The genetics of migration in the Asian houbara bustard *Chlamydotis macqueenii*

by

Albertien van Heerden

Submitted in partial fulfilment of the requirements for the degree

Master of Science

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

22 November 2022

Supervisor: Prof Thierry B Hoareau (Reneco; UP extraordinary appointment)

Co-supervisor: Prof Fourie Joubert (Center for Bioinformatics and Computational Biology, UP)

Collaborators: Dr Loïc Lesobre (Reneco), Dr Yves Hingrat (Reneco)

I, **Albertien van Heerden,** declare that the thesis/dissertation, which I hereby submit for the degree **MSc in Bionformatics** at the University of Pretoria is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: 

DATE: 22 November 2022

Summary:

Seasonal migration in birds is an important behaviour that likely developed as an adaptive trait for birds to survive cold winters and low resource availability. Asian houbara is a partial migrant bird species that is of conservation concern, but it also serves as a good model to study the evolution of migration in birds.

In my literature review, I investigated the evolution of migration from both a population and organismal perspective. I discussed the literature on Asian houbara with the purpose to use this information to investigate possible genetic mechanisms underlying migration in this species. From the literature review, I also identified candidate genes that became the focus of my investigations in my second chapter.

In my research chapter I evaluated a recent genome assembly for Asian houbara and created an annotation based on this assembly. This assembly and annotation serve as a core genomic resource for this species and can support further genomic work on Asian houbara or closely related species. A newly generated transcriptomic resource will especially be useful in the annotation of the genomes of closely related species. I used a candidate gene approach to investigate 13 gene regions previously linked to bird migration. One gene, in particular, *DIO2*, showed a significant component of variance related to migratory behaviour. This result, together with the known functional roles of *DIO2*, highlights its potential role in the migratory behaviour of Asian houbara. Most other genes also showed a significant link to sampling locations and a strong effect of geography on genetic variation.

Our new genomic resource provides us with a foundation to study the genetic mechanisms behind migration in Asian houbara in the future. Further analysis on *DIO2*, using a larger dataset and long-read sequencing, could also reveal the possible influence of this gene on bird migration. Overall, genetic information can support the ongoing conservation efforts for Asian houbara, improving the understanding and management of this species.

Acknowledgements:

I would like to thank my supervisor, Prof Thierry Hoareau, for his leadership, support and compassion. I would also like to thank my co-supervisor, Fourie Joubert, for his guidance, and the Center of Bioinformatics and Computational Biology for the use of their high-performance computing cluster on which all of the bioinformatic analysis for this project was performed. I am grateful to the Molecular Ecology and Evolution Programme for giving me a platform to ask questions and to discuss various topics. Finally, I want to thank my collaborators, Dr Loïc Lesobre and Dr Yves Hingrat for their invaluable knowledge and for the opportunity to work with them and Reneco International Wildlife Consultants LLC on this project. The University of Pretoria provided the facilities for me to do this research.

Funds and samples used in this study were provided by the International Fund for Houbara Conservation (IFHC). We are grateful to His Highness Sheikh Mohamed bin Zayed Al Nahyan, President of the United Arab Emirates and founder of the IFHC, His Highness Sheikh Theyab bin Mohamed Al Nahyan, Chairman of the IFHC, and His Excellency Mohammed Ahmed Al Bowardi, Deputy Chairman, for their support. This study was conducted under the guidance of Reneco International Wildlife Consultants LLC, a consulting company that manages the IFHC's conservation programmes. We thank Dr Frédéric Lacroix, Managing Director of Reneco, for his supervision, as well as all staff of Reneco who participated in data collection.

Contents

Chapter 1

The evolution of migration in birds, with a specific interest in the Asian houbara bustard (*Chlamydotis macqueenii*)

1.1 Introduction

Migration is a form of uninterrupted movement in a specific direction and plays an important role in the life history and ecological niche of an organism (Dingle and Drake 2007). Both a population perspective in terms of ecology and evolution, as well as an individual perspective regarding physiology, behaviour and genetics can be useful in studying the evolution of migration in birds (Dingle and Drake 2007). The ultimate and proximate drivers of migration include those that vary depending on species and ecosystem as well as interspecific drivers that determine a species' ability to adjust to environmental changes. Migrants can be seen as either highly susceptible or highly adaptable when environmental change takes place and this depends on the intensity and rate of this change (Shaw 2016). Any evolutionary adaptation exists as a trade-off between two or more factors. Beneficial factors are often traded for adaptation, and the adaptation can be correlated to detrimental factors. Migration is a complex trait that is probably comprised of various trade-offs, for example, a trade-off between wintering survival and advantages on breeding grounds (Shaw 2016). Given this trade-off, individuals will only migrate if the benefits outweigh the costs (Pulido 2011). Migration is influenced by various regions in the genome and thus it is difficult to determine the genetic basis of this trait. Nonetheless, it is important to understand the underlying genetic mechanisms of migration in order to detect signals of selection associated with migratory traits. Finding these signals will ultimately help us understand the evolutionary basis of migration.

Interspecific, intraspecific and interpopulation variation in migratory behaviour can be determined by the environment or it can be the consequence of genetic adaptations to the environment, suggesting that there are underlying controls by endogenous genetic mechanisms (Pulido 2007). External influences such as habitat, climate, season, day length, stopover ecology and competition determine the adaptations of migratory behaviour (Alerstam et al. 2003; Gwinner 1996b; Newton 2007; Shaw 2016). Genetic mechanisms control migratory ability and strategy by changing body morphology, timing, directional and navigational systems and life history traits to suit the external conditions (Gwinner 1996a; Newton 2007; Piersma and Lindström 1997; Wiltschko and Wiltschko 2003).

In this review, I will synthesize and discuss existing knowledge on the evolution of migration from both population and organismal perspectives. I will also discuss the evolution of migration in the Asian houbara bustard (*Chlamydotis macqueenii*) as a species of interest. The purpose is to help design a study to better understand the genetics of migration in this partially migratory species, by gathering knowledge on the genetic basis of migration in birds. This should help improve current and future conservation efforts for this species.

1.2 Population perspectives on the evolution of migration

Migration is considered a behavioural adaptation that results in seasonal movement ultimately driven by a cost-benefit equilibrium between an organism's growth, survival or reproduction. These three advantages separate migration into three types of movement: alimental, climatic and gametic movement. Growth and survival are dependent on the availability of food resources, which drive alimental movement (Shaw 2016). Survival and reproduction are dependent on suitable habitat and environmental conditions, such as temperature and weather, which drive climatic and gametic movement (Newton 2007). Not all species necessarily show all three types of movement, but all three can be seen in birds performing return-migration (Shaw 2016). Among the factors that influence migration at the population level is the habitat and other biotic and abiotic factors such as predation, parasitism and weather conditions. These factors are discussed below.

1.2.1 Seasonal habitat changes are critical to drive migration

During winter, migratory birds overcome low resource availability by moving towards more productive habitats closer to the equator. In this case, migration is an adaptation to use different habitats to optimize the exploitation of resources (Alerstam et al. 2003). This change in habitat must be advantageous to the population, allowing for sustainability through continuous resource exploitation and successful breeding (Dingle and Drake 2007). Given the role of habitats in the adaptation of migratory birds, it is important to describe the different types of habitats used (breeding, wintering and transient habitats) and how these habitats can influence the success of migration.

Migratory birds have two main habitats. Birds that perform return-migration have a primary habitat, which is usually also their breeding area, and a secondary habitat at a lower latitude to which they migrate when conditions in their primary habitat become unfavourable, usually during the winter. (Shaw 2016). It is less clear why birds move to a higher latitude in the spring because most birds would be able to breed in their wintering habitat. Some exceptions exist and include for instance Arctic nesting shorebirds that breed on the Arctic tundra and winter on coastal mudflats where tidal flooding makes nesting impossible. A reason for birds moving back in the spring could be to avoid competition at lower latitudes and exploit the food resources that are available at a higher latitude in the summer (Newton 2007). Birds usually leave their breeding habitat before conditions deteriorate, a sign that the movement between habitats is under genetic influence (Pulido 2007). These habitats are shaped by many different factors. Over time, the breeding and wintering ranges of migratory birds are influenced by palaeoclimatic changes and the size of these ranges have been linked to changes in effective population size (Gu et al. 2021).

The transient areas through which they migrate also influence migratory birds. Migrant populations have an advantage compared to resident conspecifics in unsuitable breeding habitats because they can still exploit resources for survival and move to a different habitat before the breeding season (Alerstam et al. 2003). Besides the wintering habitats, this includes stop-over sites which are used to replenish the energy sources depleted during migration (Newton 2007). The habitats in which birds stay and through which they move expose them to a variety of selective pressure and migration must be ultimately beneficial for the organism to endure this exposure (Dingle and Drake 2007). Even with an ultimate benefit, migration always exists as a trade-off and is perceived as a hazardous process. For instance, bad weather conditions and storms during a long migration can kill many birds simultaneously, or birds may run out of fuel or die from exhaustion. If such events happen repeatedly in the same population, it could change the migratory timing or route or even eliminate the population (Newton 2007). When migratory birds congregate in large numbers at stop-over sites, and their energy requirements are high, competition for food sources can lower the rate at which they replenish their energy stores. Without sufficient energy stores, migratory birds will possibly not complete their migration (Moore and Yong 1991).

Resident populations stay in one area throughout the year, a feature that has its benefits and costs. Residential birds stay in one habitat all year long and do not experience the same conditions linked to reproduction, migration and wintering as migratory birds. This difference in conditions can cause a difference in the population sizes of residential and migratory populations. The breeding area of residents might not be able to sustain a large population year-round, and migration would then be an adaptation to achieve higher population numbers (Newton 2007). Another reason for the difference in population size could be the difference in breeding productivity (Alerstam and Enckell 1979). Clutch size and production of fledged chicks are generally greater at higher latitudes, where migratory birds breed than at lower latitudes, where resident birds breed. Reasons for this difference include longer day-length, lower food diversity, lower predation risk and lower competition at higher latitudes than lower latitudes. Higher productivity also allows migratory birds to compete successfully with resident birds in shared breeding areas (Alerstam and Enckell 1979).

Breeding habitats are usually similar for conspecific populations, regardless of whether they are migratory or residential, but the different populations have different advantages in these areas. The individuals that arrive first at the breeding grounds have an advantage because they can exploit the most resources with the least competition (Kokko 1999). The residential population, have a competitive advantage over resources, given that they can survive the non-breeding season in their breeding habitat (Chapman et al. 2011). Site-fidelity, an individual's tendency to stay in the same area, also gives residents an advantage over migrants. When the environment is well known, it promotes more efficient gathering of food resources, finding shelter and predator avoidance. This is a big factor that works against the evolution of migration and promotes residency in populations (Alerstam and Enckell 1979). When this habitat is unstable or uncertain, this advantage disappears and the migratory population will have the advantage (Alerstam et al. 2003). The steppe, savanna and dry woodland areas of northern and eastern Africa are examples of unpredictable and unstable habitats in which migratory birds would thrive compared to resident birds. Therefore, it serves as the primary wintering area for most completely migratory land bird species in the Palearctic-African migration system. Uncertain rainfall patterns and fires in the dry season makes it hazardous for year-long residents but suitable for migrants (Alerstam and Enckell 1979). Year-long

residence persists in populations when the benefits linked to the use of only one habitat for an entire year outweigh the costs of seasonal movement (Newton 2007).

1.2.2 Other biotic and abiotic factors may influence migratory populations

Environmental conditions such as weather during the breeding season, wintering season and migratory movement influence the success rate of migration. Therefore, it influences the timing, routes and strategy of migration in a population (Newton 2007). Predators, parasites and pathogens are more examples of dangers during migration, especially because of the high density of populations during migration (Newton 2007).

Interspecific interactions, such as predator-prey interactions and parasite-host interactions affect migration success. Predation is a factor that may increase mortality during migration, especially in smaller birds. The increased risk of predation during migration leads to adaptations that protect migratory birds against predation; these include habitat selection and timing to avoid migration times of predators (Alerstam et al. 2003). For example, to avoid predation some diurnal birds choose to fly during the night (Newton 2007). The occurrence of pathogens can control the seasonal movement of individuals. Migration can therefore be an adaptation to reduce the parasite load, while these is a cost for migrants being exposed to new strains of parasites (Alerstam et al. 2003).With increased host density, parasites and pathogens also tend to spread more easily and over long distances during bird migrations (Newton 2007).

1.3 Organismal perspectives on the evolution of migration

1.3.1 Physiological and physical traits that aid migratory birds in long-distance flights

Migratory birds can change their fuel load in preparation for migration. Migratory birds are adapted to accumulate enough fuel before and during migration to complete their journey (Newton 2007), even doubling their body mass within weeks before migrating. For instance, garden warblers (*Sylvia borin*) experience a seasonal increase in body mass controlled by endogenous circannual changes, which make them change from a basic weight of 16-18g to 37g when migrating over the Sahara desert, a seasonal increase in body mass controlled by endogenous circannual changes in body mass set point. Garden warblers kept under constant environmental conditions showed changes in body mass that are similar to that of wild individuals. After being fed a deliberately mass-reducing diet in the time they would reach their highest mass, the warblers returned to this high body mass when being fed a normal diet again (Figure 1.1) (Bairlein 2002). Migratory birds' digestive systems increase in size before migration to handle the above-average food intake. During flight, a strong heart, muscles and circulatory system ease the strenuous exercise (Newton 2007). Migratory birds can change the size of their muscles and organs according to their needs, as they interchange between fuel loading and flight. An increase in organ size and basal metabolic rate is necessary to achieve peak performance, but it is very energy costly to maintain these traits at such a high level. For this reason, organs are usually held at sub-maximal sizes when birds are not migrating. For example, black-necked grebes (*Podiceps nigricollis*) have tiny wing muscles when they moult and can not to fly, but these muscles double in size in the two weeks before they migrate (Piersma and Lindström 1997).

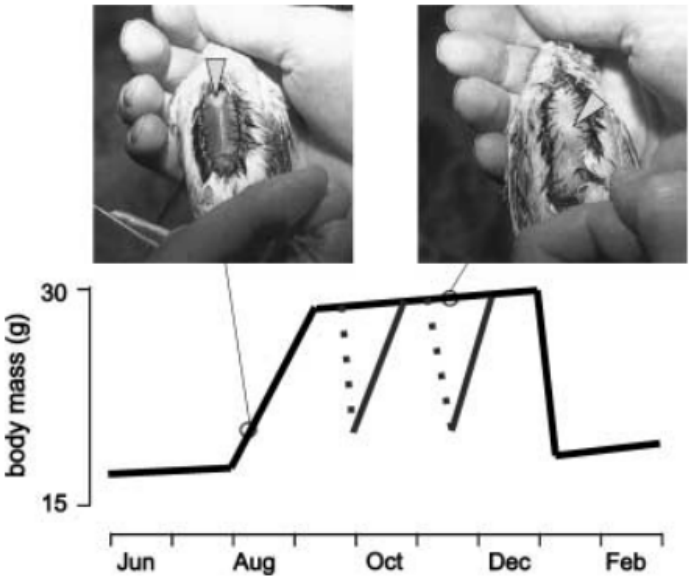


Figure 1.1: TOP LEFT: A lean garden warbler that would not have been able to cross the Sahara. TOP RIGHT: A fat garden warbler that would have been able to cross the Sahara. BOTTOM: The seasonal trend of fat accumulation in garden warblers. After food restriction (dotted line), they return to their set body mass for a specific time in the seasonal cycle (Taken from Bairlein (2002)).

A bird with a larger body mass requires more power, therefore, smaller birds can carry more fuel during migration because their overall mass will be relatively less than that of larger birds (Newton 2007). This risk of starvation decreases with an increase in energy reserves, but this increased weight comes with a cost. There is a trade-off between the risk of starvation and the risk of predation because the extra weight makes it more difficult for birds to escape predators. As a direct test of the success of predator evasion, blackcaps with a fat load between 1% and 60% were exposed to a simulated predator attack. The velocity and angle of take-off were measured to indicate the success of predator evasion. Birds with 60% fat showed a 32% slower take-off at an angle that was 17% smaller than that of birds with 1% body fat. This factor could contribute to shorter continuous flights and more regular rest stops being favoured over longer continuous flights. Fat loads of more than 50% are mostly seen in birds that have to fly over long stretches of uninhabitable areas, such as the Sahara desert. These barriers force them to carry a higher than optimal load, and a smaller load might be favoured if they could replenish energy stores during migration (Kullberg et al. 1996). Migratory birds must also carry the most energy-dense fuel during migration to harvest the most power from the smallest amount of mass. Therefore, fat is the ideal fuel because it is the most energy-dense, is stored without water, is efficiently metabolised and can be used by all the important body tissues for energy (Newton 2007).

Migratory birds have a variety of physiological changes linked to long-distance movements. The type and length of migration possible for a bird to perform depend on body size, wing shape, flight power and other features (Newton 2007). On average, larger birds fly marginally faster than smaller birds (Bruderer and Boldt 2001). A bird must exert power against gravity as well as forwards during continuous flapping flight. Breast muscles provide power for this movement and the tail works against wind drag (Newton 2007). Hormonal changes are also associated with migration, such as seasonal fluctuation in

hormonal condition which causes regression and enlargement of birds' gonads according to the breeding season. The hormones corticosterone and melatonin could also have roles in migration as they have been linked to migratory restlessness, fattening and nocturnal activity in migratory birds (Newton 2007).

1.3.2 Behavioural adaptations are involved in the timing, direction and pattern of migration

Partial migration and studies with birds in captivity provide a good study system for the genetic basis of migration. Species with partial migration provide good study systems to understand the genetic architecture of migration. The genetic basis of migration in birds is mostly studied by captive-breeding and release programmes, as well as cross-breeding experiments. The "migratory condition" of birds represents the behavioural changes that have to happen before migration and include the urge to depart and the tendency to fly in a particular direction. Migratory restlessness is particularly well recognised in captive birds as they move around in their cage and flutter their wings fast while perching. This behaviour is seen as equivalent to migration in wild individuals (Newton 2007).

Partial migration also provides a good model to study the difference between migratory and residential birds because it exists when certain individuals within a population are migratory and others are sedentary (Dingle 1996). Partial migration is divided into three types according to the cause of migration in the migratory populations: non-breeding migration, breeding migration and skipped-breeding migration (Shaw 2016). During non-breeding migration, migratory birds move away from their breeding area to a different habitat for the winter (Lundberg 1987). During breeding migration, migratory birds move away from sedentary birds during the breeding season (Gillis et al. 2008). The last type is skipped-breeding migration, where migratory birds cannot reproduce without migrating because they require a specialized habitat for reproduction (Shaw 2016). An example of a bird species that shows partial migration is the European blackcap (*Sylvia atricapilla*). The phenotype of blackcaps varies, with residential populations as well as short-, medium- and long-distance migrants (Merlin and Liedvogel 2019). This species is suited for long-term studies on the interaction between genotype and environment because they show a great variety of migratory habits (Berthold 1991). Cross-breeding experiments with *Sylvia atricapilla* form the majority of the research on the genetic basis of migration in birds.

Migratory birds have an inherent timing system for migration. The intrinsic migration condition of birds is recognised to have more control over migration than the conditions of their surrounding environment. Birds will migrate regardless of extreme conditions such as weather, even though it might cause delays in their migration. The internal condition of a bird is the determining factor for migration and determines the departure date through interaction with external factors (Newton 2007). The length of migration is also determined by these endogenous mechanisms. German blackcaps that perform longer migration show significantly more migratory restlessness than blackcaps from the Canary Islands that are nearly non-migratory. F1-hybrids between the two forms also show an intermediate pattern, which indicates that this feature is genetically determined (Berthold and Querner 1981). The timing program of migration is also a fixed trait. Repression of migratory restlessness by keeping birds in the dark or instating a feeding schedule did not influence the restlessness of the birds after the treatment (Gwinner 1996a). The timing of migration, breeding and moult is controlled by a circannual clock in both wild

and caged birds. Correlative studies of a candidate-gene approach have identified the *CLOCK* gene as a possible candidate for the timing of migration (Merlin and Liedvogel 2019).

In migratory birds, the length of daylight is known to trigger migratory behaviour. Birds that migrate away from their breeding sites to winter in more suitable conditions leave their breeding areas when day length shortens and conditions deteriorate. After winter they leave their wintering areas when day length lengthens and conditions improve (Newton 2007). Circannual programs, primarily the photoperiod cycle, are involved in the timing, course and duration of migration (Gwinner 1996a). Habitat also plays a role in the timing of migration, as discussed in Section 2.1. A trade-off exists between the benefit of a lengthened breeding period and the cost of late migration, with exhausted resources and the possibility that the wintering habitat is already inhabited (Newton 2007).

In nature, circannual rhythms are always a year long and mainly determined by the photoperiod cycle. In controlled conditions and constant photoperiod, the circannual rhythms can deviate from the natural length and occur multiple times in one year. By exposing Garden Warblers (*Sylvia borin*) to a shortened photoperiod cycle, they showed migratory restlessness up to 4 times a year and went through two gonadal cycles instead of one. Therefore, the cycles of some birds can be synchronised with the photoperiod cycle (Gwinner 1996a).

Migration influences the life history cycle of birds. Most birds experience one breeding season and moult a year. Together with migration, these factors form the yearly life cycle of the majority of migratory birds. Ultimate factors, such as the seasonal change in food availability, and proximate causes, such as day length, determines the inherent timing according to which birds perform this cycle. Birds can not breed and moult at the same time, the two processes are mutually exclusive. Migration could not have evolved without breeding and moult and the three factors are highly correlated. These processes must take place in the correct order at the correct time, which is why birds have an internal regulator. Birds in captivity, who do not experience the same environmental changes, also show the correct order and timing for these processes. They experience constant daylength but still perform these cyclic processes, which illustrates that this endogenous mechanism is genetically controlled (Newton 2007).

