. Mol. Ecol. 19, 431-435 (2010).

179. Bloomquist, E. W., Lemey, P. & Suchard, M. A. Three roads diverged? Routes to phylogeographic inference. Trends Ecol. Evol. 25, 626-632 (2010).

180. Kingman, J. F. C. The coalescent. Stoch. Proc. Appl. 13, 235-248 (1982).

181. Kingman, J. F. C. On the genealogy of large populations. J. Appl. Probab. 19, 27-43 (1982).

182. Beerli, P. & Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. U.S.A. 98, 4563-4568 (2001).

183. Kuhner, M. K. LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22, 768-770 (2006).

184. Wang, Y. & Hey, J. Estimating divergence parameters with small samples from a large number of loci. Genetics 184, 363-379 (2010).

185. Estoup, A., Jarne, P. & Cornuet, J. M. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. 11, 1591-1604 (2002).

186. Selkoe, K. A. & Toonen, R. J. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol. Lett. 9, 615-629 (2006).

187. Black IV, W. C., Baer, C. F., Antolin, M. F. & DuTeau, N. M. Population genomics: Genome-wide sampling of insect populations. Ann. Rev. Entomol. 46, 441-469 (2001).

188. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: From genotyping to genome typing. Nat. Rev. Genet. 4, 981-994 (2003).

189. Campbell, N. R. & Narum, S. R. Development of 54 novel single-nucleotide polymorphism (SNP) assays for sockeye and coho salmon and assessment of available SNPs to differentiate stocks within the Columbia River. Mol. Ecol. Res. 11, 20-30 (2011).

190. Mesnick, S. L. *et al.* Sperm whale population structure in the eastern and central North Pacific inferred by the use of single-nucleotide polymorphisms, microsatellites and mitochondrial DNA. Mol. Ecol. Res. 11, 278-298 (2011).

191. Sacks, B. N., Moore, M., Statham, M. J. & Wittmer, H. U. A restricted hybrid zone between native and introduced red fox (*Vulpes vulpes*) populations suggests reproductive barriers and competitive exclusion. Mol. Ecol. 20, 326-341 (2011).

192. Kovach, A. I., Breton, T. S., Berlinsky, D. L., Maceda, L. & Wirgin, I. Fine-scale spatial and temporal genetic structure of Atlantic cod off the Atlantic coast of the USA. Mar. Ecol. Prog. Ser. 410, 177-195 (2010).

193. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135-1145 (2008).

194. Metzker, M. L. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31-46 (2010).

195. Jonker, R. M. *et al.* The development of a genome wide SNP set for the Barnacle Goose *Branta leucopsis*. In *Revolutionary non-migratory migrants, Ph.D. Thesis, Wageningen University* (ed. Jonker, R. M.) p. (2011).

196. Willing, E. M. *et al.* Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. Mol. Ecol. 19, 968-984 (2010).

197. Santure, A. W. *et al*. On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Mol. Ecol. 19, 1439-1451 (2010).

198. Williams, L. M. & Oleksiak, M. F. Ecologically and evolutionarily important SNPs identified in natural populations. Mol. Biol. Evol. 28, 1817-1826 (2011).

199. Davey, J. W. *et al*. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499-510 (2011).

200. Delany, S. N. & Scott, D. A. Wetlands International's Flyway Atlas series: establishing the geographical limits of waterbird populations. In *Waterbirds around the world* (eds. Boere, G. C., Galbraith, C. A. & Stroud, D. A.) p. 574-581 (The Stationery Office, Edinburgh, UK, 2006).

201. Reudink, M. W. *et al*. Panmixia on a continental scale in a widely distributed colonial waterbird. Biol. J. Linnean Soc. 102, 583-592 (2011).

202. ISO3166. ISO 3166 Maintenance agency. (2007).

203. Posada, D. & Crandall, K. A. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. 16, 37-45 (2001).

204. Bryant, D. & Moulton, V. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255-265 (2004).

205. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68-73 (1998).

206. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254-267 (2006).

207. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945-959 (2000).

208. Wang, J. Coancestry: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. Mol. Ecol. Res. 11, 141-145 (2011).

209. Milligan, B. G. Maximum-likelihood estimation of relatedness. Genetics 163, 1153-1167 (2003).

210. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14, 2611-2620 (2005).

211. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genet. 11, 94 (2010).

212. Jombart, T. Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403-1405 (2008).

213. Beerli, P. & Felsenstein, J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152, 763-773 (1999).

214. Kass, R. E. & Raftery, A. E. Bayes Factors. J. Amer. Statist. Assoc. 90, 773-795 (1995).

215. Beerli, P. & Palczewski, M. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics 185, 313-326 (2010).

216. Beerli, P. Tutorial: Comparison of gene flow models using Bayes Factors. Web Tutorial http://popgen.sc.fsu.edu/Migrate/Tutorials/Entries/2010/7/12_Day_of_longboarding.html (2010).

217. Latch, E. K., Dharmarajan, G., Glaubitz, J. C. & Rhodes Jr, O. E. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Cons. Gen. 7, 295-302 (2006).

218. Whitlock, M. C. & McCauley, D. E. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. Heredity 82, 117-125 (1999).

219. Pearse, D. E. & Crandall, K. A. Beyond F_{ST}: Analysis of population genetic data for conservation. Cons. Gen. 5, 585-602 (2004).

220. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156, 439-447 (2000).

221. Brumfield, R. T., Beerli, P., Nickerson, D. A. & Edwards, S. V. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol. Evol. 18, 249-256 (2003).

222. Bradbury, I. R. *et al.* Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, *Gadus morhua*. Mol. Ecol. Res. 11, 218-225 (2011).

223. Rhodes Jr., O. E., Smith, L. M. & Chesser, R. K. Apportionment of genetic variance in migrating and wintering Mallards. Can. J. Zool. 73, 1182-1185 (1995).

224. Waples, R. S. & Gaggiotti, O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15, 1419-1439 (2006).

225. Rypien, K. L., Andras, J. P. & Harvell, C. D. Globally panmictic population structure in the opportunistic fungal pathogen *Aspergillus sydowii*. Mol. Ecol. 17, 4068-4078 (2008).

226. Jorgensen, S. J. *et al.* Philopatry and migration of Pacific white sharks. Proc. R. Soc. B. 277, 679-688 (2010).

227. Han, Y. S., Hung, C. L., Liao, Y. F. & Tzeng, W. N. Population genetic structure of the Japanese eel *Anguilla japonica*: Panmixia at spatial and temporal scales. Mar. Ecol. Prog. Ser. 401, 221-232 (2010).

228. Als, T. D. *et al.* All roads lead to home: Panmixia of European eel in the Sargasso Sea. Mol. Ecol. 20, 1333-1346 (2011).