The experiences of migratory birds during migration affects their breeding season because migration and breeding are connected. Greater snow geese *(Chen caerulescens atlantica)* in Canada were hunted at one of their stop-over sites due to the high population numbers of the species. This hunt had a smaller effect on the number of birds killed in the hunt than it did on the ability of the birds to build up energy stores and on their successive breeding success (Mainguy et al. 2002). It is also strongly suggested that genetic factors are responsible for the differences in moult in subspecies. Experiments with blackcaps showed that F1-hybrids have an intermediate pattern of moult (Berthold and Querner 1982).

Resident birds' life history is also determined by endogenous mechanisms. An endogenous timing factor was found to have a significant influence on the control of yearly breeding and moult cycles in tropic birds that do not experience high seasonal fluctuation. The African stonechat (*Saxicola torquata axillaris*) is a non-migratory bird that occurs in tropic East Africa. They have a yearly breeding and moult cycle linked to the dry and rainy seasons. Their gonads enlarge during the dry season before the breeding season. They were studied in captivity to determine if their yearly cycle is determined by endogenous mechanisms. It was found that their yearly cycles are not only endogenous but innate as well.

For example, the difference in the "reproduction window" (the interval between phases of gonadal growth and regression) between European and African stonechats, that were held in the constant equatorial photoperiod, was determined. The difference correlated with the length of the breeding season of the respective species in the wild. Another experiment showed that this trait is genetically determined, with F1-hybrids showing intermediate features. Experiments with African stonechats showed that individual reproductive ability can not change the range of their reproductive window (Gwinner 1996b).

The same endogenous mechanisms are seen in moult. When European and African stonechats are kept in the same constant equatorial photoperiod, there is a difference in the duration of their moult. This may be connected to the fact that European stonechats are migratory birds and have less time available for moult before they have to migrate (Gwinner 1996b).

Migratory birds have specialised navigation systems. Studies on birds' navigation systems are mainly done with homing pigeons (*Columbia livia f. domestica*), but data from wild birds indicate that they use the same navigation systems. There is a difference in behaviour between migratory birds and homing pigeons because first-time migrants have a destination unfamiliar to them. An endogenous migration program seems to provide the direction of migration, as well as the distance of migration, to first-time migrants. The information about the direction is determined before migration and given to the bird as a compass course. The geomagnetic field that indicates magnetic north, and the celestial rotation which indicates geographical north, are used to establish this course. The celestial rotation is indicated by the rotation of the stars at night and the rotation pattern of polarised light during the day. Migratory birds can find their course through their compasses as soon as it is established (Wiltschko and Wiltschko 2003).

Migration to a familiar destination, such as in the spring, is the same as homing movement seen in homing pigeons. Migratory birds have a map of their local area, which is probably their breeding area. They are familiar with the local geography and navigation system. This familiar information includes the terrain over which they fly during migration. After their first migration, migratory birds do not exclusively rely on inherited information. The inherent course persists, but experience plays a more important role. It allows migratory birds to adapt their route to avoid negative effects such as wind (Wiltschko and Wiltschko 2003). Migratory birds react to specific geographical cues to complete their migration. Birds can navigate with the stars at night and topographic features such as mountains and coastlines during the day. Some birds can also use their olfactory and auditory senses for navigation (Newton 2007).

Juvenile birds can find their wintering areas without help from experienced adults. Birds must not only find the direction to their end destination but also know when to stop to eat or change direction (Newton 2007). The direction of migration is inherited. In a cross-breeding experiment between a migratory and a residential blackcap population, the migratory F1 individuals preferred the same migratory direction as their migratory parent. Cross-breeding experiments with two blackcap populations that migrate in different directions delivered F1-progeny that show intermediate behaviour. Changing direction during migration is also inherited (Berthold 1991). The direction of migration is also controlled by circannual programs. Garden warblers (*Sylvia borin*) kept at a constant photoperiod changed their direction-preferences during nocturnal migratory restlessness at the same time as they would have in the wild. During the experiments, the birds were not exposed to star patterns as guides, but still to the earth's magnetic fields, which probably influenced their endogenous mechanisms to determine direction (Gwinner 1996a).

1.3.3 Patterns of migration differ among populations and among individuals within populations

Migration routes are specifically suited to the needs of a population or individuals in a population. Evolutionary changes like those affecting the degree and pattern of migration can take place without phylogenetic constraints. In other words, we see different migration patterns in species that are closely related and even in different populations from the same species (Alerstam et al. 2003). Migratory birds take detours regularly from their migration route to get the maximum benefit from continuous habitat where they can rest and eat. They also detour to avoid high-risk weather conditions or predators and to conserve energy through improved wind conditions. These detours are part of a trade-off between the energy and time conservation of the direct route and the avoidance of risk when using the detours. This difference in migration routes depending on the conditions can occur as loop migration where one route is used in the autumn and another in the spring (Newton 2007).

Migratory restlessness has been used to determine the length of migration on population level. The length of the time that an individual bird shows migratory restlessness depends on the population from which it comes and the migratory behaviour of that population. The longer the population migrates, the longer the bird will show migration restlessness. F1-progeny of two birds on opposite sides of the spectrum show intermediate behaviour that strongly support the heritability of this behaviour (Newton 2007).

Migratory behaviour can change rapidly under strong selection pressure. The migration behaviour of birds can change rapidly under strong selective pressure, such as the pressure coming from climate change and hunting (Newton 2007). Therefore, the changes in migration habits are of interest in the short and long term. Given that failed migration has a direct impact on the fitness of migratory birds, natural selection has the potential to change migratory populations and migration phenotypes fast. The selection pressure of failed migration can lead to changes in population structure, bottlenecks, inbreeding depression and even extinction (Lennox et al. 2016).

In a cross-breeding experiment between a migratory and residential blackcap population, 40% of the F1-hybrids were migrants. Therefore, the urge to migrate can be bred into a non-migratory population. Furthermore, the fact that all F1-hybrids were not migrants showed that the urge to migrate is probably a multi-locus system with a threshold (Figure 1.2) (Berthold 1991). By selecting a partially migratory population for high and low migratory restlessness and interbreeding individuals from the two categories, an exclusively migratory population was achieved in 5-6 generations and an exclusively residential population in 3 generations. Therefore, migratory behaviour can change rapidly when under high selective pressure. In nature, selective pressure is never this high, and such changes will probably take place over decades. Migratory behaviour can also change to such an extent that there is a conversion between a migratory population and a residential population. For example, for a partially migratory population that experiences abnormally harsh winters where the migratory individuals survive better than the residential individuals, the expectation is that the proportion of the population made up of migratory individuals will increase with each generation. The same applies to mild winters, where residential birds have the advantage of a good breeding site and deliver more progeny than migratory birds. In such a situation the proportion of residential birds in the population will increase(Newton 2007). Overall, the fraction of migrants and residents in a partially migratory population can change over time as this equilibrium is influenced by many factors.

If the urge to migrate can be achieved in a residential population within a few generations, the genetic variation necessary for migration must be present in the residential population. The threshold model of quantitative genetics provides a good explanation for how it is possible. The urge to migrate is a continuous variable and there is a threshold that determines if it is expressed. Residence persists when the urge to migrate falls beneath the threshold and migratory behaviour takes place as a result of strong directional selection (Figure 1.2) (Pulido 2007).

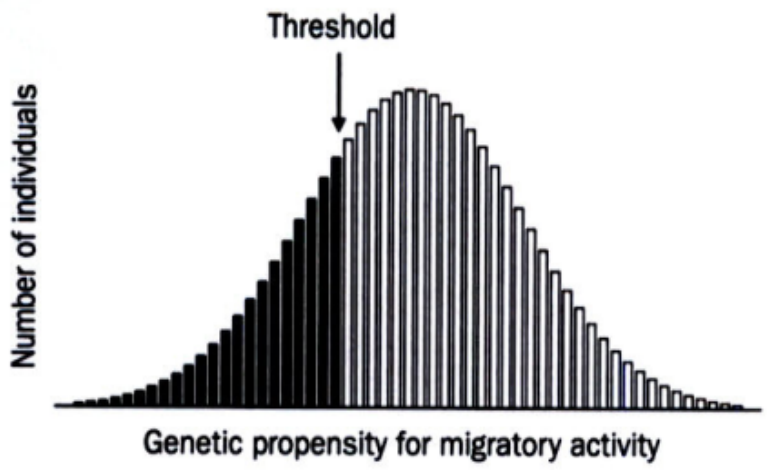


Figure 1.2: The threshold model for migratory behaviour in a partially migratory species. Residency exists when migratory propensity is beneath the threshold, and migration exists when migratory propensity exceeds the threshold (Taken from Pulido 2007) (Pulido 2007).

The rate of evolutionary change in any population depends on three factors: the amount of genetic variation in the population, the strength and consistency of selection pressure and the degree to which selection for one trait causes a parallel change in another trait. Selection pressure must be consistent in the same direction over multiple generations for permanent change to take place. Most factors that form part of the "migratory syndrome" are dependent on each other. A change in one factor will cause a change in another. In case such a change is detrimental and fixed in the population due to strong selective pressure, it would take a long time to make a positive change in the opposite direction (Newton 2007).

Migratory traits are probably controlled by the selection process on relatively few genome regions or by changes in gene expression. This hypothesis is supported by the finding that only a few genes are involved in the expression of migratory traits and the strong genetic correlation between migratory traits for which the variation in one trait is dependent on the variation in another trait (Merlin and Liedvogel 2019).

1.3.4 Perspectives from genes potentially involved in bird migration

***CLOCK* and *ADCYAP1* are the most commonly studied genes involved in migration.** The vertebrate circadian gene, *CLOCK,* was found to be involved in the circadian rhythms of humans (Katzenberg et al. 1998; Mishima et al. 2005, as cited by (Johnsen et al. 2007) and trout (Leder et al. 2006, as cited by (Johnsen et al. 2007). This prompted Johnson et al. (2007) to investigate whether allelic length variation in the *CLOCK* poly-Q coding sequence is associated with migratory behaviour (breeding

latitude) in the migratory bluethroat (*Luscinia svecica*) and largely non-migratory blue tit (*Cyanistes caeruleus*). They found a statistically significant correlation between average allele length and latitude among the blue tit populations but found no significant correlation for bluethroat populations. They did not make any firm conclusions on whether *CLOCK* gene allelic variation is under selection in these species, but their work prompted others to do similar studies. Liedvogel et al. (2009) found a general, weak trend for blue tits with fewer *CLOCK* poly-Q repeats to breed earlier (Liedvogel et al. 2009), but no evidence was found for an association between genotype and reproductive timing in a sympatric great tit (*Parus major*) population (Liedvogel and Sheldon 2010). Similarly to Johnson et al. (2007), Mueller et al. (2011) found an association between allele length and migratory restlessness in the *ADCYAP1* gene, but not in the other 5 genes they investigated, which include *CLOCK* (Mueller et al. 2011).

In a study testing whether the findings of Liedvogel et al. (2009) can also be observed in barn swallows (*Hirundo rustica*), Caprioli et al. (2012) compared the mean breeding dates of four *CLOCK* poly-Q genotypes in both yearlings and older individuals (Caprioli et al. 2012). They only found a significant difference in the mean breeding dates of one pair of genotypes and only for yearling females, a similar result to previous studies . In further studies on barn swallows, Bazzi et al. (2015) found a significant association between a rare *CLOCK* genotype and delayed timing of autumn migration (Bazzi et al. 2015). Bazzi et al. (2016) implemented a between-species comparative approach to investigate whether polymorphism at *CLOCK* and *ADCYAP1* are associated with traits related to migration and geographic distribution (Bazzi et al. 2016). They found a strong phylogenetic signal in both *CLOCK* allele size and gene diversity, as well as a positive association between maximum *CLOCK* allele size and breeding latitude, but they did not find any significant association between migration or distribution traits and *ADCYAP1* gene diversity. Saino et al. (2015) investigated the association between migratory timing and allele length at *CLOCK* poly-Q and *ADCYAP1* in four trans-Saharan long-distance migrants. They found, depending on sex and whether mean allele length or the length of the longer allele was considered, that individuals with more glutamine residues in *CLOCK* poly-Q migrated significantly later (Saino et al. 2015). They, however, did not find any association between migration date and *ADCYAP1* polymorphism. Other studies also found no significant association between *ADCYAP1* mean population allele length and the migratory status of the population (Peterson et al. 2013, Contina et al. 2018), but Bourret and Garant (2015) found a significant relationship between laying date and female *ADCYAP1* genotypes in interaction with latitude, highlighting the importance of including genotype-environment interactions (GxE) in the analysis (Bourret and Garant 2015).

A variety of other genes have also been studied, but not extensively. *CLOCK* and *ADCYAP1* are much more prevalent in the literature than any single other gene, with *CLOCK* showing much more promising results than *ADCYAP1*. There are, however, many other genes that have been investigated once or a few times that are worth mentioning. The genes *NPAS2*, *CRAB1* and *DRD4* were all investigated more than once. Fidler et al. (2007) found a significant association between *DRD4* genotype and novelty seeking behaviour in great tits (Fidler et al. 2007). Steinmeyer et al. (2009) followed a candidate gene approach to detect potentially functional polymorphisms in blue tits in *CLOCK*, *ADCYAP1, NPAS2*, *CRAB1* and other genes linked to a circadian phenotype in other organisms. They detected polymorphism in some of the genes, but the SNPs were silent and could not be linked to any function or phenotypic trait (Steinmeyer et al. 2009). Mueller et al. (2011) also investigated these three genes and did not find any significant associations between them and migratory behaviour (Mueller et al. 2011). Most recently,

Gu et al. (2021) found a significant signature of selection at the *ADCY8* locus in peregrine falcons (*Falco peregrinus*). They showed that this locus could be involved in the long-term memory of birds, which is important for long-distance migrants (Gu et al. 2021).

Transcriptomic approaches have also been utilised to study the genetic basis of migration in birds. Johnston et al. (2016) performed RNA-seq on the ventral hypothalamus and optic chiasma of captive Swainson's thrushes (*Catharus ustulatus*) in migratory condition and found 188 differentially expressed genes. Of these, they identified two particularly strong candidate genes, *CRABP1* and *DIO*2, that might be drivers of neural plasticity associated with migratory behaviour (Johnston et al. 2016). Francini et al. (2017) used high-throughput transcriptomics in blood samples to compare gene expression in resident and migrant European blackbirds (*Turdus merula*). They detected only four differentially expressed genes between residents and migrants: *MLNR, TOP2B, ST6GALNAC2* and *TGFBR1;* and five genes differentially expressed in groups with varying migratory behaviour: *HERC4, RAD51D, HIST1H2A4, TSPO2* and *FECH* (Franchini et al. 2017).

1.4 The evolution of migration in the Asian houbara bustard (*Chlamydotis macqueenii*)

1.4.1 Background

The Asian houbara bustard (*Chlamydotis macqueenii*) is a medium-sized ground-dwelling bird that occurs from the Nile river to the Gobi desert in Mongolia (Figure 1.3) (Cramp and Simmons 1980). It was recently declared a separate species from the African houbara bustard (*Chlamydotis undulata*) (Knox et al. 2002). African houbara share a recent ancestor with Asian houbara, which probably moved to developing steppe and drier land in the north at the end of the last glacial period and developed migratory behaviour (Combreau et al. 2011a). Some individuals of Asian houbara travel more than 7500 km a year and are seen as long-distance migrants while others are resident, therefore, the species is considered as partially migratory.

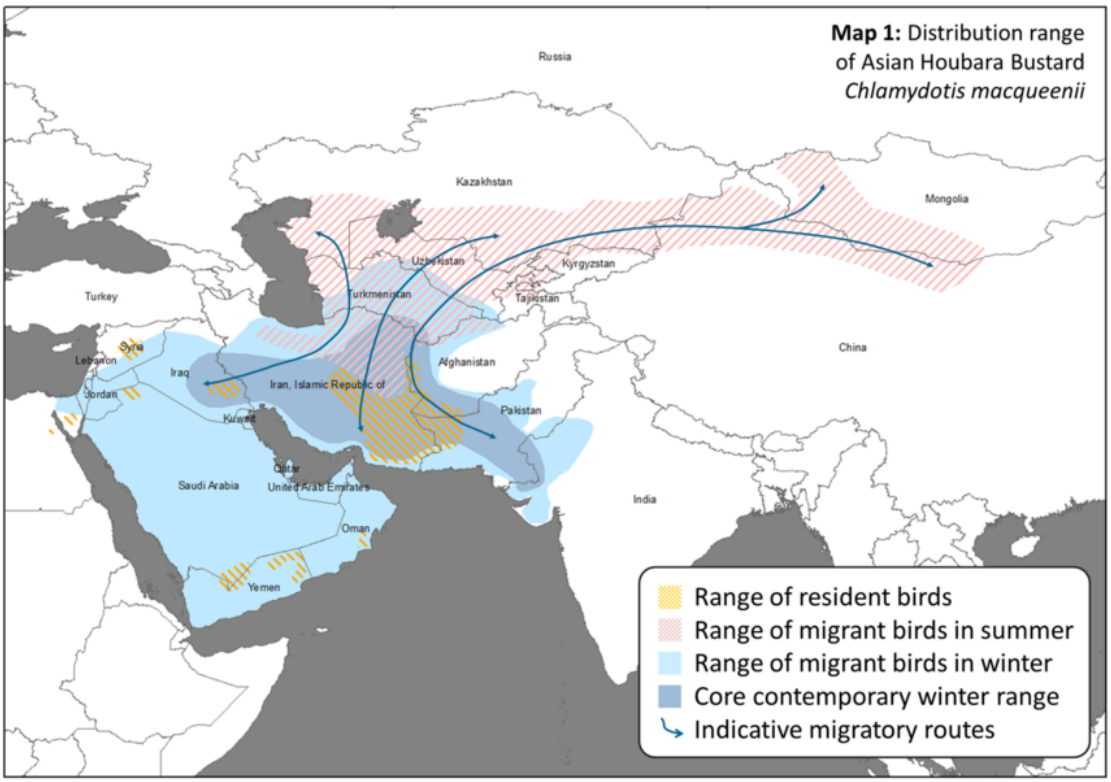


Figure 1.3: Map of Asian Houbara bustard distribution range (Bird Life International 2014).

Asian houbara's habitat comprises steppe, desert and semi-desert (Combreau et al. 2011a). A study on migratory populations in Iran showed that climate (temperature and rainfall), followed by topographic factors (terrain roughness and vegetation), has the biggest influence on the distribution of breeding and wintering populations. Breeding populations decreased with an increase in temperature and wintering populations increased with an increase in temperature, a finding that coincide with the migration patterns of Asian houbara, with breeding and wintering populations found in the northern and southern latitudes of the range respectively. Variation in vegetation and landscape provides hiding places for bustards, but too much roughness in terrain limits their sight and puts them at high risk of predation. Habitats with 15% to 40% vegetation coverage which are distant from rural areas and towns are favourable by Asian houbara (Pakniat et al. 2020). During migration, bustards generally avoid large physical barriers, such as water masses and high mountains, and prefer to fly further at lower elevations. This strategy is often used by birds that use flapping flight during migration (Combreau et al. 2011b). Asian houbara are migratory across the whole of Central Asia, which represents a large part of their distribution range. In the rest of their distribution range that include South-West Asia, Afghanistan and Pakistan, Asian houbara are found in fragmented populations that are mostly residents, except for a population characterised by migratory behaviour in North-West Iran. Residential populations on the Arabian Peninsula declined drastically due to pressure from humans and only small populations remain in Yemen, the United Arab Emirates and Southern Iran.

The Asian houbara bustard is of conservation interest. The Asian houbara bustard is listed as vulnerable on the IUCN (International Union for Conservation of Nature) red list due to overhunting and habitat loss (Combreau et al. 2011b). The species is the most sought-after prey in Arabian falconry, which makes them of economic and cultural significance. The bustards are mostly hunted in their wintering areas and are captured in large numbers to be used in training falcons. This type of overexploitation has led to a drastic decrease in population numbers (International 2014). The Asian houbara population is estimated to consist of between 50,000 and 99,000 individuals and has decreased at a rate of 30 to 49% over 20 years, which represents only a three-generation window . The main conservation method for Asian houbara is captive breeding to support both release and regulated hunting programmes. At first, only the residential population in the Arabian Peninsula and Africa was supported by this process, but due to the increased hunting of the bustards, the breeding programme was expanded to support migratory populations in Asia as well (Burnside et al. 2017).

1.4.2 Migratory populations

Asian houbara bustard migratory populations are possibly highly connected, even though they follow various migration strategies. Migratory populations follow various routes, a feature that is thought to be genetically determined (Combreau et al. 2011a). Translocation experiments with three Asian houbara populations (over a latitudinal gradient in Uzbekistan) were done to study how the genetic control of migration influences individuals and populations (Burnside et al. 2020). The three populations showed similar timing and orientation of migration but differed in terms of migration distance and latitude of wintering area. Translocated bustards migrated in the same direction as their source population and kept migrating longer than individuals in their receiver population as they originated from higher latitudes than their receiver population. They even migrated past wintering areas used by their receiver population. The migratory behaviour of translocated individuals indicates a potential significant genetic component linked to the different migration strategies (Burnside et al. 2020).

To study the influence of sex, age, geographical origin and season on the migration strategies of Asian houbara, Madon et al. (2015) performed satellite tracking on 414 wild migratory houbara. Seeing as Asian houbara has a polygynous mating system and male bustards show high site fidelity to the display territories for which they compete, differences in migratory strategy according to sex are expected. Male Asian houbara has a migration with shorter duration than females, partly because they spend less time at stopover sites. Male Asian houbara also migrate earlier in the spring to arrive at their breeding grounds as soon as possible to find the best high-quality territories. As in many migratory bird species, juvenile Asian houbara migrate before adults. Juveniles spend more time at stopover sites and have a higher mortality rate than adults, especially during their first autumn migration. Regarding geography, Asian houbara from breeding sites at lower latitudes had shorter migration distances than those from breeding sites at higher latitudes. Asian houbara with longer journeys were also more likely to stop over (Madon et al. 2015).

Asian houbara migratory populations pose a unique challenge for captive-breeding and release programmes. It is challenging to establish captive-bred birds in migratory populations, given the higher mortality risk and the potential erosion of migratory behaviour during captivity. For instance, individuals that are translocated from residential populations to migratory populations will likely fail

to migrate. To overcome the decline of populations in the wild, released captive-bred individuals must complete their yearly migration. Therefore, they have to be physiologically and behaviourally adapted to follow the migration patterns of the population into which they are released (Burnside et al. 2017).

Unintended selection of features that are beneficial in captivity can increase the occurrence of alleles that may be deleterious in the wild. For example, physiological and morphological traits that influence flight performance may be counter-selected in captivity, which could lead to lower survival and breeding rates of captive individuals in the wild (Burnside et al. 2017). During a study in Uzbekistan, Burnside et al. (2016) used satellite tracking to monitor 65 captive-bred Asian houbara after their release in April with the objective of determining their survival rate. Of the 65 birds, 27 died in summer, resulting in a survival rate of 59.5% six months after the release. Of the 21 birds that migrated, only 6 survived and returned after winter, which illustrate the general trend of lower survival rates in captive-bred birds compared to their wild conspecifics (Burnside et al. 2016). To survive their first migration, captive-bred birds must follow the correct migration strategy, which includes timing and direction among other factors. Birds that fly outside of the set direction of migration will die off and the average direction of migration will persist in the rest of the population (Burnside et al. 2017). Selection after release can fail to eradicate maladaptations which could introgress into the wild population and the offspring of captive-bred individuals can have lower fitness than those of wild individuals (Dolman et al. 2018). However, given the low survival rate of captive-bred individuals in the wild and the disadvantage of the maladapted trait, this is unlikely to happen. Long-term survival of captive-bred Asian houbara is influenced by the quality of habitat, resources, climate, predation rates and density-related mechanisms (Azar et al. 2016).

Contrasted migration strategies can exist between wild individuals and captive-bred individuals, even if maladaptations are not established in captive-bred individuals. To test this hypothesis, Burnside et al. (2017) compared the migration strategies of 46 wild Asian houbara bustards and 29 captive-bred bustards. The birds that were bred in captivity were descendant from the wild population in Uzbekistan from which the wild individuals were taken. Captive-bred individuals had a later departure date than wild individuals, but a similar initial direction of migration and arrival date in the wintering area. Because of this, captive-bred individuals wintered closer to their breeding area than wild individuals. The higher latitude wintering area of captive-bred individuals and strong wintering site fidelity can have long-term consequences on the survival of these individuals, but the true effect is still uncertain (Burnside et al. 2017).

Strong site fidelity is observed in Asian houbara populations, both for the breeding area in which they are released and the wintering areas (Burnside et al. 2017; Burnside et al. 2020). In a species where migration is genetically determined, translocation from an allopatric population can disturb the receiver population's migration strategies and hamper the population's fitness (Burnside et al. 2020). Therefore, it is important that birds used in breeding programmes are released in the same populations from which they descend and that follow the same endogenous migrations strategy as them (Burnside et al. 2017). For conservation, it is suggested that residential birds and migratory birds are not bred together and that birds should not be released outside of their geographical origin to retain the population structure necessary to preserve migration ability in the population (Burnside et al. 2020). Understanding the genetic basis of migration in Asian houbara will aid in the identification of migratory status that is important for conservation to take place successfully.

The migration of Asian houbara is controlled by genetic components. The presence of both migratory and resident populations in one species, as well as the separation of migration routes and high site fidelity, can limit gene flow between populations which can lead to significant genetic differentiation across the distribution range. To study this genetic differentiation, Riou et al. (2012) took samples from 108 Asian houbara individuals over their distribution range. They genotyped the individuals at 17 polymorphic microsatellite loci and analysed them. Genetic differentiation was most apparent between populations in the Arabian Peninsula, especially between Yemen and Central Asia. This differentiation indicates limited gene flow between residential Arabian populations and the migratory Central Asian populations. Negligible population structure in Central Asia indicates high gene flow between migratory populations that follow various migration routes (Riou et al. 2012), even though migratory route is genetically determined to a certain extent (Helbig 1996). Microsatellite markers probably have a resolution that is too low to detect genetic structure between migration routes and a whole-genome approach could uncover previously undetected structure.

Another way to study the genetic control mechanisms of migration is to look at the patterns that juvenile individuals follow during their first migration. Satellite tracking of both captive-bred individuals descendant from migratory birds and wild juveniles indicated various genetic components of Asian houbara migration (Burnside et al. 2017). The majority of captive-bred individuals migrated which indicated that migratory behaviour is indeed genetically determined. The initial direction and duration of migration were similar between captive-bred and wild juveniles, which indicates the genetic determinism of migratory strategy. Similar arrival dates in captive-bred and wild juveniles, which were independent of their departure dates, indicate an endogenous cue to stop migration, such as day-length. A difference in stop-over schedule between juveniles and adult individuals possibly indicates that the decision to stop over is facultative but restarting migration has a strong genetic component (Burnside et al. 2017).

To study these genetic components further, loci possibly involved in the migration of Asian houbara can be identified with a candidate gene approach to genome analysis. Additionally, the genetic structure and status of the species over its distribution range can also be determined by analysing informative sites across the genome.

1.5 Conclusion

Extensive research has been done on bird migration, both in terms of environmental pressures faced during migration and endogenous mechanisms controlling migration. This research is mostly based on experiments using cross-breeding or a controlled environment in captivity to measure birds' response to environmental change. Studies that evaluated the link between specific genes and migration have identified the *CLOCK* gene as the candidate with the most potential to advance knowledge in the genetics of migration. To date, the candidate gene approach is still inconsistent and limited, probably because migratory behaviour is under complex multi-locus influence. Candidate genes will likely remain species-specific or at least limited to certain clades. Most of the work done to date involves small passerine birds, and the results are difficult to generalize to underlying control mechanisms of migration in larger birds with contrasting behaviour. Very few studies have addressed the genetic basis of migration in a large, non-social, non-passerine, terrestrial long-distance migrant such as the Asian houbara bustard.

Regarding Asian houbara, the main research methods remain satellite tracking of wild and captive-bred

individuals for migratory populations. Studies on genetic differentiation are done on resident and migratory populations, but there is a lack of research on the genetic basis of migration in this species that separates residents from migrants. With four complete Asian houbara bustard genomes available on GenBank (BioProject PRJNA473697), from two residents and two migrants, genomic analysis is the next easy step in researching the genetic basis of migration in this species. Analysis of these genomes, combined with resequencing data from individuals across the distribution range of Asian houbara, will aid in choosing suitable populations for captive-breeding programmes and the release of captive-bred individuals in the wild. This will also open up opportunities for similar studies in species that face the same challenges as the Asian houbara bustard in the future.

1.6 References

Alerstam T., and P. H. Enckell, 1979 Unpredictable habitats and evolution of bird migration. Oikos 33: 228-232. https://doi.org/10.2307/3543999

Alerstam T., A. Hedenström, and S. Åkesson, 2003 Long-distance migration: evolution and determinants. Oikos 103: 247–260. https://doi.org/10.1034/j.1600-0706.2003.12559.x

Azar J. F., P. Rautureau, M. Lawrence, G. Calabuig, and Y. Hingrat, 2016 Survival of reintroduced Asian houbara in United Arab Emirates' reserves. The Journal of Wildlife Management 80: 1031–1039. https://doi.org/10.1002/jwmg.21085

Bairlein F., 2002 How to get fat: nutritional mechanisms of seasonal fat accumulation in migratory songbirds. Naturwissenschaften 89: 1–10. https://doi.org/10.1007/s00114-001-0279-6

Bazzi G., R. Ambrosini, M. Caprioli, A. Costanzo, F. Liechti, *et al.*, 2015 Clock gene polymorphism and scheduling of migration: a geolocator study of the barn swallow *Hirundo rustica*. Scientific Reports 5: 12443.
https://doi.org/10.1038/srep12443

Bazzi G., J. G. Cecere, M. Caprioli, E. Gatti, L. Gianfranceschi, *et al.*, 2016 Clock gene polymorphism, migratory behaviour and geographic distribution: a comparative study of trans-Saharan migratory birds. Molecular Ecology 25: 6077–6091. https://doi.org/10.1111/mec.13913

Berthold P., and U. Querner, 1981 Genetic basis of migratory behavior in European warblers. Science 212: 77–79. https://doi-org./10.1126/science.212.4490.77

Berthold P., and U. Querner, 1982 Genetic basis of moult, wing length, and body weight in a migratory bird species, *Sylvia atricapilla*. Experientia 38: 801–802. https://doi.org/10.1007/BF01972279

Berthold P., 1991 Genetic control of migratory behaviour in birds. Trends in Ecology & Evolution 6: 254–257. https://doi.org/10.1016/0169-5347(91)90072-6

Bird Life International, 2014 Review of the global conservation status of the Asian Houbara Bustard *Chlamydotis macqueenii*. BirdLife International.

Bourret A., and D. Garant, 2015 Candidate gene–environment interactions and their relationships with timing of breeding in a wild bird population. Ecology and Evolution 5: 3628–3641. https://doi.org/10.1002/ece3.1630

Bruderer B., and A. Boldt, 2001 Flight characteristics of birds: Ibis 143: 178–204. https://doi.org/10.1111/j.1474-919X.2001.tb04475.x

Burnside R. J., N. J. Collar, K. M. Scotland, and P. M. Dolman, 2016 Survival rates of captive-bred Asian Houbara *Chlamydotis macqueenii* in a hunted migratory population. Ibis 158: 353–361. https://doi.org/10.1111/ibi.12349

Burnside R. J., N. J. Collar, and P. M. Dolman, 2017 Comparative migration strategies of wild and captive-bred Asian Houbara *Chlamydotis macqueenii*. Ibis 159: 374–389. https://doi.org/10.1111/ibi.12462

Burnside R. J., C. Buchan, D. Salliss, N. J. Collar, and P. M. Dolman, 2020 Releases of Asian houbara must respect genetic and geographic origin to preserve inherited migration behaviour: evidence from a translocation experiment. Royal Society Open Science 7. https://doi.org/10.1098/rsos.200250

Caprioli M., R. Ambrosini, G. Boncoraglio, E. Gatti, A. Romano, *et al.*, 2012 Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. PLOS ONE 7: e35140. https://doi.org/10.1371/journal.pone.0035140

Chapman B. B., C. Brönmark, J.-Å. Nilsson, and L.-A. Hansson, 2011 The ecology and evolution of partial migration. Oikos 120: 1764–1775. https://doi.org/10.1111/j.1600-0706.2011.20131.x

Combreau O., F. Launay, M. A. Bowardi, and B. Gubin, 1999 Outward migration of Houbara Bustards from two breeding areas in Kazakhstan. The Condor 101: 159–164. https://doi-org.uplib.idm.oclc.org/10.2307/1370458

Combreau O., S. Riou, J. Judas, and M. Lawrence, 2011a Population structure, migratory connectivity and inference on gene exchange mechanisms in the Asian Houbara Bustard *Chlamydotis macqueenii*: a summary of recent findings. Zoology in the Middle East 54: 107–110. https://doi.org/10.1080/09397140.2011.10648902

Combreau O., S. Riou, J. Judas, M. Lawrence, and F. Launay, 2011b Migratory pathways and connectivity in Asian Houbara Bustards: evidence from 15 Years of satellite tracking. PLOS ONE 6: e20570. http://dx.doi.org.uplib.idm.oclc.org/10.1371/journal.pone.0020570

Contina A., E. S. Bridge, J. D. Ross, J. R. Shipley, and J. F. Kelly, 2018 Examination of Clock and Adcyap1 gene variation in a neotropical migratory passerine. PLOS ONE 13: e0190859. https://doi.org/10.1371/journal.pone.0190859

Cramp, S., and K. E. L. Simmons, ed. 1980 *Handbook of the birds of Europe, The Middle East and North Africa.* Oxford University Press, Incorporated, New York.

Dingle H., 1996 *Migration: The Biology of Life on the Move.* Oxford University Press, Incorporated, New York.

Dingle H., and V. A. Drake, 2007 What Is Migration? Bioscience; Oxford 57: 113–121. http://dx.doi.org.uplib.idm.oclc.org/10.1641/B570206

Dolman P. M., N. J. Collar, and R. J. Burnside, 2018 Captive breeding cannot sustain migratory Asian houbara Chlamydotis macqueenii without hunting controls. Biological Conservation 228: 357–366. https://doi.org/10.1016/j.biocon.2018.10.001

Fidler A. E., K. van Oers, P. J. Drent, S. Kuhn, J. C. Mueller, *et al.*, 2007 Drd4 Gene polymorphisms are associated with personality variation in a passerine bird. Proceedings: Biological Sciences 274: 1685–1691.

Franchini P., I. Irisarri, A. Fudickar, A. Schmidt, A. Meyer, *et al.*, 2017 Animal tracking meets migration genomics: transcriptomic analysis of a partially migratory bird species. Molecular Ecology 26: 3204–3216. https://doi.org/10.1111/mec.14108

Gillis E. A., D. J. Green, H. A. Middleton, and C. A. Morrissey, 2008 Life history correlates of alternative migratory strategies in American Dippers. Ecology 89: 1687–1695. https://doi.org/10.1890/07-1122.1

Gu Z., S. Pan, Z. Lin, L. Hu, X. Dai, *et al.*, 2021 Climate-driven flyway changes and memory-based long-distance migration. Nature 591: 259–264. https://doi.org/10.1038/s41586-021-03265-0

Gwinner E., 1996a Circadian and circannual programmes in avian migration. Journal of Experimental Biology 199: 39–48. https://doi.org./10.1242/jeb.199.1.39

Gwinner E., 1996b Circannual clocks in avian reproduction and migration. Ibis 138: 47–63. https://doi.org/10.1111/j.1474-919X.1996.tb04312.x

Helbig A., 1996 Genetic basis, mode of inheritance and evolutionary changes of migratory directions in palaearctic warblers (Aves: Sylviidae). Journal of Experimental Biology 199: 49–55.

Johnsen A., A. E. Fidler, S. Kuhn, K. L. Carter, A. Hoffmann, *et al.*, 2007 Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. Molecular Ecology 16: 4867–4880. https://doi.org/10.1111/j.1365-294X.2007.03552.x

Johnston R. A., K. L. Paxton, F. R. Moore, R. K. Wayne, and T. B. Smith, 2016 Seasonal gene expression in a migratory songbird. Molecular Ecology 25: 5680–5691. https://doi.org/10.1111/mec.13879

Knox A. G., M. Collinson, A. J. Helbig, D. T. Parkin, and G. Sangster, 2002 Taxonomic recommendations for British birds. Ibis 144: 707–710. https://doi.org/10.1046/j.1474-919X.2002.00110.x

Kokko H., 1999 Competition for early arrival in migratory birds. Journal of Animal Ecology 68: 940–950. https://doi.org/10.1046/j.1365-2656.1999.00343.x

Kullberg C., T. Fransson, and S. Jakobsson, 1996 Impaired predator evasion in fat blackcaps (*Sylvia atricapilla*). Proceedings: Biological Sciences 263: 1671–1675. https://doi-org.uplib.idm.oclc.org/10.1098/rspb.1996.0244

Lennox R. J., J. M. Chapman, C. M. Souliere, C. Tudorache, M. Wikelski, *et al.*, 2016 Conservation physiology of animal migration. Conserv Physiol 4. https://doi.org/10.1093/conphys/cov072

Liedvogel M., M. Szulkin, S. C. L. Knowles, M. J. Wood, and B. C. Sheldon, 2009 Phenotypic correlates of Clock gene variation in a wild blue tit population: evidence for a role in seasonal timing of reproduction. Molecular Ecology 18: 2444–2456. https://doi.org/10.1111/j.1365-294X.2009.04204.x

Liedvogel M., and B. C. Sheldon, 2010 Low variability and absence of phenotypic correlates of Clock gene variation in a great tit *Parus major* population. Journal of Avian Biology 41: 543–550. https://doi.org/10.1111/j.1600-048X.2010.05055.x

Lundberg P., 1987 Partial bird migration and evolutionarily stable strategies. Journal of Theoretical Biology 125: 351–360. https://doi.org/10.1016/S0022-5193(87)80067-X

Madon B., E. L. Nuz, C. Ferlat, and Y. Hingrat, 2015 Insights into the phenology of migration and survival of a long migrant land bird. bioRxiv 028597. https://doi.org/10.1101/028597

Moore F. R., and W. Yong, 1991 Evidence of food-based competition among passerine migrants during stopover. Behav Ecol Sociobiol 28: 85–90. https://doi.org/10.1007/BF00180984

Mueller J. C., F. Pulido, and B. Kempenaers, 2011 Identification of a gene associated with avian migratory behaviour. Proceedings: Biological Sciences 278: 2848–2856. https://doi-org.uplib.idm.oclc.org/10.1098/rspb.2010.2567

Newton I., 2007 *The Migration Ecology of Birds*. Elsevier.

Pakniat D., M.-R. Hemami, G. Shahnaseri, S. Maleki, M.-A. Adibi, *et al.*, 2020 The potential distribution of wintering and breeding populations of Asian Houbara Chlamydotis macqueenii in Iran. Bird Conservation International 1–15. https://doi.org/10.1017/S0959270920000167

Peterson M. P., M. Abolins-Abols, J. W. Atwell, R. J. Rice, B. Milá, *et al.*, 2013 Variation in candidate genes CLOCK and ADCYAP1 does not consistently predict differences in migratory behavior in the songbird genus Junco. F1000Res 2: 115. https://doi.org/10.12688/f1000research.2-115.v1

Piersma T., and Å. Lindström, 1997 Rapid reversible changes in organ size as a component of adaptive behaviour. Trends in Ecology & Evolution 12: 134–138. https://doi.org/10.1016/S0169-5347(97)01003-3

Pulido F., 2007 The genetics and evolution of avian migration. Bioscience; Oxford 57: 165–174. https://doi-org.uplib.idm.oclc.org/10.1641/B570211

Pulido F., 2011 Evolutionary genetics of partial migration – the threshold model of migration revisited. Oikos 120: 1776–1783. https://doi.org/10.1111/j.1600-0706.2011.19844.x

Riou S., J. Judas, M. Lawrence, S. Pole, and O. Combreau, 2011 A 10-year assessment of Asian Houbara Bustard populations: trends in Kazakhstan reveal important regional differences. Bird Conservation International 21: 134–141. https://doi.org/10.1017/S0959270910000377

Riou S., O. Combreau, J. Judas, M. Lawrence, M. S. Al Baidani, *et al.*, 2012 Genetic differentiation among migrant and resident populations of the threatened Asian houbara bustard. Journal of Heredity 103: 64–70. https://doi.org/10.1093/jhered/esr113

Saino N., G. Bazzi, E. Gatti, M. Caprioli, J. G. Cecere, *et al.*, 2015 Polymorphism at the Clock gene predicts phenology of long-distance migration in birds. Molecular Ecology 24: 1758–1773. https://doi.org/10.1111/mec.13159

Shaw A. K., 2016 Drivers of animal migration and implications in changing environments. Evolutionary Ecology 30: 991–1007. https://doi.org/10.1007/s10682-016-9860-5

Steinmeyer C., J. C. Mueller, and B. Kempenaers, 2009 Search for informative polymorphisms in candidate genes: clock genes and circadian behaviour in blue tits. Genetica 136: 109–117. https://doi.org/10.1007/s10709-008-9318-y

Wiltschko R., and W. Wiltschko, 2003 Avian navigation: from historical to modern concepts. Animal Behaviour 65: 257–272. https://doi.org/10.1006/anbe.2003.2054

Chapter 2

Building a genomic resource and investigating the genetics of migration in Asian houbara (*Chlamydotis macqueenii*) using a candidate gene approach

2.1 Abstract

Migration allows birds to adapt to environmental conditions by moving to lower latitudes in winter and therefore overcoming low resource availability at their breeding grounds. This movement is optimized by population-specific strategies regarding timing, direction and duration of migration. Variation in migratory behaviour can be a consequence of adaptations to the environment driven by endogenous genetic mechanisms, but much uncertainty exists around the potential influence of these mechanisms on the migration process.In Asian houbara, both resident and migrant individuals occur across the distribution range. This partial migrant species can give precious insights on the genetic mechanisms underlying migration. To address these evolutionary questions on Asian houbara, it is crucial to have a solid genomic foundation with a high-quality genome assembly and a robust annotation.

To explore the influence of genetic mechanisms on the migration of Asian houbara, we evaluated and compared the genomic resource already available for the species, we provided an improved annotation of its genome, and finally investigated genomic regions potentially involved in the evolution of migratory behaviour (using a candidate gene approach) using samples distributed across the whole distribution range and representing individuals with contrasted migratory behaviour.

We first evaluated the quality of the recent assembly and showed that it is 10-fold more contiguous (N50=2.88Mb) than the previous one and provides a completeness exceeding 96% based on the BUSCO protocol. Combining this genome assembly with four transcriptomic datasets (blood and embryos), we achieved a more complete annotation for Asian houbara reaching 17,992 genes, which is close to the 20,000 genes found in chicken. From the literature, we identified 14 candidate genes that have been linked to migration in birds, including the well-studied *CLOCK* and *ADCYAP1* genes. Mining the new genome assembly, we found 13 of these candidate gene regions. We used hierarchical groupings (AMOVA), to

evaluate the effect of migratory behaviour (migrant/resident) and geographic locations on the genetic variance. Sampling locations played a significant role on the genetic variation at seven candidate genes, and migratory behaviour had a significant effect on the variance of at least one gene (*DIO2*), highlighting its potential role in the migration of Asian houbara. The *DIO2* gene has been shown to play a significant role in the metabolism of long-distance migrants in Swainson's thrush. Creating haplotype networks and haplotype heatmaps with the coding sequences of 13 genes confirm the role of geography, and reveal the possible role of migratory behaviour in other genes. Further investigation of amino acid changes in the coding sequences also supports these findings. The improved genomic resource for Asian houbara provides us with the potential to further investigate genomic variation in this species. It also provides a reference that can be used in the genome assembly and annotation of closely related species. The candidate gene approach in the present study gives further insights into genes that are potentially linked to migration which opens perspectives for future investigations and the conservation of Asian houbara.