229. Friesen, V. L., Burg, T. M. & McCoy, K. D. Mechanisms of population differentiation in seabirds: Invited review. Mol. Ecol. 16, 1765-1785 (2007).

230. Rhymer, J. M. Extinction by hybridization and introgression in anatine ducks. Acta Zool. Sin. 52(Supplement), 583-585 (2006).

231. IUCN. IUCN Red List of Threatened Species. Downloaded on 13 July 2011, Version 2011.1 (2011).

232. Eaton, M. A. *et al.* Birds of conservation concern 3 the population status of birds in the united kingdom, channel islands and isle of man. Br. Birds 102, 296-341 (2009).

233. Mills, L. S. & Allendorf, F. W. The one-migrant-per-generation rule in conservation and management. Conserv. Biol. 10, 1509-1518 (1996).

234. Stallknecht, D. E., Shane, S. M., Kearney, M. T. & Zwank, P. J. Persistence of avian influenza viruses in water. Avian Dis. 34, 406-411 (1990).

235. Stallknecht, D. E., Kearney, M. T., Shane, S. M. & Zwank, P. J. Effects of pH, temperature, and salinity on persistence of avian influenza viruses in water. Avian Dis. 34, 412-418 (1990).

236. Ito, T. *et al.* Perpetuation of influenza A viruses in Alaskan waterfowl reservoirs. Arch. Virol. 140, 1163-1172 (1995).

237. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154, 931-942 (2000).

238. Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. The discovery of single-nucleotide polymorphisms - And inferences about human demographic history. Am. J. Hum. Genet. 69, 1332-1347 (2001).

239. Nielsen, R., Hubisz, M. J. & Clark, A. G. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168, 2373-2382 (2004).

240. Rosenblum, E. B. & Novembre, J. Ascertainment bias in spatially structured populations: A case study in the Eastern Fence Lizard. J. Hered. 98, 331-336 (2007).

241. Guillot, G. & Foll, M. Correcting for ascertainment bias in the inference of population structure. Bioinformatics 25, 552-554 (2009).

242. Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Res. 11, 117-122 (2011).

243. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. Genetics 162, 2025-2035 (2002).

244. Lynch, M. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25, 2409-2419 (2008).

245. Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol. Ecol. 19, 277-284 (2010).

246. Gourbiére, S. & Mallet, J. Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". Evolution 64, 1-24 (2010).

247. Grant, P. R. & Grant, B. R. Hybridization of bird species. Science 256, 193-197 (1992).

248. Tubaro, P. L. & Lijtmaer, D. A. Hybridization patterns and the evolution of reproductive isolation in ducks. Biol. Jour. Linn. Soc. 77, 193-200 (2002).

249. Scherer, S. & Hilsberg, T. Hybridisierung und Verwandtschaftsgrade innerhalb der Anatidae - eine systematische und evolutionstheoretische Betrachtung [in German with English summary]. J. Ornith. 123, 357-380 (1982).

250. Mallet, J. A species definition for the modern synthesis. Trends Ecol. Evol. 10, 294-299 (1995).

251. Dobzhansky, T. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics 21, 113-135 (1936).

252. Muller, H. J. Isolating mechanisms, evolution and temperature. In *Temperature, Evolution, Development* (ed. Dobzhansky, T.) p. 71-125 (Jaques Cattell Press, Lancaster, PA, 1942).

253. Prager, E. M. & Wilson, A. C. Slow evolutionary loss of the potential for interspecific hybridization in birds: a manifestation of slow regulatory evolution. Proc. Natl. Acad. Sci. U.S.A. 72, 200-204 (1975).

254. Grant, P. R. & Grant, B. R. Genetics and the origin of bird species. Proc. Natl. Acad. Sci. U.S.A. 94, 7768-7775 (1997).

255. Michelizzi, V. N. *et al.* A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to Water Buffalo. Int. J. Biol. Sci. 7, 18-27 (2011).

256. Miller, J. M., Poissant, J., Kijas, J. W. & Coltman, D. W. A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. Mol. Ecol. Res. 11, 314 (2011).

257. Charlesworth, B., Bartolomé, C. & Noël, V. The detection of shared and ancestral polymorphisms. Genet. Res. 86, 149-157 (2005).

258. Ramos-Onsins, S. E., Stranger, B. E., Mitchell-Olds, T. & Aguadé, M. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166, 373-388 (2004).

259. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, MA, USA, 2007).

260. Clark, A. G. Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. U.S.A. 94, 7730-7734 (1997).

261. Brodin, A. & Haas, F. Speciation by perception. Anim. Behav. 72, 139-146 (2006).

262. Immelmann, K. Ecological significance of imprinting and early learning. Annu. Rev. Ecol. Syst. 6, 15-37 (1975).

263. Zachos, J., Pagani, H., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292, 686-693 (2001).

264. Böhme, M., Ilg, A. & Winklhofer, M. Late Miocene "washhouse" climate in Europe. Earth Planet Sc. Lett. 275, 393-401 (2008).

265. Harzhauser, M. & Piller, W. E. Benchmark data of a changing sea - Palaeogeography, Palaeobiogeography and events in the Central Paratethys during the Miocene. Palaeogeogr. Palaeocl. 253, 8-31 (2007).

266. Cerling, T. E. *et al.* Global vegetation change through the Miocene/Pliocene boundary. Nature 389, 153 (1997).

267. Worthy, T. H. Pliocene waterfowl (Aves:Anseriformes) from South Australia and a new genus and species. Emu 108, 153-165 (2008).

268. Mlíkovský, J. *Cenozoic birds of the world. Part 1: Europe* (Ninox Press, Prague, Czech Republic, 2002).

269. Seiger, M. B. A computer simulation study of influence of imprinting on population structure. Am. Nat. 101, 47-57 (1967).

270. Schutz, F. Objektfixierung geschlechtlicher Reaktionen bei Anatiden und Hühnern [in German]. Naturwissenschaften 50, 624-625 (1963).

271. Excoffier, L. & Slatkin, M. Incorporating genotypes of relatives into a test of linkage disequilibrium. Am. J. Hum. Genet. 62, 171-180 (1998).

272. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12, 921-927 (1995).

273. Birney, E. *et al.* An overview of Ensembl. Genome Res. 14, 925-928 (2004).

274. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004).

275. Yates, F. Contingency table involving small numbers and the χ^2 test. J. Roy. Statistical Society (Supplement) 1, 217-235 (1934).

276. Kimura, M. & Ohta, T. Average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763-771 (1969).

277. Kimura, M. & Ohta, T. Average number of generations until extinction of an individual mutant gene in a finite population. Genetics 63, 701-709 (1969).

278. BirdLife International. Species factsheets. Downloaded from http://www.birdlife.org on 20/10/2010. (2010).

279. Frankham, R. Effective population size/adult population size ratios in wildlife: A review. Genet. Res. 66, 95-107 (1995).

280. Tomlinson, C., Mace, G. M., Black, J. M. & Hewston, N. Improving the management of a highly inbred species: the case of the white-winged wood duck *Cairina scutulata* in captivity. Wildfowl 42, 123-133 (1991).

281. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904-909 (2006).

282. Nudds, T. D. Niche dynamics and organization of waterfowl guilds in variable environments. Ecology 64, 319-330 (1983).

283. Patterson, J. H. Can ducks be managed by regulation? Experiences in Canada. Trans. North Am. Wildl. and Nat. Resour. Conf. 44, 130-139 (1979).

284. Whitlock, M. C. & Barton, N. H. The effective size of a subdivided population. Genetics 146, 427-441 (1997).

285. Gao, H., Williamson, S. & Bustamante, C. D. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176, 1635-1651 (2007).

286. Mayr, E. *Systematics and the origin of Species* (Columbia University Press, New York, 1942).

287. Mayr, E. Birds collected during the Whitney South Sea Expedition. XII. Notes on *Halcyon chloris* and some of its subspecies. Am. Mus. Novit. 469, 1-10 (1931).

288. Rensch, B. Grenzfälle von Rasse und Art. Journ. f. Ornith. 76, 222-231 (1928).

289. Urban, E. K., Fry, C. H. & Keith, S. Introduction to "Birds of Africa". In *Birds of Africa, Vol. II* (eds. Urban, E. K., Fry, C. H. & Keith, S.) p. xi-xvi (Academic Press, Inc., London, 1986).

290. Amadon, D. The superspecies concept. Syst. Zool. 15, 245-249 (1966).

291. Haffer, J. Superspecies and species limits in vertebrates. Z. Zool. Syst. Evol. 24, 169-190 (1986).

292. Kiriakoff, S. G. On the nomenclature of the superspecies. Syst. Zool. 16, 281-282 (1967).

293. Mayr, E. & Short, L. L. *Species Taxa of North-American Birds. A Contribution to Comparative Systematics.* (Publications of the Nuttall Ornithological Club, No 9, Cambridge, Mass., 1970).

294. Short, L. L. Taxonomic aspects of avian hybridization. Auk 86, 84-105 (1969).

295. Amadon, D. & Short, L. L. Treatment of subspecies approaching species status. Syst. Zool. 25, 161-167 (1976).

296. Dubois, A. New proposals for naming lower-ranked taxa within the frame of the *International Code of Zoological Nomenclature*. C. R. Biol. 329, 823-840 (2006).

297. Manegold, A. & Brink, J. S. Descriptions and palaeoecological implications of bird remains from the Middle Pleistocene of Florisbad, South Africa. Paläontol. Z. 85, 19-32 (2010).

298. Olson, S. L. The identity of the fossil ducks described from Australia by C.W. De Vis. Emu 77, 127-131 (1977).

299. Worthy, T. H. Descriptions and phylogenetic relationships of two new genera and four new species of Oligo-Miocene waterfowl (Aves: Anatidae) from Australia. Zool. J. Linn. Soc. 156, 411-454 (2009).

300. Seehausen, O., Takimoto, G., Roy, D. & Jokela, J. Speciation reversal and biodiversity dynamics with hybridization in changing environments. Mol. Ecol. 17, 30-44 (2008).

301. Shiina, T. *et al.* Comparative genomic analysis of two avian (quail and chicken) MHC regions. J. Immunol. 172, 6751-6763 (2004).

302. Burt, D. W. Chicken genome: Current status and future opportunities. Genome Res. 15, 1692-1698 (2005).

303. Serratosa, J., Ribó, O., Correia, S. & Pittman, M. EFSA scientific risk assessment on animal health and welfare aspects of avian influenza (EFSA-Q-2004-075). Avian Dis. 51, 501-503 (2007).

304. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. Nucleic Acids Res. 38, D46-D51 (2009).

305. Kida, H., Yanagawa, R. & Matsuoka, Y. Duck influenza lacking evidence of disease signs and immune response. Infect. Immun. 30, 547-553 (1980).

306. van Gils, J. A. *et al.* Hampered foraging and migratory performance in swans infected with low-pathogenic avian influenza A virus. PLoS ONE 2, e184 (2007).

307. Latorre-Margalef, N. *et al.* Effects of influenza A virus infection on migrating mallard ducks. Proc. R. Soc. B. 276, 1029-1036 (2009).

308. Feare, C. J. & Yasué, M. Asymptomatic infection with highly pathogenic avian influenza H5N1 in wild birds: How sound is the evidence? Virol. J. 3, 96 (2006).

309. Gilbert, M. *et al.* Anatidae migration in the western Palearctic and spread of highly pathogenic avian influenza H5N1 virus. Emerg. Infect. Dis. 12, 1650-1656 (2006).

310. Wallensten, A. *et al.* High prevalence of influenza A virus in ducks caught during spring migration through Sweden. Vaccine 24, 6734-6735 (2006).

311. Munster, V. J. *et al*. Spatial, temporal, and species variation in prevalence of influenza a viruses in wild migratory birds. PLoS Pathog. 3, e61 (2007).

312. Weber, T. P. & Stilianakis, N. I. Ecologic immunology of avian influenza (H5N1) in migratory birds. Emerg. Infect. Dis. 13, 1139-1143 (2007).

313. Cromie, R. L., Lee, R. & Baz, H. Avian influenza: A short review of the disease in wild birds, and of European wild bird surveillance during winter 2005/06. Wildfowl 56, 197-202 (2006).

314. Munster, V. J. *et al*. Practical considerations for high-throughput influenza A virus surveillance studies of wild birds by use of molecular diagnostic tests. J. Clin. Microbiol. 47, 666-673 (2009).

315. Li, C. C., Beck, I. A., Seidel, K. D. & Frenkel, L. M. Persistence of human immunodeficiency virus type 1 subtype B DNA in dried-blood samples on FTA filter paper. J. Clin. Microbiol. 42, 3847-3849 (2004).

316. Moscoso, H., Raybon, E. O., Thayer, S. G. & Hofacre, C. L. Molecular detection and serotyping of infectious bronchitis virus from FTA® filter paper. Avian Dis. 49, 24-29 (2005).

317. Ndunguru, J. *et al*. Application of FTA technology for sampling, recovery and molecular characterization of viral pathogens and virus-derived transgenes from plant tissues. Virol. J. 2, 45 (2005).

318. Perozo, F., Villegas, P., Estevez, C., Alvarado, I. & Purvis, L. B. Use of FTA® filter paper for the molecular detection of Newcastle disease virus. Avian Pathol. 35, 93-98 (2006).

319. Purvis, L. B., Villegas, P. & Perozo, F. Evaluation of FTA® paper and phenol for storage, extraction and molecular characterization of infectious bursal disease virus. J. Virol. Methods 138, 66-69 (2006).