2.2 Introduction

The Asian houbara bustard (*Chlamydotis macqueenii*) is a partial migrant bird species with a large distribution range where migrant and resident populations co-occur. Individuals with resident behaviour spend their breeding and wintering seasons in the same region, in the wintering range of migrants (1.3). They occur in Iran, Pakistan and the Arabian Peninsula (Combreau et al. 2011a). Individuals identified as migrants have their breeding grounds ranging from northern Iran and Uzbekistan eastward up through Kazakhstan to the Gobi Desert in Mongolia and China. In autumn, these individuals move southward to spend winter in Iran, Iraq, Pakistan and the Arabian Peninsula (Combreau et al. 1999; Combreau et al. 2001; Judas et al. 2006) using three different flyways. The eastern flyway stretches from the Gobi desert through eastern and central Kazakhstan to Iran and Pakistan, with birds migrating as far South as the Kingdom of Saudi Arabia, Oman and India. The western flyway stretches from West Kazakhstan, through Iran to Iraq. The central flyway is in the middle of these two, stretching from central Kazakhstan and Uzbekistan to southern Iran (Combreau et al. 2011b).

From 1998 to 2002, dedicated population surveys have identified dramatic declines in spring and autumn abundance of Asian houbara in China, Kazakhstan and Oman, with an average annual decline of 27-30% (Tourenq et al. 2004; Tourenq et al. 2005). Similar trends were observed from 2000 to 2009, with declines as high as 93% in some regions. A study by Riou et al. (2001) also found regional differences in declines, and higher declines were observed in regions with higher accessibility where hunting and poaching is easier and more common. High densities of Asian houbara is still observed in vast undisturbed areas and regions with more frequent passage of migrants (Riou et al. 2011).

The overall population reduction was estimated to range between 30-45% over a period of three generations (more than a decade), which led the community to classify Asian houbara as vulnerable on the IUCN red list in 2014 (Bird Life International 2014). This survey indicates that the main cause of mortality is hunting and poaching, as well as degradation of suitable habitat in some areas. The large declines in abundance of Asian houbara creates conservation concern for this species (IFHC 2020). Captive-breeding and release programs form a major part of the conservation of Asian houbara, with approximately 170 000 Asian houbara individuals released into the wild since the establishment of these programs (L. Lesobre, internal communication). To release captive-bred individuals into the appropriate wild populations, it is important to correctly identify their migratory behaviour and assign them to the correct geographical region as individuals could have lower survival if they don't follow the behaviour of the rest of their population (Burnside et al. 2017).

Even though different populations of Asian houbara have drastically different migratory behaviour, the genetic differentiation between them is low but significant, even when using a combination of mitochondrial and microsatellite markers (Pitra et al. 2004; Riou et al. 2012; Haghani et al. 2018). These studies showed that the highest differentiation involved is between resident individuals from Egypt and Yemen and long-distance migrant individuals from all other locations. These results suggest limited or the absence of gene flow between resident and migrant individuals. A significant amount of genetic differentiation was also found between migrant individuals from West Kazakhstan and Mongolia, who have different flyways (Riou et al. 2012). Therefore, migratory behaviour (migrant vs. resident) seem to be the main factor that drive the population structure in Asian houbara.

A high-quality genome assembly and annotation provides a solid foundation to investigate genetic features

and phylogenetic information for non-model species (Zimin et al. 2009). This resource allows researchers to find and identify genes which have a functional or regulatory role on specific phenotypes of the species of interest, including their behaviour (Ejigu and Jung 2020). An accurate assembly and annotation also provides information about gene models and features such as alternative splice sites, which are highly beneficial in gene expression analysis (Zimin et al. 2009) and allows the investigation of genome structure and evolution (Abdellah et al. 2004). The current reference genome for Asian houbara (GenBank accession GCA_000695195.1; BioProject PRJNA212891) was assembled and annotated as part of a larger comparative genomics project involving avian genomes (Zhang et al. 2014) and submitted to the National Center for Biotechnology Information (NCBI) by the Beijing Genome Institute (BGI). More recently, a new assembly produced by GenoScreen has yet to be annotated and compared to the previous assembly.

When considering reference genomes, the chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*) genomes are the two most complete in birds (Knief and Forstmeier 2016) and they represent the state of the art in terms of genomic resource. The genomes of these species are assembled to chromosome level, with 39 and 37 assembled chromosomes available (on the NCBI genome database) for chicken and zebra finch respectively, including assembled sex chromosomes for both. In comparison, the current Asian houbara genome has 59,693 scaffolds which means it is far from a chromosome level assembly. Moreover, there are 27,036 and 21,481 gene models annotated for chicken and zebra finch respectively, compared to 13,996 for Asian houbara. From the six bustard genome assemblies available at the time of this project, an annotation was available only for the current Asian houbara reference genome.

Transcriptome data from the species of interest or a closely related species is usually required to generate a robust genome annotation. Before the release of another bustard species, the great bustard *Otis tarda*, transcriptome at the end of 2021 (Kuhl et al. 2021), no mRNA data was available for any bustard species, apart from the predicted mRNA from the current Asian houbara annotation. A good quality transcriptome assembly for Asian houbara would also be useful in the annotation of the 5 other bustard species genomes that are available on NCBI as well as for any future genome assemblies for the other 20 bustard species. Candidate gene approaches have been previously used to identify possible genetic influence on migration in birds (Fidler et al. 2007; Caprioli et al. 2012; Contina et al. 2018). These targeted studies focus on specific genes chosen according to their function or role in similar behaviour or species. This is a good approach for a pilot study before doing whole-genome analysis.

The aim of the present study is to improve the genome resource of the Asian houbara bustard with the purpose of investigating genomic regions potentially involved in the evolution of migratory behaviour of this species. First, we evaluated the quality of the new genome assembly for Asian houbara and compared it to the previous assembly. Then, we provided a new annotation of the Asian houbara genome and we compared it to the previous annotation to assess the quality of the new genome assembly. Finally, we searched for candidate genes previously identified as having a role in bird migration to investigate possible associations between variations at these genes, migratory behaviour and geographic locations.

2.3 Materials and Methods

2.3.1 Sample selection

2.3.1.1 Sample selection for genome assembly

We took two samples with the purpose of assembling their genomes. The first was taken from a wild male, collected as an egg, from Yemen (latitude 16.86447, longitude 52.02142) in April 2005. The genome from this sample is further referred to as REN_Cmacq_1.0 (Genbank accession: GCA_011799995.1). The second sample was taken from a captive-bred male whose ancestry can be completely traced back to founders collected in Pakistan. The genome from this sample is further referred to as REN_Cmacq_2.0 (GenBank accession: GCA_011800025.1). The two birds originate from populations composed of resident individuals.

2.3.1.2 Sample selection for transcriptome assembly

Blood samples and embryos were selected for transcriptome analysis. The blood samples were collected from two adults (Individual IDs: M06N07795 & M12N18705) and directly preserved in dry ice after being flash frozen. Based on previous studies that showed how embryonic samples can provide a very good transcriptome, we included two embryonic samples. We harvested two embryos from eggs of the 2021 cohort (Individual IDs: EM21N01538 & EM21N01566) that were directly put onto and kept on dry ice until the extraction step.

2.3.1.3 Sample selection for genomic investigation

We designed a strategy to select samples that represent the full distribution range and variation in migratory behaviour of Asian houbara. Starting with a large set of samples taken from wild Asian houbara individuals, we first chose locations across the distribution range and chose the best samples in each location.

We used the sampling coordinates to plot each sample on a map of the distribution range. We identified 10 geographical locations where samples occur in high concentrations without obvious geographical separation between them. We defined these locations by a range of latitude and longitude values. We grouped all of the available samples of wild Asian houbara into these locations and identified the migratory behaviour of each location based on satellite tracking data or known behaviour of the individuals from the location (Figure 2.1).

We chose 10 individuals for each location. We chose samples by prioritizing individuals with the most metadata (nest ID, mother ID and collection date, site and type). Our primary strategy was to choose all individuals with satellite tracking data, exclude those without additional metadata and exclude any related individuals (based on brood number and mother ID). (Figure 2.1). If a location did not have 10 individuals after this step, we followed a secondary strategy. In this case, we included individuals without tracking data, still excluding those without additional metadata and related individuals.

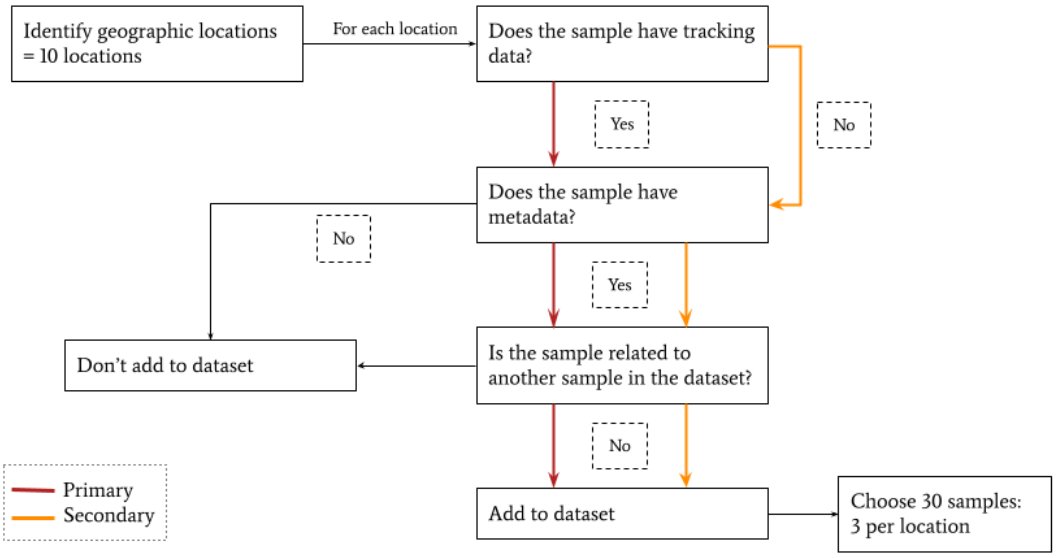


Figure 2.1: Flow diagram illustrating the sample selection strategy used for the population genomics investigation. This strategy was used to select 111 samples from 10 geographic locations. Metadata includes: nest ID, mother ID and collection date, site and type.

2.3.2 Laboratory processing of samples

2.3.2.1 Short-read sequencing for genome assembly

The samples selected to produce the reference genome were sequenced by GenoScreen on two Illumina runs at a read length of 250 bp using both paired-end (PE) and mate-pair (MP) reads with several insert sizes (Figure 2.2). The reads were cleaned by removing adapter sequences (Nextera Transposase) and undetermined bases which were located at both ends of the sequences. After this step, more than 80% of good-quality reads were retained in the initial dataset. The filtered dataset was then validated with FastQC (v0.11.9) (Andrews 2010) (Figure 2.2).

2.3.2.2 RNA sequencing for transcriptome assembly

Total RNA was extracted from the whole blood and embryo samples by BGI. They isolated the mRNA by treating the total RNA with DNaseI and did Oligo dT-based mRNA enrichment. They created cDNA from this mRNA using reverse transcription. Thereafter they added adapters to the ends of the mRNA fragments and performed PCR. They purified the PCR products with Ampuure XP Beads (AGENCOURT) and dissolved them in EB solution. They validated the library on the Agilent Technologies 2100 bioanalyzer. Finally, they circularized the double stranded PCR products which left a single strand circle DNA (ssCir DNA) as the final library.

The transcriptomes of the samples were sequenced by BGI on the DNBseq platform at a read-length of 150bp. The adapter sequences were removed and the reads were filtered to remove contamination and low-quality reads (base quality<Q15 in 40% of the read). The reads with more than 5% unknown bases (N's)

were also removed. We implemented BUSCO (v3.0.2) to measure the completeness of the transcriptome assemblies, using the *vertebrata_odb9* gene dataset.

2.3.2.3 Whole-genome sequencing

We sent 111 Asian houbara samples to BGI for whole-genome sequencing. These included blood on FTA cards, blood in ethanol and muscle in ethanol samples.

First, they tested the sample concentration, integrity and purity. They tested concentration with a fluorometer or Microplate Reader and sample integrity and purity with Agarose Gel electrophoresis (Concentration of Agarose Gel: 1%; Voltage: 150 V: Electrophoresis Time: 40 min). Thereafter they fragmented 1μg genomic DNA by Covaris Adaptive Focused Acoustics and selected the fragments of 200-400bp with the Agencourt AMPure XP-Medium kit. They performed end repair and 3' adenylation on the fragments and ligated adapters to the fragments.

They amplified the fragments with PCR and purified the PCR products with the Agencourt AMPure XP-Medium kit. The circularized the double stranded DNA and formatted the single strand circle DNA (ssCir DNA) as the final library. They validated the libraries with standard QC.

The validated libraries were sequenced at a read-length of 150bp on the BGISEQ-500 platform. For each geographic location, one sample was sequenced at 30× coverage whereas the rest of the samples were sequenced at 10× coverage. BGI also performed filtering for quality control, which includes removing adapter sequences, contamination and low-quality reads.

2.3.3 Genome and transcriptome assembly

2.3.3.1 Assembly process with MaSuRCA

GenoScreen used MaSuRCA (v3.4.2) (Zimin et al. 2013) to generate a *de novo* assembly (Figure 2.2). MaSuRCA is a genome assembling software based on a hybrid method which uses both paired-end (PE) and mate-pair (MP) reads to construct contigs. PE reads are used to create contigs, and MP reads are used to connect the contigs into larger scaffolds. MaSuRCA lowers the complexity of short-read data by constructing super-reads, which have a much lower coverage than the raw reads. Super-reads are PE reads which are uniquely extended on both sides. Multiple PE reads can form the same super-read, thus lowering the coverage. These super-reads are connected by the MP reads to form the assembly. MaSuRCA is a good assembler when working with large genomes and it usually provides large contigs and high contiguity (Zimin et al. 2013) (Sohn and Nam 2018).

The RMBC (Reads Mapped Back onto Contigs) were computed by aligning the reads along the scaffolds to evaluate the fraction of the total dataset represented in the assembly (Figure 2.2). This was done using BOWTIE2 (v2.4.1) (Langmead and Salzberg 2012) which only keeps the reads that align on their entire length without any gaps and whose respective pair also aligns.

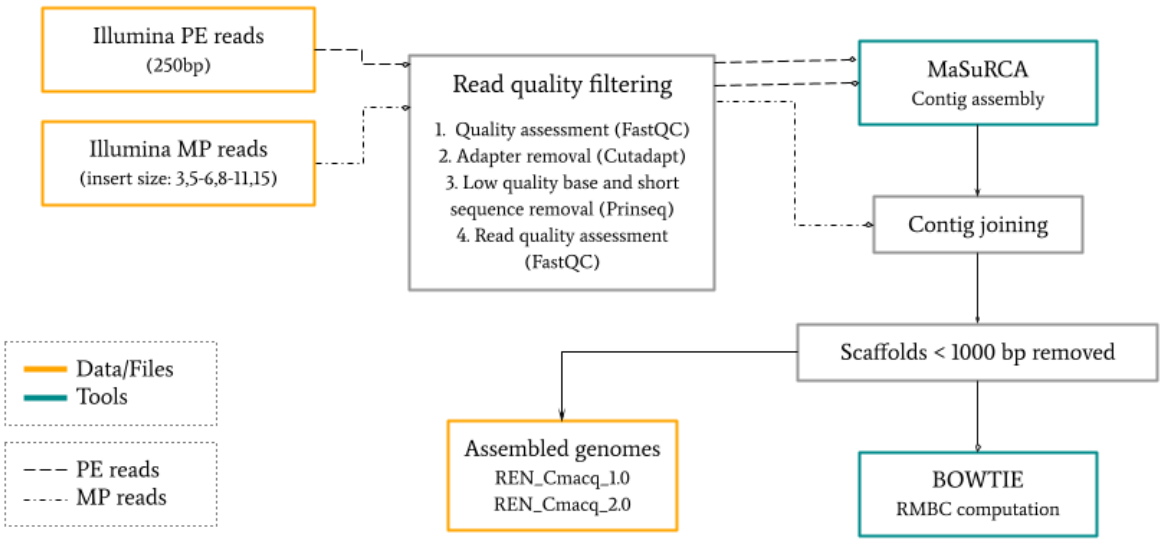


Figure 2.2: Flow diagram illustrating the steps for the new *de novo* assembly for Asian houbara. This pipeline was implemented by GenoScreen to assemble the genome.

After an investigation by GenoScreen, they found that they used the wrong insert size when they generated the assemblies. They then redid the assembly with MaSuRCA using the correct insert size.

2.3.3.2 Comparing genome assemblies

We compared the assemblies generated by GenoScreen (REN_Cmacq_1.0 and REN_Cmacq_2.0) to the reference assembly available on NCBI (ASM69519v1), using the number of contigs, N50, L50 and the number of mismatches as quality measures. We used Quast (v5.0.2) (Gurevich et al. 2013) to calculate these statistics and generated an N50 plot with Python. Fewer scaffolds indicate a better assembly because the genome has longer contiguous stretches which increase the ease of performing computational analyses on the genome. N50 is the smallest contig above which 50% of the assembly would be represented, meaning half of the genome sequence is covered by contigs larger or equal to the N50 size (Baker 2012). L50 is the minimum number of contigs needed to cover 50% of the genome. A larger N50 and smaller L50 usually indicate a better assembly, because it suggests a larger mean scaffold length.

We also measured the completeness of genome assemblies with BUSCO (v3.0.2) (Benchmarking Universal Single-Copy Orthologs) (Simão et al. 2015). BUSCO compares the assembly to a dataset of single-copy orthologs for a group of organisms, which is adjusted according to the study organism. The BUSCO dataset is used to quantify the completeness of a genome or transcriptome assembly by quantifying the proportions of 'complete', 'fragmented', 'missing' and 'duplicated' genes in the assembly. We used the *vertebrata_odb9* gene dataset for this test.

2.3.3.3 Transcriptome assembly

We generated a single *de novo* transcriptome assembly using the reads from all four samples using Trinity (v6.2.5) (Grabherr et al. 2011). We also mapped the reads to the reference genome using HISAT2 (v2.1.0) (Kim et al. 2019).

2.3.4 Genome annotation

2.3.4.1 Overview of MAKER2 pipeline

We applied the MAKER2 (v2.31.9)(Cantarel et al. 2008; Holt and Yandell 2011) genome annotation pipeline to complete the annotation of the Asian houbara genome (Figure 2.3). We performed the annotation on the assembly with the wrong insert size first and later repeated the annotation on the corrected assembly. MAKER2 integrates *ab initio* gene prediction tools, allowing annotation even when a reference genome for the species is not available. In the first step of the MAKER2 pipeline, empty config files *maker_exe*, *maker_opts* and *maker_bopts* are created. These files are edited to specify the paths to the input files and the tools used during the annotation. The input files include the FASTA files with the genomic sequence that needs to be annotated as well as the mRNA and protein evidence. The output from MAKER2 includes a GFF3 file and a FASTA file for annotated proteins and transcripts respectively. These output files are created for every scaffold and are concatenated after the initial MAKER2 run. MAKER2 can be run successively to improve the annotation accuracy by providing the output GFF3 file from the previous run as input for the following run.

MAKER2 starts the annotation process by masking repeats with RepeatMasker (http://repeatmasker.org), which screens the genome for low-complexity repeats, followed by BLASTX (Altschul et al. 1990), which identifies mobile elements. BLAST is also used to identify expressed sequence tags (ESTs), mRNAs and proteins that are significantly similar to the input genome sequence.

The MAKER2 output files are used to run Interproscan (Zdobnov and Apweiler 2001), a protein-domain finder. It is also used to run a BLAST homology report of the annotated proteins against the UniProt/Swiss-Prot database (Bairoch and Apweiler 2000). The results from MAKER2, Interproscan and the genome assembly are loaded into WebApollo (Lee et al. 2013), a web-based interface used to visually inspect and curate annotated genomes. Transcriptome data can also be added to aid in the manual curation of the annotated genes.

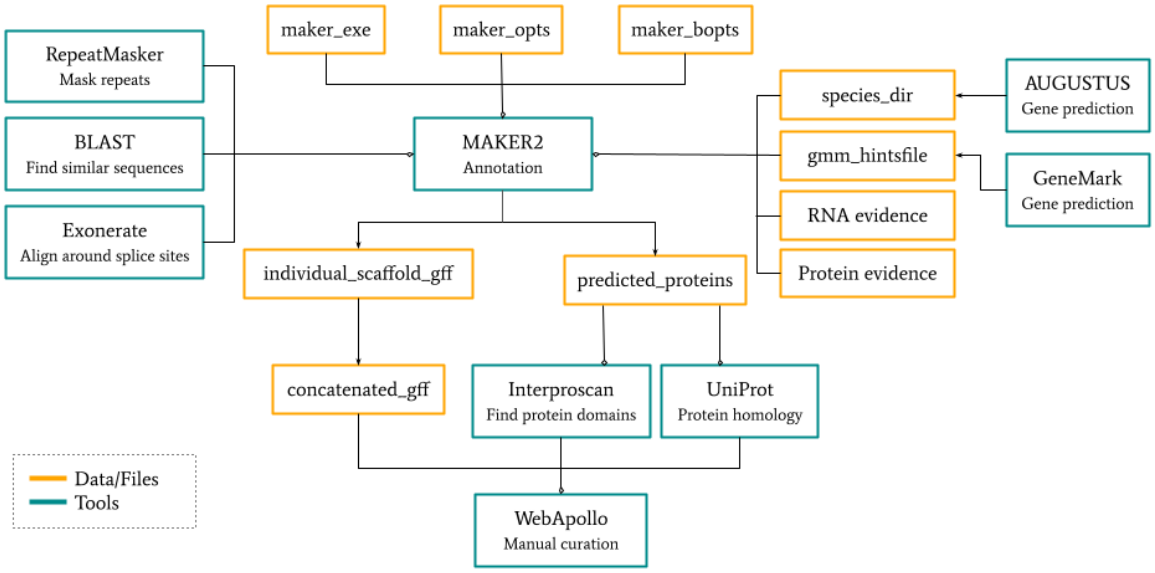


Figure 2.3: Flow diagram of the MAKER2 annotation pipeline. The details of the input files and tools are explained in the text.