320. Inoue, R., Tsukahara, T., Sunaba, C., Itoh, M. & Ushida, K. Simple and rapid detection of the porcine reproductive and respiratory syndrome virus from pig whole blood using filter paper. J. Virol. Methods 141, 102-106 (2007).

321. Nuchprayoon, S., Saksirisampant, W., Jaijakul, S. & Nuchprayoon, I. Flinders Technology Associates (FTA) filter paper-based DNA extraction with Polymerase Chain Reaction (PCR) for detection of *Pneumocystis jirovecii* from respiratory specimens of immunocompromised patients. J. Clin. Lab. Anal. 21, 382-386 (2007).

322. Picard-Meyer, E., Barrat, J. & Cliquet, F. Use of filter paper (FTA®) technology for sampling, recovery and molecular characterisation of rabies viruses. J. Virol. Methods 140, 174-182 (2007).

323. Muthukrishnan, M., Singanallur, N. B., Ralla, K. & Villuppanoor, S. A. Evaluation of FTA® cards as a laboratory and field sampling device for the detection of foot-and-mouth disease virus and serotyping by RT-PCR and real-time RT-PCR. J. Virol. Methods 151, 311-316 (2008).

324. Rogers, C. D. G. & Burgoyne, L. Bacterial typing: Storing and processing of stabilized reference bacteria for polymerase chain reaction without preparing DNA - An example of an automatable procedure. Anal. Biochem. 247, 223-227 (1997).

325. Wallensten, A. *et al.* Surveillance of influenza A virus in migratory waterfowl in northern Europe. Emerg. Infect. Dis. 13, 404-411 (2007).

326. Burgoyne, L. A. Solid medium and method for DNA storage. U.S. patent 5,496,562. (1996).

327. Fouchier, R. A. M. *et al.* Detection of influenza a viruses from different species by PCR amplification of conserved sequences in the matrix gene. J. Clin. Microbiol. 38, 4096-4101 (2000).

328. Spackman, E. *et al.* Development of a real-time reverse transcriptase PCR assay for type A influenza virus and the avian H5 and H7 hemagglutinin subtypes. J. Clin. Microbiol. 40, 3256-3260 (2002).

329. Wang, R. *et al.* Examining the hemagglutinin subtype diversity among wild duck-origin influenza A viruses using ethanol-fixed cloacal swabs and a novel RT-PCR method. Virology 375, 182-189 (2008).

330. Rogers, C. D. G. & Burgoyne, L. A. Reverse transcription of an RNA genome from databasing paper (FTA®). Biotechnol. Appl. Biochem. 31, 219-224 (2000).

331. Runstadler, J. A. *et al.* Using RRT-PCR analysis and virus isolation to determine the prevalence of avian influenza virus infections in ducks at Minto Flats State Game Refuge, Alaska, during August 2005. Arch. Virol. 152, 1901-1910 (2007).

332. Pulido, F. The genetics and evolution of avian migration. BioScience 57, 165-174 (2007).

333. Taubenberger, J. K. The origin and virulence of the 1918 "Spanish" influenza virus. Proc. Am. Philos. Soc. 150, 86-112 (2006).

334. Chen, H. *et al.* Properties and dissemination of H5N1 viruses isolated during an influenza outbreak in migratory waterfowl in Western China. J. Virol. 80, 5976-5983 (2006).

335. Gaidet, N. *et al.* Avian influenza viruses in water birds, Africa. Emerg. Infect. Dis. 13, 626-629 (2007).

336. Krauss, S. *et al.* Influenza in migratory birds and evidence of limited intercontinental virus exchange. PLoS Pathog. 3, e167 (2007).

337. Parmley, E. J. *et al.* Wild bird influenza survey, Canada, 2005. Emerg. Infect. Dis. 14, 84-87 (2008).

338. Evers, D. L., Slemons, R. D. & Taubenberger, J. K. Effect of preservative on recoverable RT-PCR amplicon length from influenza A virus in bird feces. Avian Dis. 51, 965-968 (2007).

339. Forster, J. L., Harkin, V. B., Graham, D. A. & McCullough, S. J. The effect of sample type, temperature and RNAlater™ on the stability of avian influenza virus RNA. J. Virol. Methods 149, 190-194 (2008).

340. Lack, D. *Evolution Illustrated by Waterfowl* (Blackwell Scientific Publishing, Oxford, 1974).

341. Rhymer, J. M. & Simberloff, D. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27, 83-109 (1996).

342. MacDonald, M. R. W. *et al.* Genomics of antiviral defenses in the duck, a natural host of influenza and hepatitis B viruses. Cytogenet. Genome Res. 117, 195-206 (2007).

343. Magor, K. E. Immunoglobulin genetics and antibody responses to influenza in ducks. Dev. Comp. Immunol. 35, 1008-1016 (2011).

344. Barber, M. R. W., Aldridge Jr, J. R., Webster, R. G. & Magor, K. E. Association of RIG-I with innate immunity of ducks to influenza. Proc. Natl. Acad. Sci. U.S.A. 107, 5913-5918 (2010).

345. Huang, Y., Li, N., Burt, D. W. & Wu, F. Genomic research and applications in the duck (*Anas platyrhynchos*). Worlds Poult. Sci. J. 64, 329-341 (2008).

346. Rhodes Jr., O. E., Smith, L. M. & Chesser, R. K. Temporal components of genetic variation in migrating and wintering American Wigeon. Can. J. Zool. 71, 2229-2235 (1993).

347. Rhymer, J. M., Murray, J. T. & Braun, M. J. Mitochondrial analysis of gene flow between New Zealand Mallards (*Anas platyrhynchos*) and Grey Ducks (*A. superciliosa*). Auk 111, 970-978 (1994).

348. Rhymer, J. M., Williams, M. J. & Kingsford, R. T. Implications of phylogeography and population genetics for subspecies taxonomy of Grey (Pacific Black) Duck *Anas superciliosa* and its conservation in New Zealand. Pac. Conserv. Biol. 10, 57-66 (2004).

349. Browne, R. A., Griffin, C. R., Chang, P. R., Hubley, M. & Martin, A. E. Genetic divergence among populations of the Hawaiian Duck, Laysan Duck, and Mallard. Auk 110, 49-56 (1993).

350. Livezey, B. C. A phylogenetic classification of waterfowl (Aves: Anseriformes), including selected fossil species. Ann. Carnegie Mus. 66, 457-496 (1997).

351. Kulikova, I. V., Chelomina, G. N. & Zhuravlev, Y. N. Low genetic differentiation of and close evolutionary relationships between *Anas platyrhynchos* and *Anas poecilorhyncha*: RAPD-PCR evidence. Russ. J. Genet. 39, 1143-1151 (2003).

352. Kulikova, I. V. & Zhuravlev, Y. N. Genetic structure of the Far Eastern population of Eurasian wigeon *Anas penelope* inferred from sequencing of the mitochondrial DNA control region. Russ. J. Genet. 46, 976-981 (2010).

353. Kreisinger, J., Munclinger, P., Jav rková, V. & Albrecht, T. Analysis of extra-pair paternity and conspecific brood parasitism in mallards *Anas platyrhynchos* using non-invasive techniques. J. Avian Biol. 41, 551-557 (2010).

354. Hoeh, W. R., Blakley, K. H. & Brown, W. M. Heteroplasmy suggests limited biparental inheritance of *Mytilus* mitochondrial DNA. Science 251, 1488-1490 (1991).

355. Meusel, M. S. & Moritz, R. F. A. Transfer of paternal mitochondrial DNA during fertilization of honeybee (*Apis mellifera* L.) eggs. Curr. Genet. 24, 539-543 (1993).

356. Gyllensten, U., Wharton, D. & Wilson, A. C. Maternal inheritance of mitochondrial DNA during backcrossing of two species of mice. J. Hered. 76, 321-324 (1985).

357. Schwartz, M. & Vissing, J. Paternal inheritance of mitochondrial DNA. New Eng. J. Med. 347, 576-580 (2002).

358. Wolff, J. N. & Gemmell, N. J. Lost in the zygote: The dilution of paternal mtDNA upon fertilization. Heredity 101, 429-434 (2008).

359. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729-744 (2004).

360. Maak, S., Neumann, K., Von Lengerken, G. & Gattermann, R. First seven microsatellites developed for the Peking duck (*Anas platyrhynchos*). Anim. Genet. 31, 233 (2000).

361. Maak, S., Wimmers, K., Weigend, S. & Neumann, K. Isolation and characterization of 18 microsatellites in the Peking duck (*Anas platyrhynchos*) and their application in other waterfowl species. Mol. Ecol. Notes 3, 224-227 (2003).

362. Huang, Y. *et al.* Characterization of 35 novel microsatellite DNA markers from the duck (*Anas platyrhynchos*) genome and cross-amplification in other birds. Genet Sel Evol 37, 455-472 (2005).

363. May, R. M. Network structure and the biology of populations. Trends Ecol. Evol. 21, 394-399 (2006).

364. Matthews, L. & Woolhouse, M. New approaches to quantifying the spread of infection. Nat. Rev. Microbiol. 3, 529-536 (2005).

365. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. Nature 393, 440-442 (1998).

366. Taubenberger, J. K., Hultin, J. V. & Morens, D. M. Discovery and characterization of the 1918 pandemic influenza virus in historical context. Antivir. Ther. 12, 581-491 (2007).

367. Taubenberger, J. K. & Reid, A. H. The 1918 'Spanish' influenza pandemic and characterization of the virus that caused it. Perspect. Med. Virol. 7, 101-122 (2002).

368. Caron, A., Gaidet, N., de Garine-Wichatitsky, M., Morand, S. & Cameron, E. Z. Evolutionary biology, community ecology and avian influenza research. Infect. Genet. Evol. 9, 298-303 (2009).

369. Stallknecht, D. E. & Shane, S. M. Host range of avian influenza virus in free-living birds. Vet. Res. Commun. 12, 125-141 (1988).

370. Kawaoka, Y., Chambers, T. M., Sladen, W. L. & Webster, R. G. Is the gene pool of influenza viruses in shorebirds and gulls different from that in wild ducks? Virology 163, 247-250 (1988).

371. Slemons, R. D. & Easterday, B. C. Virus replication in the digestive tract of ducks exposed by aerosol to type-A influenza. Avian Dis. 22, 367-377 (1978).

372. Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. 22, 1561-1568 (2005).

373. Geraldes, A. *et al.* Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. 17, 5349-5363 (2008).

374. del Hoyo, J., Elliot, A. & Christie, D. *Handbook of the Birds of the World* (Lynx Edicions, Barcelona, Spain, 2006).

375. Sekercioglu, C. H. Foreword. In *Handbook of the Birds of the World, Volume 11: Old World Flycatchers to Old World Warblers* (eds. del Hoyo, J., Elliot, A. & Christie, D.) p. 48 (Lynx Edicions, Barcelona, Spain, 2006).

376. Dallimer, M., Jones, P. J., Pemberton, J. M. & Cheke, R. A. Lack of genetic and plumage differentiation in the red-billed quelea *Quelea quelea* across a migratory divide in southern Africa. Mol. Ecol. 12, 345-353 (2003).

377. Dugan, V. G. *et al*. The evolutionary genetics and emergence of avian influenza viruses in wild birds. PLoS Pathog. 4, e1000076 (2008).

378. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227-R240 (2010).

379. Tautz, D., Ellegren, H. & Weigel, D. Next generation molecular ecology. Mol. Ecol. 19, 1-3 (2010).

380. Kingman, J. F. C. Origins of the coalescent: 1974-1982. Genetics 156, 1461-1463 (2000).

381. Cohen, J. Bioinformatics - An introduction for computer scientists. ACM Comput. Surv. 36, 122-158 (2004).

382. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142-149 (2008).

383. Moore, G. Cramming more components onto integrated circuits. Electronics 38, 114-117 (1965).

384. Post, D. The future of computing performance. Comp. Sci. Eng. 13, 4-5 (2011).

385. Artois, M. *et al*. Outbreaks of highly pathogenic avian influenza in Europe: The risks associated with wild birds. OIE Rev. Sci. Tech. 28, 69-92 (2009).

Summary

Birds, in particular poultry and ducks, are a source of many infectious diseases, such as those caused by influenza viruses. These viruses are a threat not only to the birds themselves but also to poultry farming and human health, as forms that can infect humans are known to have evolved. It is believed that migratory birds in general play an important role in the global spread of avian influenza (AI). However, it is still debated how large this role precisely is and whether other modes of spread may be more important. The mallard (*Anas platyrhynchos*) is the world's most abundant and well-studied waterfowl species. Besides being an important game and agricultural species, it is also a flagship species in wetland conservation and restoration. Waterfowl (Anseriformes: Anatidae) and especially ducks currently are the focal bird group in long distance dispersal of Avian Influenza in the wild, and the mallard has been identified as the most likely species to transport this virus.

In my thesis I report aspects of the biology of this important host species of AI by molecular ecological means. As molecular marker system I established a genome-wide set of more than 100,000 SNPs of which I developed a subset of 384 SNPs into an assay to genotype about 1,000 ducks. This subset was employed to study the evolutionary history and speciation processes in the *Anas* genus. Further investigations into the world-wide mallard population structure on a species level were based not only on this set of 384 SNPs but also on mitochondrial DNA sequences. Last but not last, I investigated an option of AI sampling and detection from duck faeces by technology that is safe from a biohazard perspective, and solves transportation issues related to cold chains.