2.3.4.2 Training Augustus with BRAKER

We implemented a BRAKER (v2.1.6) pipeline to add RNA evidence to the annotation of the Asian houbara genome, as this is lacking in the annotation of the current reference assembly. The pipeline starts with mapping the RNA-seq reads with HISAT2 (v2.1.0) (Kim et al. 2019) to generate BAM (binary alignment map) files which are used as the input for BRAKER. When RNA-seq alignments are used to train BRAKER, GeneMark-ET (Brůna et al. 2020) is executed. GeneMark-ET uses the RNA-seq alignments to inform BRAKER on where the splice sites are located. The set of predicted genes from GeneMark-ET is used by AUGUSTUS for training, after which it predicts genes using the RNA-seq alignments as evidence for intron position.

The output from this training includes the GeneMark hintsfile, AUGUSTUS Species directory and AUGUSTUS GFF3. These can be included as input in the MAKER2 config file and serve as RNA evidence in the genome annotation.

2.3.4.3 Evidence specified for MAKER2

We edited the MAKER2 config file *maker_opts* to include the GeneMark and AUGUSTUS output from BRAKER and the assembled transcriptome of Asian Houbara as RNA evidence (Figure 2.3).

Based on the quality of the previous annotation by Zhang et al. (2014) using only protein homology evidence, we decided to use the same sets of proteins that they used in their annotation. All of these gene sets were sourced from the Ensembl (Aken et al. 2016) database (release 60). This includes orthologous gene pairs of chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*), non-orthologous gene sets from chicken and zebra finch, and the complete human gene set.

We enabled the est2genome and protein2genome options, which predict genes and proteins directly from the evidence provided. We also enabled the option to include quality scores to have the AED scores for each gene prediction in the MAKER2 output.

We ran Interproscan (Jones et al. 2014) using the protein prediction FASTA file as input. We also ran the MAKER2 post-processing scripts to include the Interproscan results in the GFF3 file.

2.3.4.4 Assessing the quality of genome annotation

Sensitivity, specificity and accuracy are common metrics used to measure annotation quality when a high-quality reference genome is available (Holt and Yandell 2011) (Table 1). Sensitivity refers to the proportion of a reference that overlaps a prediction (Figure 2.4) and is calculated as the number of overlapping nucleotides between the reference and the prediction $[i \cap j]$ divided by the total number of nucleotides in the reference $[j]$. Specificity refers to the proportion of a prediction that overlaps a reference (Figure 2.4) and is calculated as the number of overlapping nucleotides $[i \cap j]$ divided by the total number of nucleotides in the prediction $[i]$. Accuracy is the average of sensitivity and specificity (Table 2.1).

Figure 2.4: Basic parameters used in the calculation of sensitivity and specificity, where j = the total number of nucleotides in the reference and i = the total number of nucleotides in the prediction.

When a high-quality reference genome is not available the reference is approximated with a cluster of experimental evidence aligned to the genome (Holt and Yandell 2011). The overlap $[i \cap j]$ is then measured as the number of nucleotides in the prediction that overlaps with the experimental evidence. In this case, the accuracy is referred to as congruency and the distance between the prediction $[i]$ and the experimental evidence $[j]$ is the incongruency (Table 2.1). This distance is also called the Annotation Edit Distance (AED). An AED score of 0 indicates complete agreement between the annotation and the evidence, while an AED score of 1 indicates complete disagreement. To measure the performance of different genome annotation strategies, we measured the AED scores.

Table 2.1: Formulas used to calculate parameters of annotation quality

Metric	Formula
Sensitivity	$SN = \|i \cap j\|/\|j\|$
Specificity	$SP = \|i \cap j\|/\|i\|$
Accuracy/ Congruency	$C = (SN+SP)/2$
Incongruency/ AED	$D = 1\text{-}C$

Variables used in formulas: SN = Sensitivity; SP = Specificity; j = the total number of nucleotides in the reference; i = the total number of nucleotides in the prediction; C = Congruency; D = Incongruency. AED=Annotation Edit Distance (agreement between prediction and evidence).

2.3.5 Candidate gene identification

2.3.5.1 Identification of genes from the literature

We conducted a literature search to find genes previously linked to seasonal migration in birds. We used keyword searches with wildcards (e.g. "BIRD + MIGRAT* + GENE*") on databases such as Web of Science to gather relevant papers. We chose the most recent and highest cited papers as a starting point for the literature search. We also followed citations within papers to find additional studies on the subject. We ignored all of the genes that were investigated but did not show a significant link to migratory behaviour in birds.

2.3.5.2 Mining the candidate genes in the new Asian houbara reference genome

We used the output from the annotation with MAKER2 to confirm the presence of the candidate genes in the Asian houbara genome assembly. For each candidate gene, we retrieved the UniProt (The UniProt Consortium 2021) and Ensembl (Aken et al. 2016) codes from the GeneCard (Stelzer et al. 2016) database. These codes are used in the GFF3 file from MAKER2 which indicates gene and protein predictions. For each prediction, the scaffold number and position of the prediction on the scaffold are also included in the GFF3 file. The GFF3 is therefore searchable using the codes to find the positions of the genes in the reference assembly.

We used the codes for the candidate genes to extract the positional information of the genes from the GFF3 file (using custom Python scripts). We extracted the sequences of each gene from all genomes analysed and created an alignment for each gene. This alignment included the following sequences:

1. *C. macqueenii* Ensembl prediction (REN_Cmacq_1.0 assembly)
2. *C. macqueenii* UniProt prediction (REN_Cmacq_1.0 assembly)
3. *C. macqueenii* Ensembl prediction (current reference assembly, ASM69519v1)
4. *G. gallus* reference (From NCBI: GRCg6a, Genome Reference Consortium Chicken Build 6a)
5. *C. macqueenii* mapped reference (REN_Cmacq_1.0 assembly)

Using these alignments, we chose the best predictions based on which covered the highest number of candidate genes and how they compared with the *G. gallus* and *C. macqueenii* reference genes. We also used the annotation to assess the quality of the genome assembly by investigating the candidate gene regions.

2.3.6 Processing of genomic data

2.3.6.1 Pipeline overview

We designed a pipeline that provides two haplotypes for each gene for every individual in our dataset. The pipeline is implemented sequentially. Firstly, we create a reference gene for Asian houbara using a

reference from *G. gallus* and high-coverage Asian houbara reads (from REN_Cmacq_1.0; wild male from Yemen; resident). Secondly, we use the new reference gene for Asian houbara to map the reads from all individuals from our dataset to in order to create two haplotypes for each individual (Figure 2.5). We ran our pipeline using short-read data from 30 Asian houbara individuals representing 10 geographic locations.

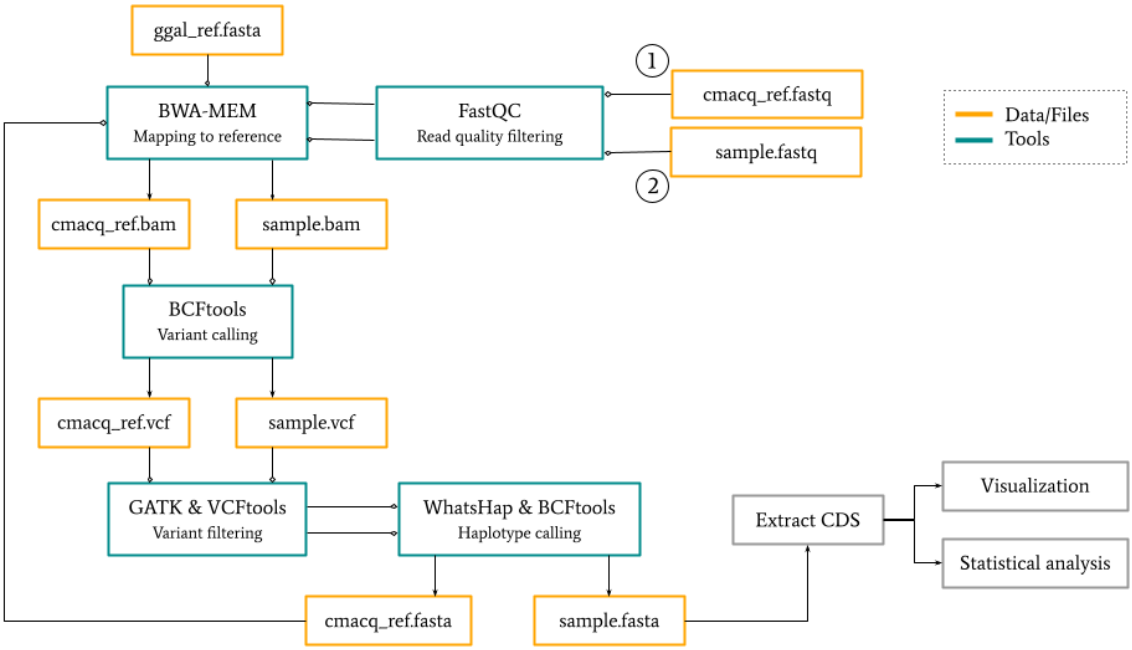


Figure 2.5: Flow diagram illustrating the different steps followed for generating the coding sequence (CDS) for each *Chlamydotis macqueenii* individual and each candidate gene. [1] The pipeline was first followed with the read data for the *C. macqueenii* reference (cmacq_ref.fastq) to create a reference haplotype sequence for *C. macqueenii* (cmacq_ref.fasta). [2] This reference was used in the mapping step to create haplotype sequences for the individuals in the dataset (sample.fasta).

2.3.6.2 Mapping and variant calling

We mapped the reads to the reference with BWA-MEM (Li and Durbin 2009), which outputs a BAM file (binary alignment map). We sorted the BAM file with SAMtools (Danecek et al. 2021) and removed the duplicates using Picard tools (http://broadinstitute.github.io/picard/). We called the variants with BCFtools (Danecek et al. 2021)(Figure 2.5).

2.3.6.3 Variant filtration (criteria selection and implementation)

We chose the parameters for variant filtering based on common use in literature. We used read depth, mapping quality and strand bias to filter the variants for the individuals. We chose the cut-off values of these parameters by assessing the distribution of each parameter for each gene, therefore having a set of parameters for each candidate gene. We filtered the variants for each individual according to these values. After this, we applied a filter for minimum allele count (mac=3) across all individuals for each gene. This step ensures that no singletons are present in the final haplotypes. We used GATK VariantFiltration (McKenna et al. 2010), and VCFtools (Danecek et al. 2011) to perform these filtering steps.

2.3.6.4 Haplotype calling and phasing

We used WhatsHap (Martin et al. 2016) to phase the haplotypes and BCFtools (Danecek et al. 2021) to call the haplotypes for each individual (Figure 2.5). WhatsHap infers the correct relationship (cis or trans) between alleles at heterozygous loci using a read-base phasing method (Martin et al. 2016). It requires the BAM and VCF files of each individual and the reference genome used in mapping as input. We ran WhatsHap for each candidate gene to end up with one FASTA file for each gene containing both haplotypes for each individual.

2.3.7 Genomic analyses

We decided to analyze only the coding sequences (CDS) of the candidate genes. Our haplotype calling pipeline outputs FASTA files with the haplotypes for the complete candidate genes, so we needed to extract the CDS for each gene from these FASTA files. As we did in the mapping step of our pipeline, we use *G. gallus* as a reference to identify the CDS of each gene. We retrieved the information of the positions of the exons for each *G. gallus* gene from the NCBI database. For each gene, we extracted the CDS from our haplotype FASTA files into a new FASTA file. This new FASTA file was used in all downstream analyses.

2.3.7.1 Standard statistics

We calculated standard statistics of the haplotype CDS sequences using the R package 'pegas' (v1.1) (Paradis 2010). This includes the nucleotide and haplotype diversities for each candidate gene.AMOVA in R

We ran an AMOVA (analysis of molecular variance) with the haplotype CDS sequences of each gene using the R package 'poppr' (v2.9.3) (Kamvar et al. 2014; Kamvar et al. 2015) and did a Randomization test to test for significance.

2.3.7.2 Haplotype heatmaps

We created haplotype heatmaps for each candidate gene CDS to represent the variant sites in the sequences. We used a custom Python script and Matplotlib heatmap function (Hunter 2007) to visualize the heatmaps.

2.3.7.3 Haplotype networks in PopArt and HaploViewer

We used PopArt (Leigh and Bryant 2015) to draw both minimum spanning and median-joining networks for each candidate gene using the CDS sequences. We used these networks to identify patterns in haplotype frequency between the geographic locations.

We also created haplotype networks with the amino acid sequences of the candidate genes using Haplotype Viewer (http://www.cibiv.at/~greg/haploviewer). We used RAxML (v8.2.12) (Stamatakis 2014) with the PROTGAMMADAYHOFF model of substitution to create trees with the amino acid sequences. Haplotype Viewer uses the amino acid sequences together with these trees to draw haplotype networks.

2.3.7.4 Principal component analyses

Using the same pipeline as for the candidate gene analysis, we did variant calling for the complete genomes of the 30 selected Asian houbara individuals. We performed a principle component analysis (PCA) with this whole-genome information. We used PLINK (v1.9) (http://pngu.mgh.harvard.edu/purcell/plink) (Purcell et al. 2007) to do linkage pruning and calculate the eigenvectors and eigenvalues. We visualized the PCA using Matplotlib scatterplots.

2.4 Results

2.4.1 Genome assembly

The assemblies REN_Cmacq_1.0 and REN_Cmacq_2.0 were compared to the previous reference assembly (ASM69519v1) obtained from the NCBI Genome database (Table 2.2). The REN_Cmacq_1.0 and REN_Cmacq_2.0 assemblies are 103.6 and 83.9 Mb longer than ASM69519v1, respectively. these assemblies were also sequenced at a higher coverage with 114× and 96× for REN_Cmacq_1.0 and REN_Cmacq_2.0 respectively, compared to 27× for ASM69519v1. Compared to ASM69519v1, the new assemblies have a lower number of scaffolds and approximately a 100-fold longer N50 length (Figure 2.6). Additionally, the lower L50 indicates that a smaller number of scaffolds were needed to cover 50% of the genome (132 for REN_Cmacq_1.0 and 161 for REN_Cmacq_2.0 compared to 6,641 for ASM69519v1).

One problem with genome assembly is that contiguity can be falsely increased by using unassigned bases (N's, also called mismatches) to connect contigs into scaffolds. This leads to a pseudo-improvement of the N50, L50 and scaffold number, but increases the number of unassigned bases (N's) in the assembly. We ran a QUAST (Gurevich et al. 2013) assessment on the assembled genomes that we received from GenoScreen. This showed that REN_Cmacq_1.0 and REN_Cmacq_2.0 had 12,715 and 5,515 N's per 100 kbp respectively. This means 12% and 5% of the bases in the respective assemblies were unassigned. After redoing the assembly with the correct insert size, the number of N's per 100 kbp was 2,352.13 and 3,225.13 in REN_Cmacq_1.0 and REN_Cmacq_2.0 respectively (Table 2.2). Even though this is still much higher than in ASM69519v1 (271.98 N's per 100 kbp), it is much better than the assemblies generated with the incorrect insert size.

Table 2.2: Genome assembly summary for *Chlamydotis macqueenii* genomes

Assembly	ASM69519v1	REN_Cmacq_1.0	REN_Cmacq_2.0
GenBank accession	GCA_000695195.1	GCA_011799995.1[a]	GCA_011800025.1[a]
Total length	1,086.6 Mb	1,190.2 Mb	1,170.5 Mb
Coverage	27×	114×	96×
Number of scaffolds	59,693	2,667	2,839
N50	0.046 Mb	2.88 Mb	2.4 Mb
L50	6,641	132	161
#N's per 100 kbp	271.98	2,352.44	3,225.13

[a]The genomes available on GenBank are previous versions where the wrong insert size was used during the assembly. The genomes used in this study are those with the correct insert sizes.

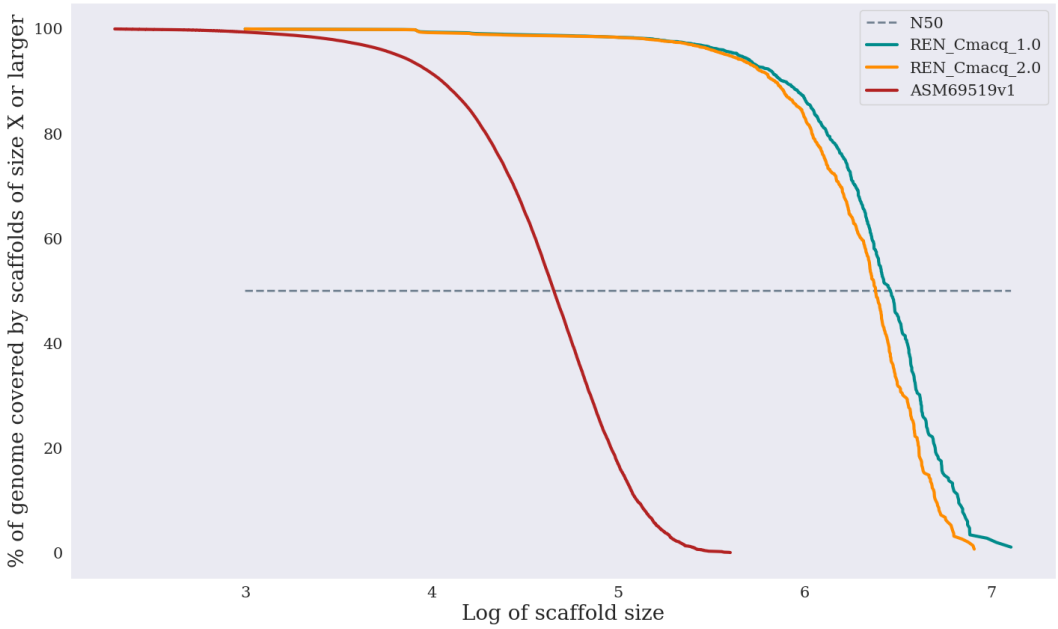


Figure 2.6: Comparison of scaffolding contiguity for the different assemblies of *Chlamydotis macqueenii* using cumulative plots of the N-statistic. N50: REN_Cmacq_1.0 = 0.046Mb, REN_Cmacq_2.0 = 2.88Mb, ASM69519v1 = 2.4Mb.

We used BUSCO (Simão et al. 2015) to estimate the completeness of the assemblies, using a set of core genes of vertebrates (*vertebrata_odb9* dataset). The set comprises a total of 2,586 genes which could be either missing, fragmented, complete and duplicated or complete and single-copy in the assembly of interest. In general, a more complete, non-duplicated and non-fragmented assembly is preferred. Both the new assemblies showed a higher fraction of complete single copy genes (REN_Cmacq_1.0: 97.2%; REN_Cmacq_2.0:

96.2%) compared to the previous reference (ASM69519v1: 78.7%) (Figure 2.7). All three assemblies have low numbers of duplicated genes (REN_Cmacq_1.0: 8; REN_Cmacq_2.0: 4; ASM69519v1: 6). The two new assemblies showed fewer fragmented genes compared to the previous reference (REN_Cmacq_1.0: 1.7%; REN_Cmacq_2.0: 2.5%; ASM69519v1: 16.4%). Finally, the new assemblies showed fewer missing genes than the previous reference (REN_Cmacq_1.0: 0.8%; REN_Cmacq_2.0: 1.2%; ASM69519v1: 4.7%) (Figure 2.7).

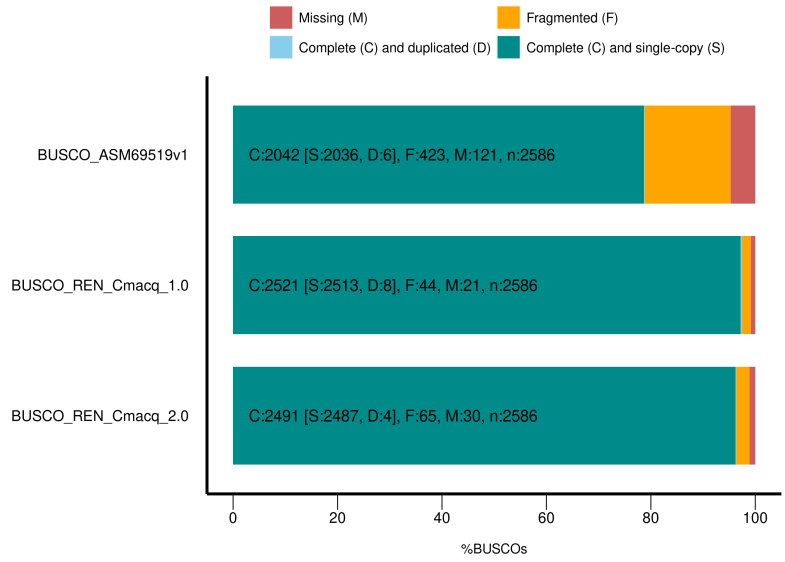


Figure 2.7: Results of BUSCO analyses of the genome assemblies for *Chlamydotis macqueenii* illustrating their completeness based on the vertebrate gene set.