The main results of my thesis include the development of a generally applicable improved analysis pipeline to develop genome-wide SNP sets for non-model organisms. Further, my results show that, from a migration system perspective, mallard flyways/populations can hardly be delineated from a biological point of view. Detailed phylogenetic, population genetic and coalescent analyses of a data set of samples spanning the whole northern hemisphere leads me to conclude that the only firm population boundaries that I can draw are between Eurasia and North America, within which panmixia is almost achieved. Mallards' and other *Anas*-ducks' whole continental to global distribution brings them together in sympatry. I can show that a combination of sympatric distribution, conflicting genetically determined and learned mate recognition mechanisms, and genomic compatibility between species helps to explain the long-standing puzzle of waterfowl hybridisation and introgression of genes from one duck species into another. Besides obvious management implications I propose that this fact can be part of the explanation why ducks are so well adaptable and successful, as well as why they show extraordinary abilities to withstand AI infections, or its consequences for health status.

Samenvatting

Vogels, met name pluimvee en eenden, zijn een bron van vele infectieziekten, zoals de ziekten die veroorzaakt worden door influenzavirussen. Deze virussen vormen niet alleen een bedreiging voor de vogels zelf, maar ook voor de pluimvee-industrie en publieke gezondheid, aangezien het bekend is dat er varianten zijn geëvolueerd die mensen kunnen infecteren. Trekvogels worden in het algemeen verondersteld een belangrijke rol te vervullen bij de verspreiding van vogelgriep (VG). Echter, hoe groot die rol precies is en of andere manieren van verspreiding belangrijker zouden kunnen zijn, is onderwerp van discussie. De wilde eend (*Anas platyrhynchos*) is de meest voorkomende en de best bestudeerde watervogel ter wereld. Afgezien van het belang van deze soort in de jacht en de landbouw, is het ook een indicatorsoort voor de bescherming en het herstel van wetlands. Watervogels (Anseriformes: Anatidae) en vooral eenden zijn op dit moment de belangrijkste focale groep voor de lange-afstandsverspreiding van vogelgriep in het wild, en de wilde eend is geïdentificeerd als de meest waarschijnlijke soort voor transport van dit virus.

In mijn proefschrift beschrijf ik aspecten van de biologie van deze belangrijke gastheersoort van VG met behulp van moleculair ecologische methoden. Als moleculair merkersysteem heb ik een genoom-brede set van meer dan 100,000 SNPs vastgesteld, waarvan ik een subset van 384 SNPs heb ontwikkeld tot een assay voor het genotyperen van ongeveer 1,000 eenden. Deze subset is gebruikt om de evolutionaire geschiedenis en het soortvormingsproces van het genus Anas te bestuderen. Verder onderzoek naar de wereldwijde populatiestructuur van de wilde eend op soortniveau, waren niet alleen gebaseerd op deze 384 SNP set, maar ook op mitochondriale DNA sequenties. Ten slotte heb ik de optie onderzocht om het bemonsteren en detecteren van VG in feces van eenden uit te voeren met behulp van een technologie die veilig is vanuit biohazard oogpunt, en die transportproblemen oplost met betrekking tot koude ketens.

De hoofdresultaten van mijn proefschrift omvatten de ontwikkeling van een algemeen toepasbare en verbeterde pipeline voor de ontwikkeling van genoom-brede SNP sets voor niet-model organismen. Verder geven mijn resultaten aan dat, vanuit het perspectief van een migratie-systeem, de vliegroutes/populaties van wilde eenden in feite amper onderscheiden kunnen worden vanuit een biologisch oogpunt. De gedetailleerde fylogenetische, populatiegenetische en coalescent-analyses van een dataset van monsters die het volledige noordelijk halfrond omvatten, leiden mij tot de conclusie dat the enige steekhoudende populatiegrenzen die ik kan trekken zich bevinden tussen Eurazië en Noord-Amerika, waarbinnen bijna panmixia wordt bereikt. De continentale tot wereldwijde verspreiding van de wilde eend en andere Anas-eenden brengt deze samen in sympatrie. Ik kan aantonen dat een combinatie van sympatrische distributie, conflicterende genetische bepaalde en aangeleerde partner-herkenningsmechanismen, en genomische compatibiliteit tussen soorten, helpt bij het begrijpen van de lang standhoudende puzzel van watervogelhybridisatie en introgressie van genen van de ene eendensoort naar de ander. Naast de voor de hand liggende implicaties voor beheer, stel ik voor dat dit feit onderdeel kan zijn van de verklaring waarom eenden zo goed aangepast en succesvol zijn, alsook waarom zij de buitengewone eigenschappen vertonen om VG infecties, dan wel de gevolgen voor hun staat van gezondheid, te weerstaan.

Acknowledgements

The process of assembling this thesis was supported by so many people. First and foremost I would like to thank my promotors Herbert Prins and Ron Ydenberg. Herbert, already in my job interview I had the feeling that working with you would turn out to become a great experience. Besides all the usual skills that a chair of a research group most often has, such as experience, expertise, overview, etc., I recognised one set of skills that I think makes quite a difference: enthusiasm, joy, and helpfulness. Besides the usual things that PhD students do during their working day and experience during discussions with supervisors, the most important activities are those that lead to happiness in one's project. I always experienced you as a person with those special skills, leading to an extraordinarily nice working atmosphere! Additionally, I think what facilitated the success of this thesis in the past 4.5 years was the amount of trust you put into me, leaving me so much freedom that it was sometimes difficult to believe! Thanks a million for enabling me to carry out my project in the way it went. I realise that very often rules, regulations, but also traditions were bent almost until they broke. And as very often "shit happens", there is always sunshine after the rain. Ron, we didn't get to talk to each other that much. But when we had the chance I was impressed by the speed with which you absorbed my topics even though it appeared to be barely related to anything you did before. As far as I can see you have never been involved in a molecular ecology project until now (to the extent of appearing as author on a resulting paper). The details of my analysis were just loosely connected to your expertise, yet you sharply grasped what was exactly relevant, and could return valuable feedback. Thanks for this!

In a supervisory team there are not only the professors. The role of the daily supervisor is an important one, too! Pim van Hooft has the burden of being the only molecular ecology oriented staff member of REG. Pim, while performing a molecular ecological project - especially with next generation sequencing and SNPs, among the newest technologies on the scene - we often found ourselves in the interesting position of being a bit on our own. This was particularly challenging and required a lot of thinking outside the box. I admire you for your ability to look a t things in different and fresh angles, and found this very important in a setting in which there is no clear road paved yet. It often lead to disagreement, but never to misgivings. You showed your respect for me and my work in all situations and mastered the balancing act between being professional and informal. Especially towards the final bits the speed with which you returned comments was outstanding, I think. Thanks for all your support!