Overall, REN_Cmacq_1.0 and REN_Cmacq_2.0 had higher completeness (97.49% and 96.33%) than ASM69519v1 (78.96%) when taking into account both single-copy and duplicated complete genes. We chose REN_Cmacq_1.0 as the assembly to annotate in the next step of the project. REN_Cmacq_1.0 had higher completeness, fewer scaffolds and higher N50 compared to REN_Cmacq_2.0. This higher contiguity and completeness are important factors in genome annotation quality, which is why we chose this assembly.

2.4.2 Transcriptome assembly

We sent two blood (M06N07795A & M12N18705A) and two embryo (EM21N01538A & EM21N01566A) samples for sequencing. The transcriptomes yielded between 200,000 and 340,000 transcripts (Table 2.3). A BUSCO (Simão et al. 2015) analysis based on the same vertebrate gene set as the genome assemblies above (*vertebrata_odb9* dataset), shows that the transcriptomes have completeness ranging from 58% to 90%, but most of the complete genes are duplicated (Figure 2.8). The transcriptomes of the embryo samples are overall more complete with less fragmented or missing genes than the transcriptomes of the blood samples.

Table 2.3: Comparison of the assembly statistics for the different transcriptomes of *Chlamydotis macqueenii*

Sample ID	Tissue	Number of transcripts	Completeness (%)[a]
EM21N01538A	Embryo	212,009	89.1
EM21N01566A	Embryo	211,472	87.23
M06N07795A	Blood	202,173	68.14
M12N18705A	Blood	343,405	58.39

aThe percentage completeness was estimated using BUSCO with the vertebrate gene set.

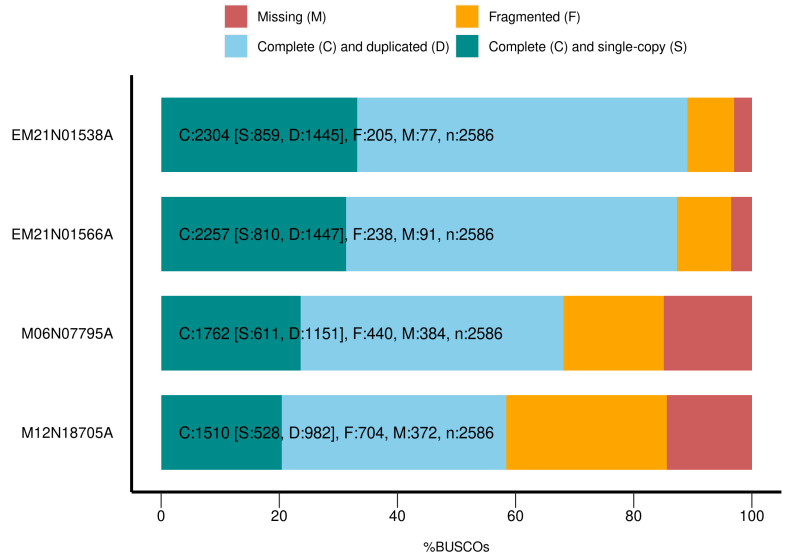


Figure 2.8: Results of BUSCO analyses of the transcriptome assemblies for *Chlamydotis macqueenii* illustrating their completeness using the vertebrate gene set. EM21N01538A and EM21N01566A are embryo samples and M06N07795A and M12N1875A are blood samples.

To test whether there were genes in the transcriptomes of some samples that were missing or fragmented in others, we made a Venn diagram of the complete genes in each transcriptome (Figure 2.9). A total of 1,188 genes were present and complete in all four transcriptomes, making up 45.94% of the vertebrate gene set. A total of 137 genes were present in only one transcriptome, with the embryonic transcriptomes both having more unique genes than the blood transcriptomes. If we combine all of the complete genes present in all four transcriptomes, we have 2,472 complete genes in total, making up 95.59% of the vertebrate gene set. By combining these transcriptomes into one assembly we would therefore have a higher number of genes covered completely than if we only chose the single most complete transcriptome (EM21N01538A).

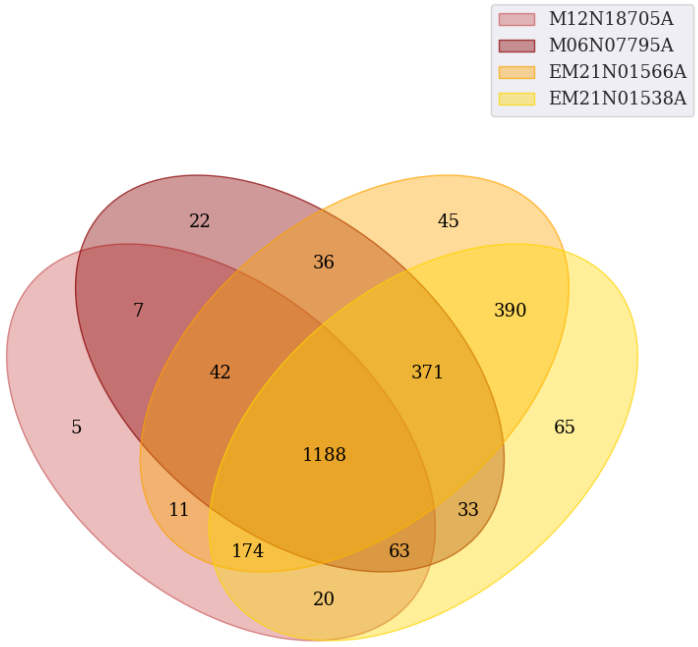


Figure 2.9: Venn diagram illustrating the distribution of complete genes across the four transcriptomes. EM21N01538A and EM21N01566A are embryo samples and M06N07795A and M12N1875A are blood samples.

2.4.3 Genome annotation

We performed annotation of the REN_Cmacq_1.0 assembly with MAKER2 (Holt and Yandell 2011). We did the annotation on the assembly with the wrong insert size using only the UniProt database (The UniProt Consortium 2021) for protein evidence. For the annotation with the correct assembly,we used both the UniProt database and Ensembl (Aken et al. 2016) gene sets from chicken, zebra finch and human in the annotation of the correct assembly. In both annotations, we included the combined transcriptome assembly from the blood and embryo samples as RNA evidence in the annotation. We also trained gene predictors with BRAKER (Brůna et al. 2021) and added the results as additional evidence in the annotation. We optimized our genome annotation by doing multiple rounds of annotation with our improved genome assembly (REN_Cmacq_1.0). We used various combinations of RNA datasets and protein evidence for the different rounds. We did sequential rounds of annotation, using the outputs (MAKER2 GFF) from previous rounds for MAKER2 in following rounds. This enables MAKER2 to be more accurate in making gene predictions.

One way to measure the quality of genome annotation is with the annotation edit distance (AED). The annotation of REN_Cmacq_1.0 had lower overall AED scores than that of ASM69519v1 (Figure 2.10). The AED scores for the correct REN_Cmacq_1.0 assembly are higher than for the annotation of the misassembly. This is an indication that the improvement of the assembly increased the AED scores of the annotated predictions. The distribution of the AED scores shows that we could not improve on the previous reference annotation when we only take into account this measure.

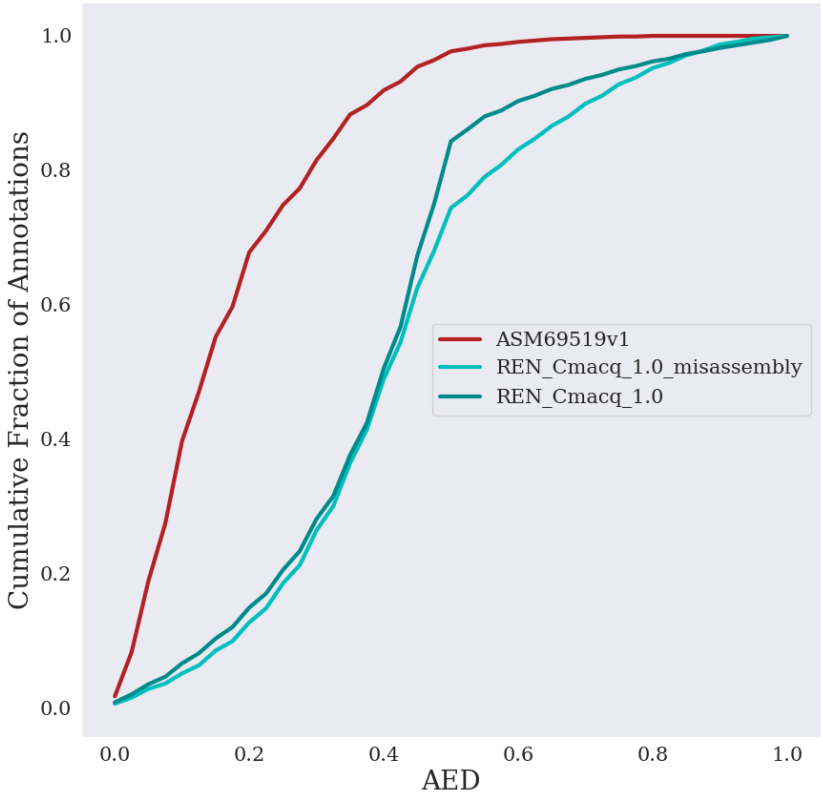


Figure 2.10: Cumulative annotation edit distance (AED) for all predicted genes obtained from each full round of MAKER2 annotation.REN_Cmacq_1.0_misassembly was assembled using the wrong insert size.

The number of gene and protein predictions is also a measure used to assess the quality of genome annotation. We compared the number of predicted genes, proteins and expressed sequence tags (ESTs) from the annotation of the new assembly to that of the previous reference (Table 2.4). The number of genes increased by 28.6% and the number of predicted proteins increased by 112.3% (Table 2.4). By adding RNA evidence to the annotation, we enabled the prediction of ESTs, which were not present in the previous annotation.

Table 2.4: Number of predicted gene models from each full round of MAKER2 annotation

Assembly	Gene number	Protein number	EST number (est2genome)	EST number (blastn)
ASM69519v1	13,996	278,170	N.A.	N.A.
REN_Cmacq_1.0_misassembly	20,860	591,427	5,744,650	5,193,797
REN_Cmacq_1.0	17,992	590,512	5,888,282	5,232,977

EST = expressed sequence tag. ESTs are either predicted directly from RNA evidence (est2genome) or compared to the BLAST nucleotide database and predicted by similarity (blastn).

2.4.4 Candidate gene identification

We identified 14 candidate genes from the literature which have been previously linked to migratory behaviour in birds (Table 2.5). They are *ADCY8*, *ADCYAP1*, *CLOCK*, *CRABP1*, *DIO2*, *DRD4*, *FECH*, *HERC4*, *MLNR*, *RAD51D*, *ST6GALNAC2*, *TGFBR1*, *TOP2B* and *TSPO2*. A significant association between migratory behaviour and these genes were found in at least one bird species. *CLOCK* (circadian locomotor output cycles kaput protein) is the gene that has been most often linked to migratory behaviour in birds (Bazzi et al. 2015; Bazzi et al. 2016; Bourret and Garant 2015; Caprioli et al. 2012; Johnsen et al. 2007; Saino et al. 2015).

The *CLOCK* gene encodes a transcription factor which is involved in interactions that are central to the regulation of circadian rhythms (according to the NCBI Entrez database). *ADCYAP1* (adenylate cyclase activating polypeptide 1) has also been studied extensively and linked to migration in birds in more than one study (Bourret and Garant 2015; Mueller et al. 2011). This gene encodes multiple peptides which are involved in the generation of cAMP (cyclic adenosine monophosphate), which leads to transcriptional activation of target genes (according to the NCBI Entrez database). Functionally, it influences neuroendocrine stress responses. Similarly, *ADCY8* (adenylate cyclase 8) encodes part of an enzyme that catalyzes the formation of cAMP from ATP (adenosine triphosphate) (according to the NCBI Entrez database). This gene is central in hormone and brain function and has been linked to long-term memory (The UniProt Consortium 2021).

Table 2.5: Genes found to be significantly linked to migratory behaviour in at least one bird species.

Gene	Species	Findings	Reference
CLOCK (Circadian Locomotor Output Cycles Kaput Protein)	Blue tit (*Cyanistes caeruleus*)	Evidence of latitudinal clines	(Johnsen et al. 2007)
	Barn swallow (*Hirundo rustica*)	Varying mean breeding date according to genotype in yearling females.	(Caprioli et al. 2012)
	Barn swallow	Rare genotype Q7/Q8 shows delayed timing of autumn migration.	(Bazzi et al. 2015)
	Tree swallow (*Tachycineta bicolor*)	Positive relationship between laying date and female genotypes in interaction with breeding diversity.	(Bourret and Garant 2015)

	Four trans-Saharan species: *Luscinia megarynchos, Ficedula hypoleuca, Anthus trivialis, Saxicola rubetra*	Individuals with more glutamine residues in the poly-Q region of the gene migrated later depending on sex and whether the within-individual mean length or the length of the longer allele was considered.	(Saino et al. 2015)
	23 Afro-Palearctic migratory bird species	Gene diversity covaried with several migration traits. Allele size increase with latitude of breeding range.	(Bazzi et al. 2016)
ADCYAP1 (Adenylate Cyclase Activating Polypeptide 1)	European blackcaps (*Sylvia atricapilla*)	Association between some genotypes and migratory restlessness.	(Mueller et al. 2011)
	Tree swallow	Relationship between laying date and female genotypes in interaction with latitude.	(Bourret and Garant 2015)
ADCY8 (Adenylate Cyclase 8)	Peregrine falcons (*Falco peregrinus*)	Significantly associated with a selective sweep. Associated with better long-term memory.	(Gu et al. 2021)
DRD4 (Dopamine Receptor D4)	Great tit (*Parus major*)	Three genotypes were associated with varying levels of early exploratory behaviour.	(Fidler et al. 2007)
CRABP1 (Cellular Retinoic Acid Binding Protein 1)	Swainson's thrush (*Catharus ustulatus*)	Gene upregulated in migratory birds.	(Johnston et al. 2016)
DIO2 (Iodothyronine Deiodinase 2)	Swainson's thrush	Gene downregulated in migratory birds.	(Johnston et al. 2016)
MLNR (Motilin Receptor)	European blackbirds (*Turdus merula*)	Differential expression in residents and migrants. Overexpressed in migrant group.	(Franchini et al. 2017)
TOP2B (DNA Topoisomerase II Beta)	European blackbirds	Differential expression in residents and migrants. Overexpressed in migrant group.	(Franchini et al. 2017)
ST6GALNAC2 (Siyaltransferase)	European blackbirds	Differential expression in residents and migrants. Overexpressed in migrant group.	(Franchini et al. 2017)
TGFBR1 (TGF-Beta Receptor)	European blackbirds	Differential expression in residents and migrants.	(Franchini et al. 2017)

		Overexpressed in migrant group.	
HERC4 (Ubiquitin-Protein Ligase)	European blackbirds	Differentially expressed in residents vs. short-term migrants and short-term vs. long-term migrants.	(Franchini et al. 2017)
RAD51D (RAD51 Paralog D)	European blackbirds	Differentially expressed in residents vs. short-term migrants and short-term vs. long-term migrants.	(Franchini et al. 2017)
TSPO2 (Translocator Protein)	European blackbirds	Differentially expressed in residents vs. short-term migrants and short-term vs. long-term migrants.	(Franchini et al. 2017)
FECH (Ferrochetelase)	European blackbirds	Differentially expressed in residents vs. short-term migrants and short-term vs. long-term migrants.	(Franchini et al. 2017)

DRD4 (dopamine receptor 4) encodes a subtype of the receptor for the neurotransmitter dopamine (NCBI Entrez). This receptor is involved in the regulation of emotion and complex behaviour, but has also been linked to circadian rhythm modulation (The UniProt Consortium 2021). *CRABP1* (cellular retinoic acid binding protein) encodes a binding protein for vitamin A and is important for processes that involve retinoic acid-mediated differentiation and proliferation (NCBI Entrez). *DIO2* (Iodothyronine Deiodinase 2) catalyses the conversion of the thyroid hormone T4 to the active hormone T3 (NCBI Entrez). Thyroid hormones are crucial in growth and development and ultimately regulate many crucial body functions such as digestion, temperature regulation and muscle contraction. *MLNR* (motilin receptor) encodes an enzyme which is involved in gastrointestinal motility (Tocris). *TOP2B* (DNA Topoisomerase II Beta) encodes the enzyme that is responsible for breaking and joining double-stranded DNA during transcription (NCBI Entrez). *ST6GALNAC2* (ST6 N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 2) encodes a protein that is involved in protein metabolism and the termination of O-glycan biosynthesis (Stelzer et al. 2016). *TGFBR1* (Transforming Growth Factor Beta Receptor 1) encodes a receptor protein which is involved in the regulation of many cellular processes (The UniProt Consortium 2021). *HERC4* (HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 4) encodes a protein for which the function is not yet confirmed. *RAD51D* (RAD51 Paralog D) is involved in DNA repair (The UniProt Consortium 2021). *TSPO2* (Translocator Protein 2) is predicted to be involved in cholesterol metabolism (NCBI Entrez). *FECH* (Ferrochelatase) encodes a protein that is localized in the mitochondrion and is involved in the heme synthesis pathway (NCBI Entrez).

We used the GeneCard database (Stelzer et al. 2016) to get the UniProt (The UniProt Consortium 2021) and Ensembl (Aken et al. 2016) codes for these 14 genes. We conducted a search of the GFF3 files which were generated by MAKER2, containing the gene, mRNA and protein predictions from the annotation. There were differences in the number of genes that are present in the annotation of the different assemblies (Table 2.6). There were also differences in the number of genes annotated with the UniProt and Ensembl evidence. Overall, the new assembly (REN_Cmacq_1.0) yielded more of the candidate genes in its annotation than the previous reference. Only 7 of the candidate genes could be found in the previous reference assembly, compared to 13 in REN_Cmacq_1.0 using the same protein evidence (Table 2.6).

Table 2.6: Number of candidate genes recovered from the different MAKER2 annotation predictions.

Genome assembly	Database	Number of candidate genes
ASM69519v1	Ensembl	7
REN_Cmacq_1.0 (misassembly)	UniProt	8
REN_Cmacq_1.0	Ensembl	13
REN_Cmacq_1.0	UniProt	10

The REN_Cmacq_1.0 assembly generated with the incorrect insert size only yielded 8 of the candidate genes, compared to 10 in the correct assembly using the same protein evidence (Table 2.6). Overall, the Ensembl gene sets led to more predicted candidate genes present in the annotation than the UniProt database. With Ensembl, 13 of the 14 chosen candidate genes were present in the annotation. *TSPO2* was the only candidate gene that we could not find in the annotation across the different predictions, therefore we excluded it from all downstream analyses.

Even though the candidate genes were found in the assembly, analysis of the predicted sequences showed that they are only partially covered (Table 2.7). We extracted the predicted gene sequences from the assembly and aligned them to the chicken reference genes. We used the chicken gene size and calculated the percentage of the genes that is covered in our assembly. These percentages range from 5.71% for *TGFBR1* to 87.40% for *TOP2B*. On average, the genes were 46.42% covered in the REN_Cmacq_1.0 assembly. Due to the limited potential of the assembly as a reference for our candidate gene analysis, we chose to use the complete chicken genes as a reference in our pipeline.

Table 2.7: Coverage of 13 candidate genes from annotation predictions using the Ensembl database and the REN_Cmacq_1.0 assembly.

Gene	Expected size in *G. gallus* (bp)	Total covered (bp)	Percentage covered
ADCY8	5064	2276	44.9
ADCYAP1	757	523	69.1
CLOCK	3560	2504	70.3
CRABP1	746	249	33.4
DIO2	6090	616	10.1
DRD4	1316	744	56.1
FECH	1500	949	63.2
HERC4	6858	2957	43.1
MLNR	1663	693	41.4
RAD51D	1456	823	56.5
ST6GALNAC2	2167	488	22.2
TGFBR1	5781	330	5.7
TOP2B	5477	4787	87.4

2.4.5 Sample selection for genomic investigation

We identified 10 geographic locations across the distribution of the Asian houbara bustard. Each location consists of a cluster of samples taken from wild Asian houbara individuals and is clearly geographically separated from any other such cluster of samples (Figure 2.11). We defined these locations with a latitude and longitude range (Table 2.8). Some samples are from founders, meaning that they were collected as eggs in the wild and were then used as founders in the captive breeding population. These eggs were hatched in breeding facility of the International Fund for Houbara Conservation (IFHC) and the chicks were raised in captivity.

Table 2.8: Details of the sample locations indicating geographic coordinates, behavioural status and sample numbers.

Location code	Location description	Migratory status	Wild/ Founders	Latitude range (°)	Longitude range (°)	Sample number
IRN1	Iran South; Herat	Resident	Wild/ founders	29 - 31	54 - 55	14 (3)
IRN2	Iran North; Semnan	Migrant	Founders	35 - 36	54 - 56	10 (3)
KZT1	Kazakhstan Central; Shymkent, Kyzylkum	Migrant	Wild	42 - 43.5	67 - 69	10 (3)
KZT2	Kazakhstan Central; Shymkent, Betpak-Dala	Migrant	Wild	44.5 - 46.5	67.4 - 72	16 (3)
KZT3	Kazakhstan East; Balkash	Migrant	Wild	46.31 - 48	78.5 - 81	12 (3)
KZT4	Kazakhstan West; Fetisovo	Migrant	Wild	42 - 44	52 - 54	10 (3)
MGL	Mongolia; Galba Gobi	Migrant	Wild	42 - 45	107 - 109	13 (3)
UZB1	Uzbekistan South; Karmana	Migrant	Wild	39 - 40.5	64 - 67	9 (3)
UZB2	Uzbekistan North; Madanyat	Migrant	Wild	40.5 - 42	54 - 67	10 (3)
YMN	Yemen Al Marah	Resident	Founders	16 - 17	51 -53	7 (3)

Sample numbers are written as the total number of samples in the dataset with the number of samples used in the analysis in brackets.

We assigned a migratory status (resident or migrant) to each location using satellite tracking information and past reports (Figure 2.11). Individuals from 8 of the 10 locations have satellite tracking data. These are from 7 locations whose individuals are known to migrate (KZT1, KZT2, KZT3, KZT4, UZB1, UZB2 and MGL) and some individuals from southern Iran (IRN1), which is known to be a resident population (Figure 2.11). All founder samples could not be used in this way to inform their migratory behaviour, as they were taken from captive birds. For the founder individuals, we assigned the known behaviour of the populations to the individuals from these sampling locations (YMN; resident, IRN1; resident and IRN2; migrant).

Individuals from Yemen and southern Iran are confirmed to be residents. Some individuals sampled in southern Iran (IRN1) were tracked by satellite and shown to migrate between eastern Kazakhstan and southern Iran. Therefore, these individuals are actually migrants from central Kazakhstan who were sampled very late in their wintering area before returning to their breeding grounds. We excluded these

individuals from further analyses as their sampling location and their population of origin are not the same. One individual was only tracked for a few weeks (Figure 2.11) and we cannot confirm whether this movement is dispersal or seasonal migration based on the tracking data alone. Based on the tracking of individuals from northern Iran (IRN2) that are not part of our dataset (Pakniat et al. 2020), we know individuals from this location are migrant.

Migrant individuals were sampled from 8 locations and follow three flyways. Individuals sampled in western Kazakhstan (KZT4) follow the western flyway and move towards Iraq and Syria in the winter. All other individuals (from KZT1, KZT2, KZT3, UZB1, UZB2, MGL) follow the central or eastern flyways and move to southern Iran, Afghanistan and Pakistan in the winter. Individuals from Uzbekistan (UZB1, UZB2) and central Kazakhstan (KZT1, KZT2) follow a more direct southward route, while individuals from eastern Kazakhstan (KZT3) and Mongolia (MGL) follow a longer route beside the mountains in Kyrgyzstan and Afghanistan.

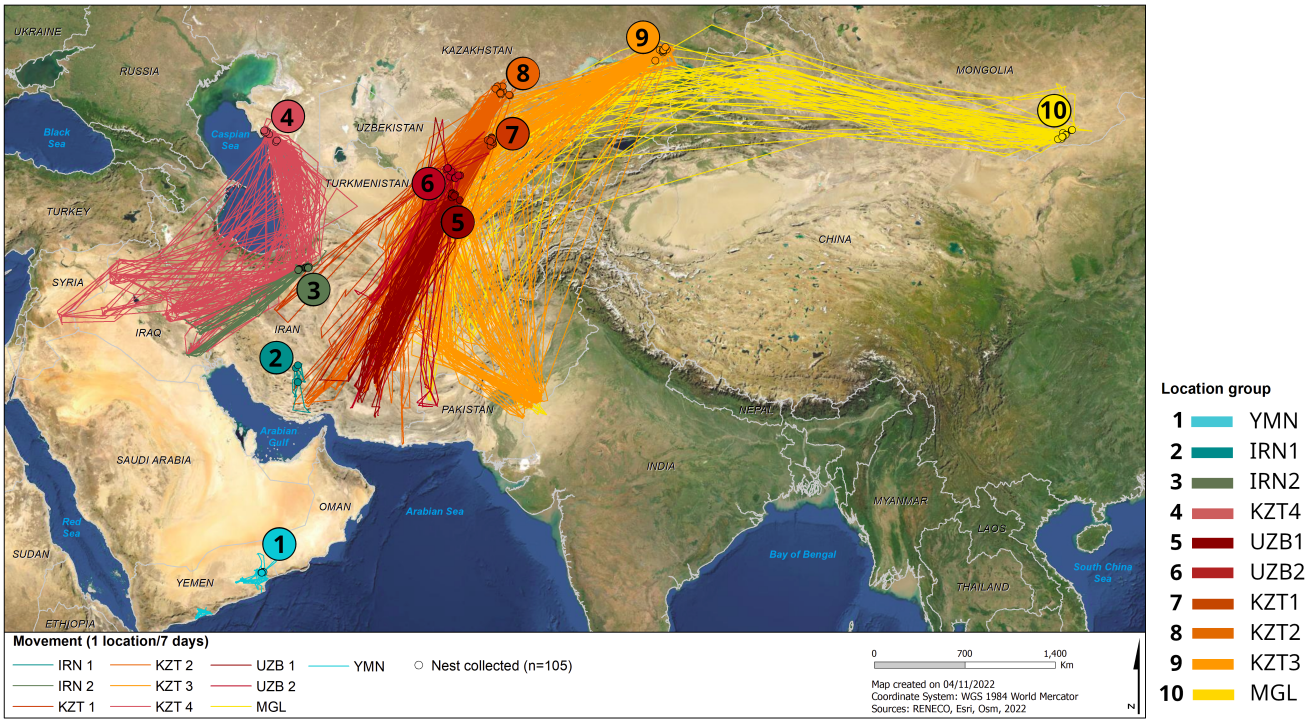
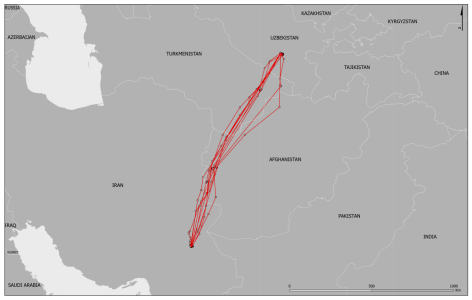


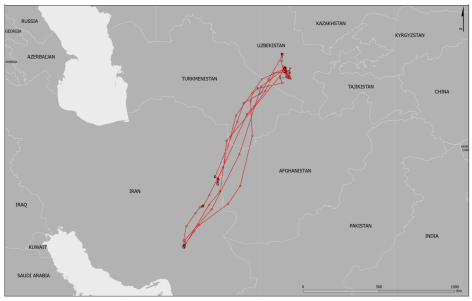
Figure 2.11: Migration route and sampling locations for selected Asian houbara individuals tracked by satellite. Tracks from birds not part of the analysed dataset are also shown for sampling sites from which we only analysed founders. The sampling locations correspond to the following codes: [1] YMN; [2] IRN1; [3] IRN2; [4] KZT4; [5] UZB1; [6] UZB2; [7] KZT1; [8] KZT2; [9] KZT3; [10] MGL.

When available, we chose at least 10 samples (if available) from each location to send for whole-genome sequencing. We used satellite tracking maps for most samples to choose the samples with the highest amount of metadata. We favoured samples from individuals with multiple years of tracking data and high route fidelity (Figure 2.12). We sent 111 samples in total for whole-genome sequencing (Table 2.9). Most of these samples were from blood on FTA cards, but we also sent blood in ethanol, muscle in ethanol and blood on dry ice samples.

[A]



[B]



[C]

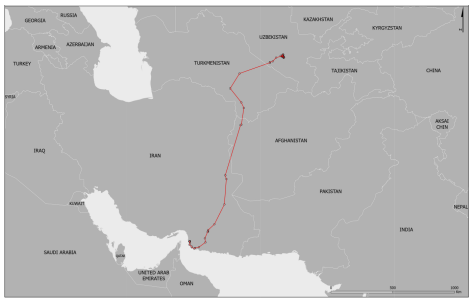


Figure 2.12: Illustration of tracking data from samples from UZB1 with high [A], medium [B] and low [C] preference for sample selection.

Table 2.9: Number of samples of different types that were sent for whole-genome sequencing.

Sample type	Number
Blood on FTA cards	83
Blood in ethanol	23
Muscle in ethanol	4
Blood in dry ice	1
Total	111

2.4.6 Processing of data for population genomic approaches

The samples were sequenced to an expected coverage of either 10× or 30× (Figure 2.13). Only one sample expected to be 30× coverage actually met this expectation. All except one of the samples expected to be 10× coverage reached the expected coverage. Some of the blood in ethanol samples yielded a much higher coverage than expected (>20× coverage). With few exceptions, the samples yielded at least 15× coverage.

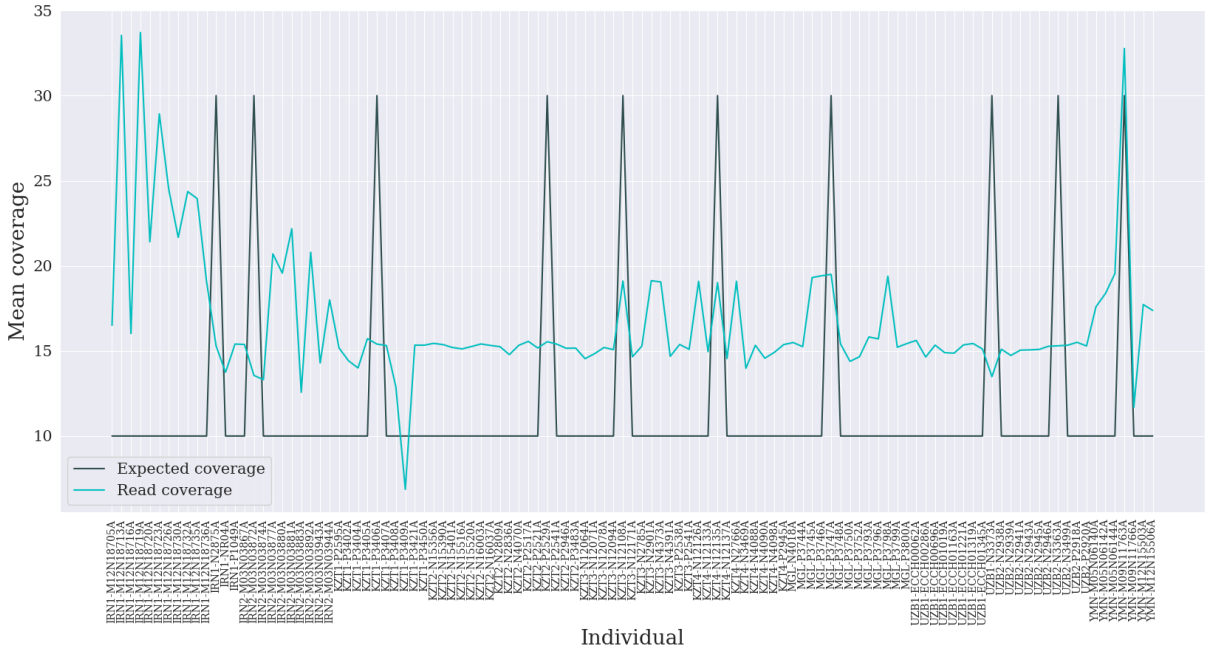
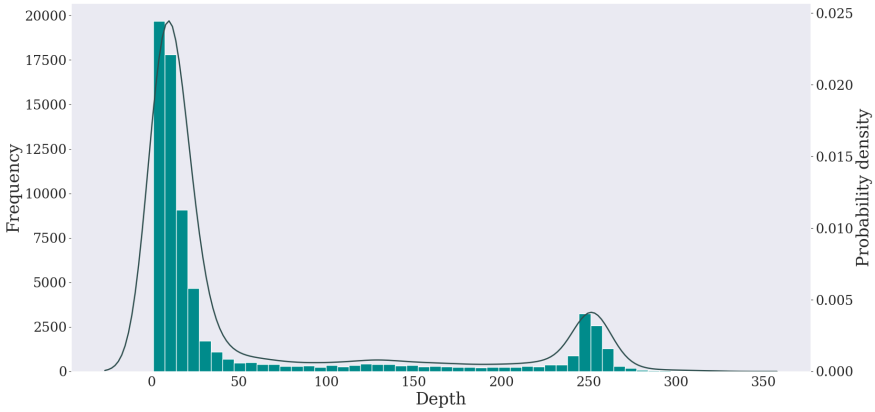


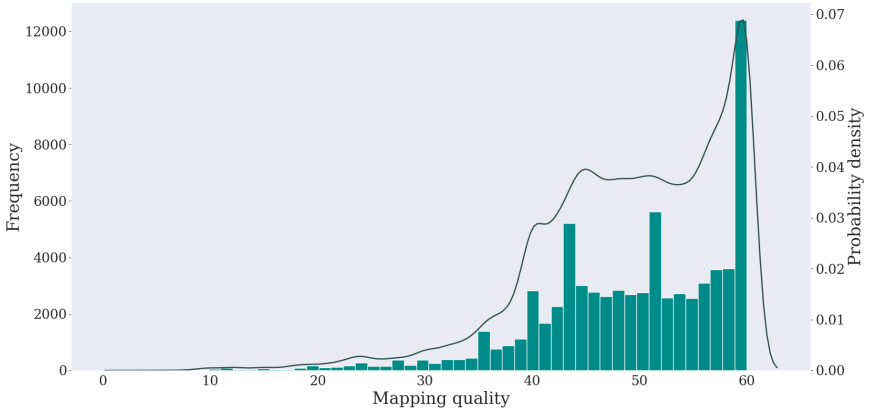
Figure 2.13: Expected and calculated mean read coverage of whole-genome sequence data. The grey line represents the expected coverage and the blue line the actual coverage.

We implemented a custom pipeline to get the haplotypes of the samples. Because of time constraints we selected 30 samples representing the 10 geographic locations out of the total of 111 genomic datasets available. We did this for the 13 candidate genes that we identified from the literature and recovered from our new annotation. Following a mapping protocol with BWA-MEM (Li and Durbin 2009) we called the variants of the candidate genes. We plotted the distributions of depth, mapping quality and strand bias to visually determine the limits of these parameters for variant filtering (Figure 2.14). For depth, the main goal was to exclude any variants with abnormally high depth. For the example of *ADCY8,* we kept all variants with a depth ≥ 3 and ≤ 50. This would include all variants with a depth that falls within the first mode of the distribution (Figure 2.14). We used a minimum mapping quality of 30 for all candidate genes, as we did not see a difference in the number of haplotypes we got with a higher minimum mapping quality and therefore more stringent filtering. Similarly, we used a maximum strand bias of 10 for all candidate genes.

[A]



[B]



[C]

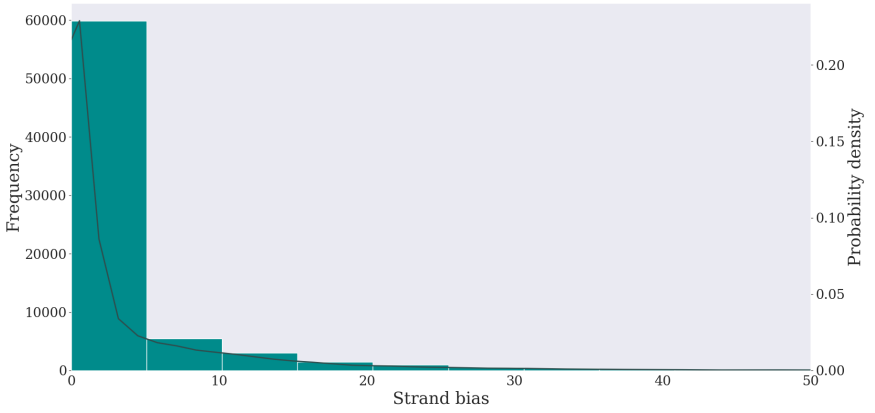


Figure 2.14: Distribution of parameters for variant filtering of the *ADCY8* gene as an example of the procedure followed for all candidate genes. [A] Depth; [B] Mapping quality; [C] Strand bias (sequencing bias where one DNA strand is favoured over the other).

2.4.7 Population genomics and analyses of candidate genes for migration in birds

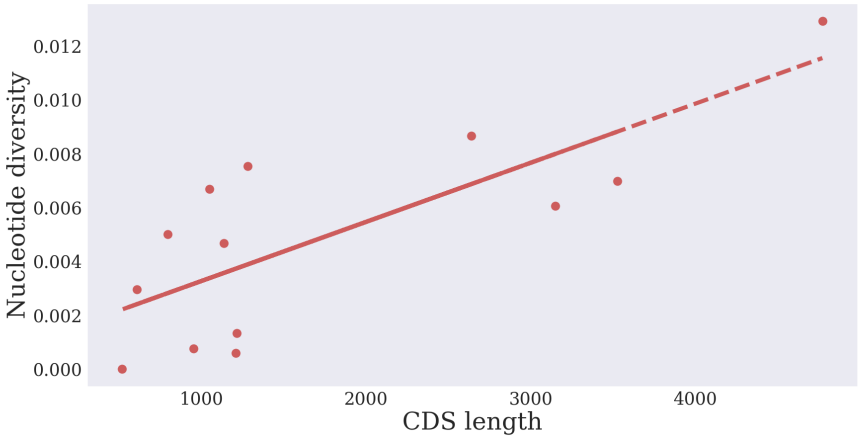
We calculated the haplotype diversity and nucleotide diversity for all of the candidate genes using only their coding sequences. The number of haplotypes for the sequences ranges from the minimum of 1 for *ADCYAP1* to the maximum of 60 (2 for each of the 30 individuals) for *ADCY8*, *CLOCK*, *HERC4* and *TOP2B* (Table 2.10). For the purposes of our study, *ADCYAP1* would be uninformative in our downstream analysis. Therefore, we did not include it in the rest of the study and only 12 candidate genes remained in the downstream analyses. The haplotype diversity is not significantly correlated to the length of the CDS (Figure 2.15). The nucleotide diversity is significantly positively correlated with the CDS length (R^2= 0.56; p=0.002).

Table 2.10: Haplotype statistics for the 13 candidate genes.

Gene	CDS (bp)	H	Hd	π
ADCY8	3525	60	1	0.007
ADCYAP1	519	1	0	0
CLOCK	2637	60	1	0.0086
CRABP1	609	13	0.893	0.003
DIO2	795	31	0.862	0.005
DRD4	1137	43	0.974	0.0047
FECH	1209	4	0.432	0.00059
HERC4	3147	60	1	0.0061
MLNR	1050	45	0.986	0.0067
RAD51D	951	6	0.567	0.00076
ST6GALNAC2	1215	5	0.597	0.0013
TGFBR1	1281	57	0.997	0.0075
TOP2B	4770	60	1	0.0129

H = Number of haplotypes; Hd = Haplotype diversity; π = Nucleotide diversity.

[A]



[B]

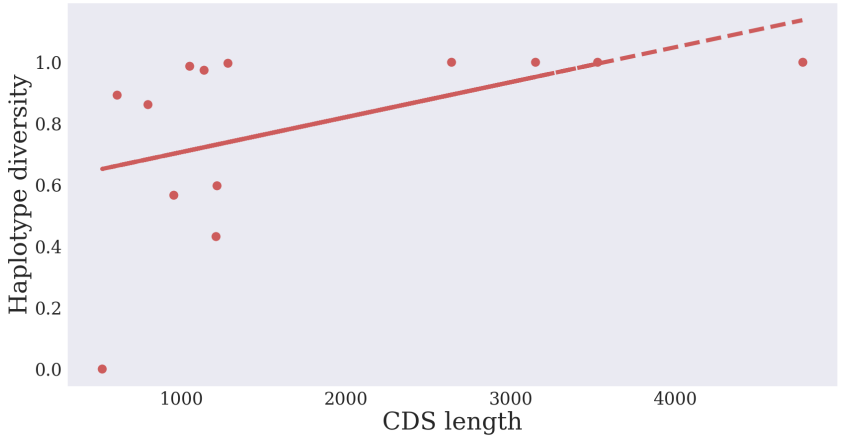


Figure 2.15: Scatterplots illustrating the relationship between haplotype statistics and candidate gene coding sequence (CDS) lengths. Haplotype diversity vs. CDS length: R^2= 0.17; p=0.09. Nucleotide diversity vs. CDS length: R^2= 0.56; p=0.002**. [A] Nucleotide diversity vs. CDs length; [B] Haplotype diversity vs. CDS length.

We found a significant effect of both migratory behaviour and geographical location based on an AMOVA with the coding sequences of candidate genes. We have 60 haplotypes for the *CLOCK* CDS, the maximum possible for our set of 30 individuals. An AMOVA for the *CLOCK* CDS seems to assign 50% of the covariance within samples and 50% of the covariance among locations within behaviour groups. AMOVA's for the other genes with 60 haplotypes (*ADCY8*; *HERC4*; *TOP2B*) gave exactly the same results as *CLOCK*, even though they have different sequence lengths and nucleotide diversity. The AMOVA results for these genes were ultimately meaningless and we excluded them from the analysis.

For all the candidate genes that have less than 60 haplotypes (*CRABP1, DIO2, DRD4*, *FECH*, *MLNR*, *RAD51D*, *ST6GALNAC2* and *TGFBR1*), the majority of variance was explained by differences among locations (50.683% - 79.548%). In the case of all the candidate genes, this variance was significantly

greater than expected if the samples were randomly distributed ($p \leq 0.001$) (Table 2.11). The second greatest percentage of variance was explained by variation within samples in all 8 candidate genes (23.4% - 49.929%). This variance was significantly less than expected with randomly distributed samples in all 8 candidate genes ($p \leq 0.001$).

DIO2 is the only gene for which differences among migrants and residents significantly contribute to the variance, with 5.037% of the variance being explained by these differences. This component is significantly greater than expected if the samples were randomly distributed ($p \leq 0.01$) (Table 2.11).

Table 2.11: AMOVA statistics for *CRABP1, DIO2, DRD4, FECH, MLNR, RAD51D*, *ST6GALNAC2* and *TGFBR1* (migrants vs. residents).