Of course, daily work was carried out in the whole research group. I enjoyed the diverse nature of this assemblage of people from so many different nationalities, and with so many backgrounds. I'd like to thank the lecturers for organising the core business of the group so well: Arend, Frank, Fred, Ignas, Milena, Pim, Sip. Keeping together the flesh and bones of the Resource Ecology Group are the administrative staff: Gerda, Patricia, many thanks for helping me with all my stupid questions and requests! Herman, thanks for doing a great job in keeping my website alive and reacting so quickly to all my request, and for checking the Dutch translation of my summary section. That was remarkable! Now, let's get to the actual bulk of the Resource Ecology Group: The PhD students and postdocs! This is quite a crowd, and even though I did not have the chance to get in touch with all of you, I'd like to express my gratitude to have met you, and have learned

from your experience in life: Alfred, Ania, Anil, Anna, Audrie, Bas, Benson, Christian, Cornelis, Daniel, Dulce, Edson, Eduardo, Edward, Emmanuel, Farshid, Geerten, Henjo, Jasja, Jasper, Jia, Kyle, Lennart, Mariaan, Nicole, Nikki, Ntuthuko, Qiong, Priya, Ralf, Rudy, Sisi, Tessema, Thomas, Tom vdH, Tom H, Tsewang, Vincent, Xavier, Yolanda, Yong, Zheng. In this group, special thanks go to Mariaan and Daniel. Mariaan, I am really happy to have met you, and enjoyed the short time that we spent together as room mates. Your plant is still alive (even now, I am sure)! It is a pity that life goes on as it does, and we hardly have the opportunity to meet. Daniel, as my second room mate you also deserve special mentioning! I enjoyed your presence and the calm atmosphere you created – this is by no means meant in a negative way! Keep the duck computer running! Thanks, too, for translating my summary into Dutch. I'd also like to pick out Rudy from that list: I think we had a great time in the last few years, and I discussed more about science, society, and work life with you than with any other colleague (except Herbert, maybe). Thanks for these discussions, they were and are still very valuable. Thanks, too, for daring to engage into our genetics project! I can't wait to see how this part of our collaboration eventually works out, when all related papers are published! Further, Daniel and Rudy, thanks for being my paranymphs! Kyle, thanks for reading and commenting upon my introduction and discussion sections in this thesis. Last but not least I'd like to mention my master students and internship students: Anne, Jacintha, Sjoerd. Anne, thanks for your hard work in creating our mtDNA data set, and for being such a motivated student with so much drive and enthusiasm. Keep up the momentum! Jacintha, Sjoerd, what started out as an interesting little idea eventually culminated in a paper! Thanks a lot for enduring the long publishing process, especially of that story! By now the three of you are PhD students yourself, I am happy to see you progressing in the scientific arena. Let's keep in touch!

My project was carried out with help of the Animal Breeding and Genomics (ABG) Group in Wageningen. I spent a lot of time there and experienced amazing support! Jan van der Poel was the first person at ABG to help me find my way. Martien Groenen and Richard Crooijmans helped understand and organise many aspects of both lab work and bioinformatics. My thanks go to the members of the Genomics Discussion Group which I attended: Andreia, Bert, Hendrik-Jan, Hinri, John, Marcos, Nikkie, Ole, Xinning, Zhao Zhen. Our joint discussions about next generation sequencing, genomics, and bioinformatics were extremely valuable to me! Special thanks go to my closest collaborator Hinri, who worked with me on my data with admirable patience. Further I wish to mention that I enjoyed inspiring discussions on life and science (and the interface thereof) with Hendrik-Jan. I am very happy to have worked with you and enjoy the ongoing finishing of bits and pieces. Surely we'll stay in touch, too. Classical lab work was facilitated by great support from the technical staff of ABG: Bert, Jan, Rosilde, Sylvia, Tineke.

One of the great adventures besides my PhD project and the numerous side projects was the creation of WEES, the Wageningen Ecology and Evolution Seminars. Giving birth to, and organising this special series, really thrilled me at times, and I enjoyed so much being part of the team. Thanks to the organising committee: Ansa, Bart, Daan, Daniel, Detmer, Erik, Hanneke, Mark, Mieke, Mirte, Ralf, Robin, Rudy, Stineke.

Projects like mine very much depend on the numerous contacts who kindly provided me with samples of mallards: A.A. Samoilov, A.N. Orlov, A.V. Karelov, A.Y. Volkov, Aaron Everingham,

Aili Lage Labansen, Alf Tore Mjøs, Alyn Walsh, Andrew Iwaniuk, Andy Richardson, Anna Palmé, Anne Zeddeman, Antti Paasivaara, Apostolos Tsiompanoudis, Arseny Tsvey, Bjorn Birgisson, Brandt Meixell, Cameron Manson, Charles Bull, D.O. Klymysmyn, Danielle Mondloch, Darren Hasson, David Lamble, David Rodrigues, David Schonberg Alm, Dmitry A. Sartakov, Dominic Berridge, Ernst Niedermayer, G.V. Gonokhin, Gábor Czirják, Garnet Baker, Garry Grigg, Hans Geisler, Hans Jörg Damm, Henry J. Bruhlman, Herman Postma, Holly Middleton, I.V. Shydlovkyy, Jan Bokdam, Jens Kjeld Jensen, Jonas Waldenström, Jonathan Rundstadler, Jordi Figuerola, Julius Morkunas, Karin Geisler, M.V. Gavrilickev, Mathieu Boos, Matthieu Guillemain, Mitja Kersnik, Muhammad Hashim, N. D. Poyarkov, Neus Latorre-Margalef, Nicolaos Kassinis, O. Tutenkov, O.V. Koshyn, Ricardas Patapavicius, Ruth Cromie, Sasan Fereidouni, Sergei Fokin, Severin Wejbora, Shah Nawaz Khan, Steven Evans, Thomas Kondratowicz, Tom Fiske, Tróndur Leivsson, Urmas Võro, V.I. Zalogin, V.N. Stepanov, V.S. Galtsov, Valery Buzun, Y. Konstantinov, Yan-Ling Son.

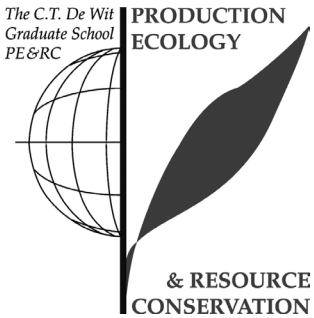
Further, due to my interest in other duck species as well, I'd like to thank the following people and institutions for providing me with samples of other duck species: Beth Rich (Tautphaus Park Zoo), Brandt Meixell (University of Alaska Fairbanks), Crystal Matthews (Virginia Aquarium & Marine Science Center), Danielle Mondloch (University of Alaska Fairbanks), David Gomis (Parc Zoologique et Botanique de Mulhouse), Dirk Ullrich (Alpenzoo Innsbruck), Dorothee Ordonneau (Parc Zoologique de Lille), Iñigo Sánchez (ZooBotánico de Jerez), Javier Gonzalez (University of Heidelberg), Jonathan Rundstadler (University of Alaska Fairbanks), Kamil ihák (Zoo Dvur Kralove), Magnus Hellstöm (Ottenby Bird Observatory), Marina Euler (Tierpark Lange Erlen), Martin Straube (Zoo Krefeld), Mathieu Boos (CNRS Strasbourg, France), Michael Wink (University of Heidelberg), Michelle O'Brian (Wildfowl & Wetlands Trust), Robert Zingg (Zoo Zürich), Roger Sweeney (Tracy Aviary), Rolik Grzegorz (Zoo Opole), Sascha Knauf (Opel Zoo), Sergey Fokin (State Informational-Analytical Centre of Game Animal and Environment of Hunting Department of Russia), Timm Spretke (Zoologischer Garten Halle), Valery Buzun, Yang Liu (University of Bern, Switzerland), Yannik Roman (Le parc de Clères). Unfortunately, for various reasons, I could not use all of the mallard and other duck samples that I obtained over the years. Thanks a million anyway for your effort and willingness to share your precious material with me. It is people like you, who facilitate scientific progress more than is often acknowledged!

Besides scientific support, I wish to thank my family who had a difficult time due to my engagement with science. Anika, thank you for your support, and Anna and Nele as well, for keeping me away from work when it was good to keep some distance to it. I often realised what I was working for when thinking of my wife and kids. I'd also like to thank Sigrun and Harald, my parents, for giving me the feeling of full support in all my actions. I also enjoyed the (often too rare occasions of) staying with the rest of my family in Germany. My apologies to all of you for not being around as much as I should have been. I never meant (and mean) no harm with not staying in touch as much as I should. The same counts for my friends in Germany. Some of them I met regularly through the past few years (mainly the Assenheimers), others I met rarely (former school and sports mates). I wouldn't do any justice to you by listing all your names here because the list would naturally be incomplete; there is no clear boundary between close friends, friends, loose friends, acquaintances, and some such – and eventually things are in flux anyway. It was just

nice to have you around every now and then, having someone else than colleagues (who know what I work on anyway) and family (who have suffered enough from my explanations about what I do) to annoy with my presence.

Finally, I wish to express my apologies to all who I might have forgotten. It is well possible that your name appears in one of the specific acknowledgements sections of the individual chapters.

PE&RC PhD Education Certificate



With the educational activities listed below the PhD candidate has complied with the educational requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

REVIEW OF LITERATURE (5.6 ECTS)

– Host migration and influenza dispersal in the mallard (*Anas platyrhynchos*) as assessed by population genetic means (2007)

WRITING OF PROJECT PROPOSAL (4 ECTS)

– Host migration and influenza dispersal in the mallard (*Anas platyrhynchos*) as assessed by population genetic means (2007)

POST-GRADUATE COURSES (6 ECTS)

– Animal handling course; Utrecht University (2007)
– Introduction to R for statistical analysis; WIAS (2008)
– Advanced course guide to scientific artwork; WUR library (2008)
– Bayesian statistics; PE&RC (2009)
– Hands-on genomic workshop; Bik-F, Frankfurt/Main, Germany (2010)

LABORATORY TRAINING AND WORKING VISITS (1.2 ECTS)

– Handling and sampling mallards; Ottenby Bird Observatory, Sweden (2007)

INVITED REVIEW OF (UNPUBLISHED) JOURNAL (6 ECTS)

– South African Journal of Animal Science: PCR protocol for detecting meat contamination (2009)
– Conservation Genetics resources: cross-species SNP testing (2010)
– Nwo Sanpad Project proposal (not a journal): predator conservation in a National Park, South Africa (2010)
– Endangered Species Research: Mammal management and its impact on evolutionary strategies (2011)
– Molecular Ecology: a bird ‘s population genetic structure (2011)

DEFICIENCY, REFRESH, BRUSH-UP COURSES (4.3 ECTS)

– Self course: linux pocket guide; DJ Barret (2008)
– Self course: beginning Perl for bioinformatics; J Tisdall (2008)
– Basic statistics (2009)

COMPETENCE STRENGTHENING / SKILLS COURSES (1.6 ECTS)

– PhD Competence assessment; WGS (2008)
– Effective behaviour in your professional surroundings; WGS (2008)
– Career assessment; WGS (2008)

PE&RC ANNUAL MEETINGS, SEMINARS AND THE PE&RC WEEKEND (1.5 ECTS)

– PE&RC Days (2007-2010)
– B.V.W. 'Biologica' symposium: from the cradle to the grave; Wageningen, the Netherlands (2008)

DISCUSSION GROUPS / LOCAL SEMINARS / OTHER SCIENTIFIC MEETINGS (14.8 ECTS)

– Dutch bird migration PhD student group meeting (2007)
– Ecological Theories and Applications discussion group (2007-2011)
– Experimental Evolution discussion group (2007-2011)
– Genomics group meetings at the Animal Breeding and genomics Centre; Wageningen (2008-2010)
– Beschermde fauna: schade, preventie en oplossing; Landelijk contact- en demodag (2009)
– Participation in six WEES (Wageningen Ecology and Evolution Seminars); master class (2009-2010)

INTERNATIONAL SYMPOSIA, WORKSHOPS AND CONFERENCES (13.3 ECTS)

– Evolutionary genetics of host-parasite relationships; poster; Roscoff, France (2007)
– Kick-off meeting animal network research group; oral presentation; REG, Wageningen, the Netherlands (2007)
– 2nd Pan-European duck symposium; two posters; Arles, France (2009)
– Frühjahrstagung der Gesellschaft für Genetik: Evolutionary Genetics-the impact of next generation sequencing technologies; oral presentation; Lutherstadt Wittenberg, Germany (2009)
– Duck Genome Meeting; oral presentation; Beijing, China (2010)
– Hands-on Genomics workshop; oral presentation; Bik-F, Frankfurt/Main, Germany (2010)

LECTURING / SUPERVISION OF PRACTICAL 'S / TUTORIALS (4.5 ECTS)

– Population genetics; 5 days (2009)
– Wildlife resource management; 5 days (2009)
– Population genetics; 5 days (2010)

SUPERVISION OF MSC STUDENT; 10 DAYS (3 ECTS)

– Mallard (*Anas platyrhynchos*) mtDNA phylogeography: genetic diversity and population structure of an important vector of Avian influenza

The research described in this thesis was financially supported by the KNJV (Royal Netherlands Hunters Association), the Dutch Ministry of Agriculture, the Faunafonds and the Stichting de Eik.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover drawing by Charlotte Grunow. Layout and setting by Max Schmidt.
Printed by Druckerei + Verlag Esser, Weilrod-Neuweilnau, Germany.