Source of variation	Df	Sum Sq	% covariance	ϕ	Direction of deviation from H_0
CRABP1					
Between behaviour	1	7.908	0.802	0.008	greater
Between locations (within behaviour)	8	60.792	56.089	0.565***	greater
Between samples (within location)	20	16.669	-1.658	-0.038	greater
Within samples	30	27	44.758	0.552***	less
Total	59	112.367	100		
DIO2					
Between behaviour	1	9.908	5.032	0.05*	greater
Between locations (within behaviour)	8	63.292	57.475	0.605***	greater
Between samples (within location)	20	15.667	0.403	0.0106	greater
Within samples	30	23	37.09	0.629***	less
Total	59	111.867	100		
DRD4					
Between behaviour	1	7.433	0.584	0.006	greater
Between locations (within behaviour)	8	57.667	52.51	0.528***	greater
Between samples (within location)	20	17.667	-2.906	-0.062	greater
Within samples	30	30	49.812	0.501***	less
Total	59	112.667	100		
FECH					

Between behaviour	1	9.717	-0.441	-0.004	greater
Between locations (within behaviour)	8	79.083	79.548	0.792***	greater
Between samples (within location)	20	7.333	-2.507	-0.12	greater
Within samples	30	14	23.4	0.766***	less
Total	59	110.133	100		
MLNR					
Between behaviour	1	7.15	0.146	0.001	greater
Between locations (within behaviour)	8	56.75	50.733	0.508***	greater
Between samples (within location)	20	20	0.833	0.017	greater
Within samples	30	29	48.288	0.517***	less
Total	59	112.9	100		
RAD51D					
Between behaviour	1	9.042	-0.367	-0.004	greater
Between locations (within behaviour)	8	73.458	73.637	0.734***	greater
Between samples (within location)	20	7.333	-8.353	-0.313	greater
Within samples	30	21	35.083	0.649***	less
Total	59	110.833	100		
ST6GALNAC2					
Between behaviour	1	8.892	-0.514	-0.005	greater
Between locations (within behaviour)	8	72.708	71.251	0.709***	greater
Between samples (within location)	20	11.333	-0.836	-0.029	greater
Within samples	30	18	30.099	0.7***	less
Total	59	110.933	100		
TGFBR1					
Between behaviour	1	7.142	0.219	0.002	greater
Between locations (within behaviour)	8	56.458	50.683	0.508***	greater
Between samples (within location)	20	19.333	-0.832	-0.017	greater
Within samples	30	30	49.929	0.501***	less
Total	59	112.933	100		

1300

We created haplotype heatmaps for each candidate gene CDS (Figure 2.16). This represents the polymorphic sites within the genes and can be used to identify linked SNPs that are associated with a certain group (location or behaviour). In most genes, we see a higher frequency of alternate alleles in individuals from YMN, IRN1 and IRN2 (represented by cool colours). This is evident by the higher density of black blocks in the haplotypes of these individuals. This splits the individuals into two groups, one that is exclusively migrants from Central Asia, and one that is a mix of residents (YMN and IRN1) and migrants (IRN2) from South-West Asia. This split occurs between geographic regions, rather than between migratory behaviours. We see this trend in *ADCY8*, *CLOCK*, *CRABP1*, *DIO2*, *HERC4*, *ST6GALNAC2*, *TGFBR1* and *TOP2B*. It is especially prevalent in *DIO2* and *HERC4*. Not all of the genes showed this pattern (cf. *DRD4*, *FECH*, *MLNR* and *RAD51D*).

Figure 2.16: Haplotype heatmaps illustrating polymorphic nucleotide sites for twelve informative candidate genes. Each row is a haplotype and each column one polymorphic site. Cream-coloured and black squares are reference and alternative alleles. Residents and individuals from IRN2 are represented by cool colours while all other migrant individuals are represented by warm colours as illustrated on Figure 2.11.

Based on the split between individuals from South-West Asia and Central Asia from the haplotype heatmaps, we performed an AMOVA using these groupings. We removed all genes with 60 haplotypes for the same reasons as in the AMOVA on migratory behaviour.

For all genes with less than 60 haplotypes (*CRABP1, DIO2, DRD4*, *FECH*, *MLNR*, *RAD51D*, *ST6GALNAC2* and *TGFBR1*), the majority of variance was explained by differences among locations (50.687% - 79.647%). In the case of all the candidate genes, this variance was significantly greater than expected if the samples were randomly distributed ($p \leq 0.001$) (Table 2.12). The second greatest percentage of variance was explained by variation within samples in all 8 candidate genes (23.401% - 49.947%). This variance was significantly less than expected with randomly distributed samples in all 8 candidate genes ($p \leq 0.001$).

For two of the candidate genes, *CRABP1* and *DIO2,* a part of the variance is explained by differences among geographic regions (0.47% for *CRABP1* and 2.90% for *DIO2*). Even though these are small percentages, in both these genes it is significantly greater than expected if the samples were randomly distributed ($p \leq 0.01$) (Table 2.12).

Table 2.12: AMOVA statistics for *CRABP1, DIO2, DRD4, FECH, MLNR, RAD51D*, *ST6GALNAC2* and *TGFBR1* (South-west Asia vs. Central Asia).

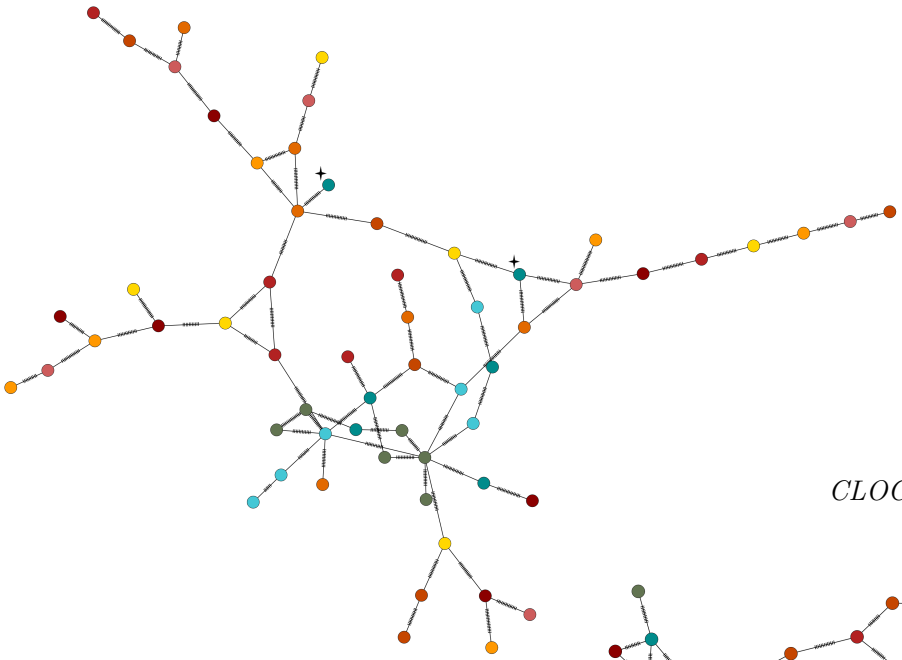
Source of variation	Df	Sum Sq	% covariance	ϕ	Direction of deviation from H_0
CRABP1					
Between geographic region	1	7.843	0.467	0.005**	greater
Between locations (within region)	8	60.857	56.308	0.566***	greater
Between samples (within location)	20	16.669	-1.663	-0.038	greater
Within samples	30	27	44.888	0.551***	less
Total	59	112.367	100		
DIO2					
Between geographic region	1	9.454	2.902	0.0290**	greater
Between locations (within region)	8	63.746	58.948	0.607***	greater
Between samples (within location)	20	15.667	0.410	0.0108	greater
Within samples	30	23	37.74	0.623***	less
Total	59	111.867	100		
DRD4					
Between geographic region	1	7.481	0.551	0.006	greater
Between locations (within region)	8	57.619	52.504	0.528***	greater

Between samples (within location)	20	17.667	-2.908	-0.062	greater
Within samples	30	30	49.853	0.501***	less
Total	59	112.767	100		
FECH					
Between geographic region	1	9.625	-0.540	-0.0054	greater
Between locations (within region)	8	79.175	79.647	0.792***	greater
Between samples (within location)	20	7.333	-2.507	-0.12	greater
Within samples	30	14	23.401	0.766***	less
Total	59	110.133	100		
MLNR					
Between geographic region	1	7.17	0.156	0.002	greater
Between locations (within region)	8	56.73	50.718	0.508***	greater
Between samples (within location)	20	20	0.833	0.017	greater
Within samples	30	29	48.293	0.517***	less
Total	59	112.9	100		
RAD51D					
Between geographic region	1	9.643	1.057	0.011	greater
Between locations (within region)	8	72.857	72.427	0.732***	greater
Between samples (within location)	20	7.333	-8.286	-0.313	greater
Within samples	30	21	34.803	0.652***	less
Total	59	111.833	100		
ST6GALNAC2					
Between geographic region	1	8.965	-0.227	-0.002	greater
Between locations (within region)	8	72.635	71.025	0.709***	greater
Between samples (within location)	20	11.333	-0.834	-0.029	greater
Within samples	30	18	30.035	0.7***	less
Total	59	110.933	100		
TGFBR1					
Between geographic region	1	7.156	0.198	0	greater
Between locations (within region)	8	56.444	50.687	0.5***	greater
Between samples (within location)	20	19.333	-0.832	0	greater

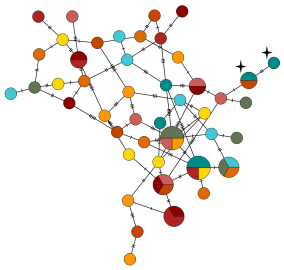
Within samples	30	30	49.947	0.5***	less
Total	59	112.933	100		

We created a minimum spanning haplotype network with the CDS of each candidate gene (Figure 2.17) and a haplotype network based on the amino acid sequence of each candidate gene (Figure 2.18). The haplotypes are coloured in the same way as the heatmaps, with the cool colours representing individuals from South-West Asia and the warm colours representing migrants from Central Asia.

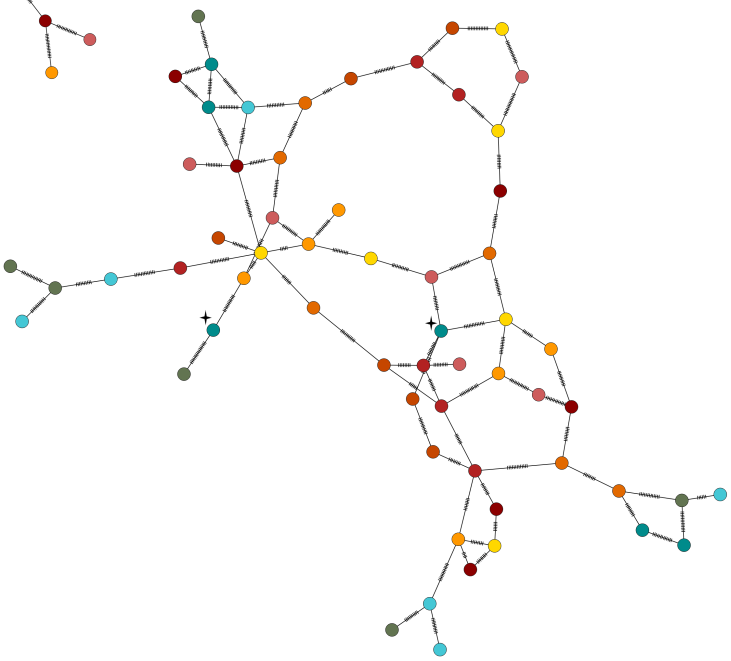
ADCY8



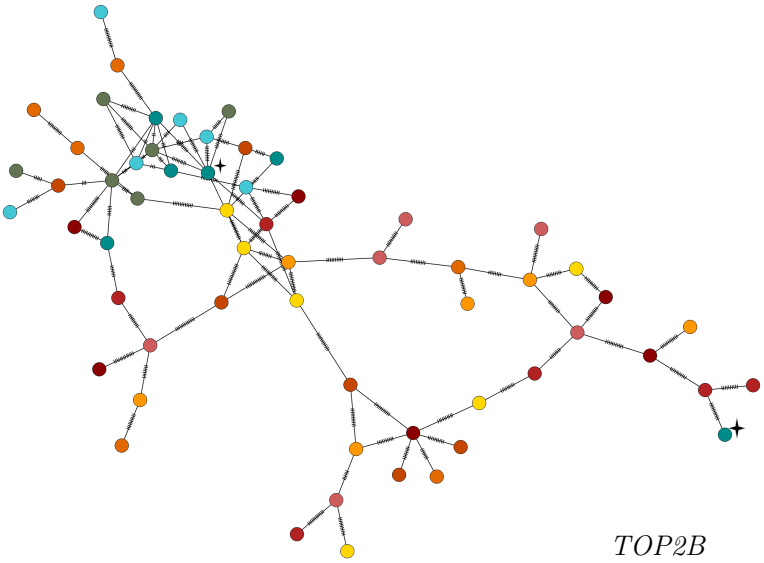
MLNR



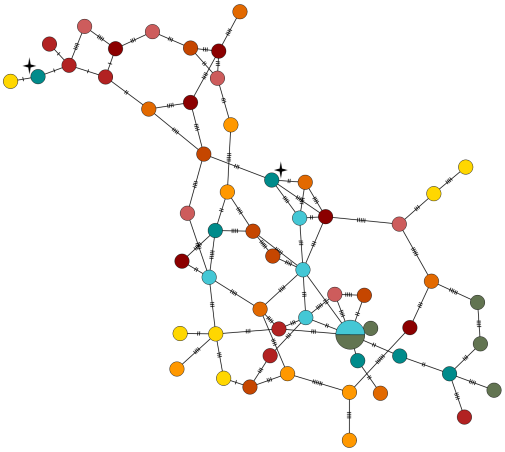
CLOCK



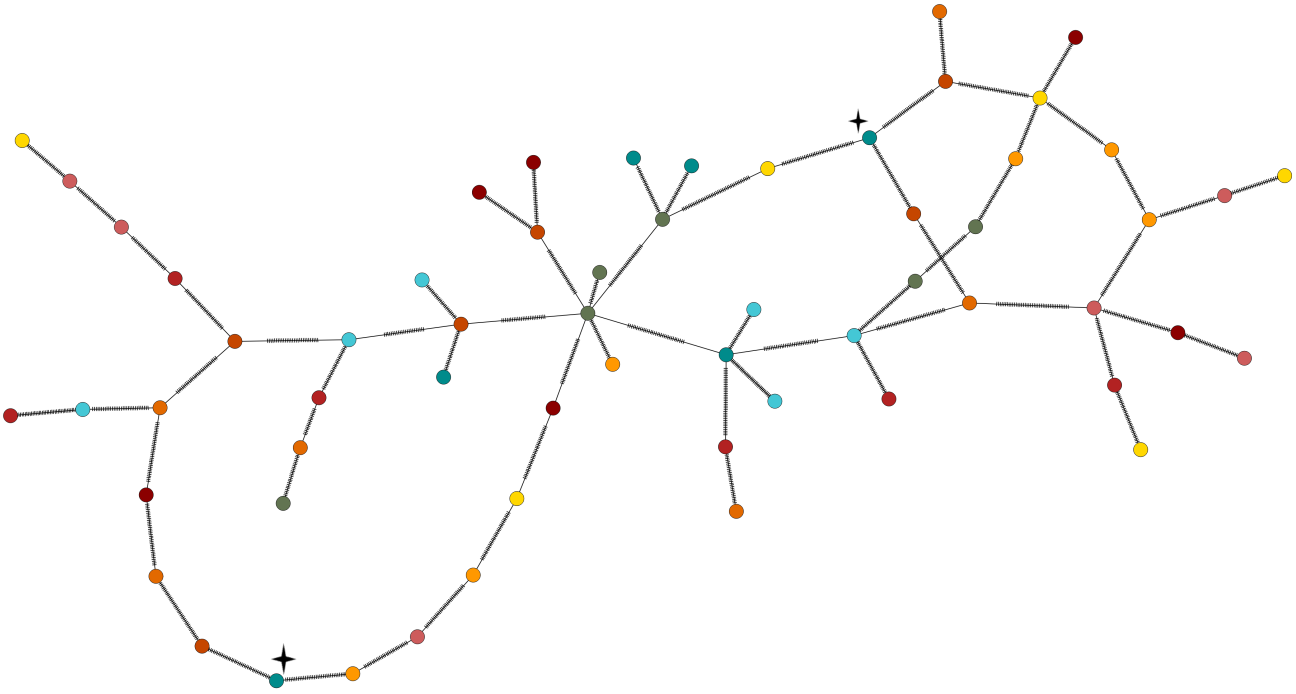
HERC4



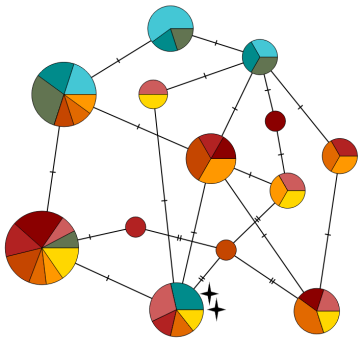
TGFBR1



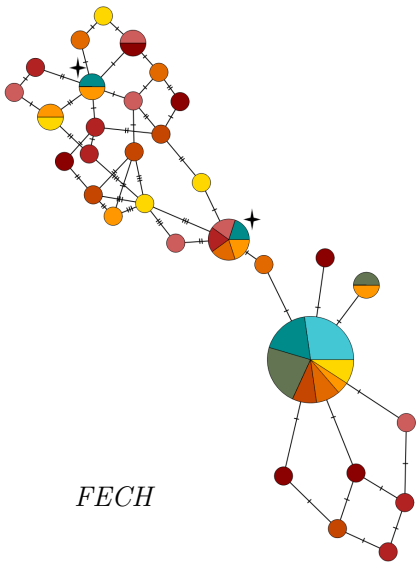
TOP2B



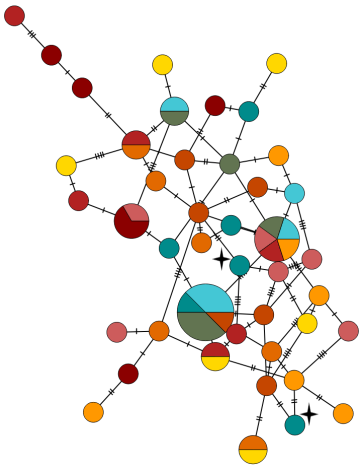
CRABP1



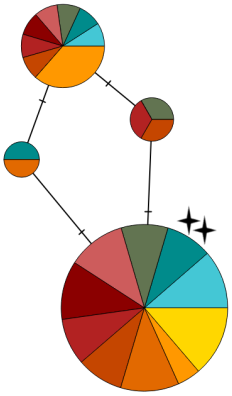
DIO2



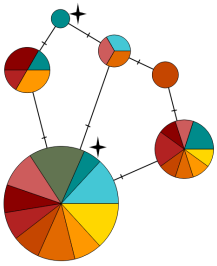
DRD4



FECH



RAD51D



ST6GALNAC2

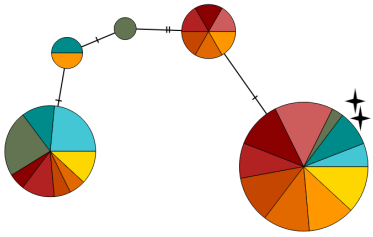
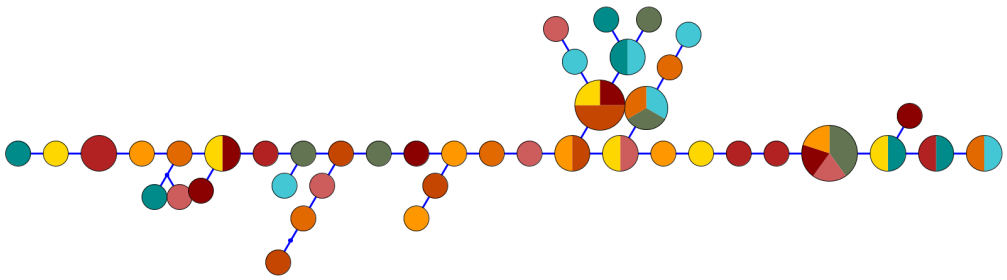


Figure 2.17: Minimum spanning haplotype networks of 12 candidate gene coding sequences. Residents and individuals from IRN2 are represented by cool colours while all other migrant individuals are represented by warm colours, as illustrated in Figure 2.11. Haplotypes from the wild individual N2875 from IRN1 are indicated with a star.

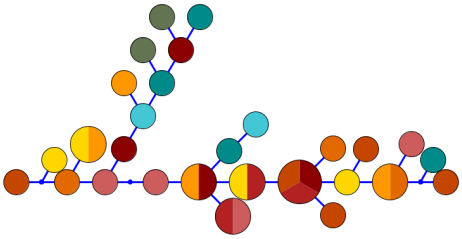
Our main candidate genes of interest are *DIO2* and *CRABP1* because the variance due to differences in geographic region was significantly greater than expected for these genes. *DIO2* shows a major haplotype in multiple individuals with many other less frequent haplotypes that are present in only one or two individuals (Figure 2.17). The majority of the individuals with the major haplotype are from South-West Asia, even this group represents the minority of the dataset. *CRABP1* shows two haplotypes that are exclusive to individuals from South-West Asia and one haplotype that is shared by a majority of individuals from South-West Asia (Figure 2.17). Only two haplotypes from this gene is shared between individuals from South-West Asia and individuals from Central Asia, and the rest of the haplotypes are exclusive to individuals from Central Asia.

We were interested to see if we could see similar patterns in the haplotype networks as we saw in the haplotype heatmaps. *ST6GALNAC2* shows 5 haplotypes with one showing a majority of individuals from South-West Asia, one showing a majority of individuals from Central Asia and one showing only individuals from Central Asia (Figure 2.17). It also seems that there is a trend with more individuals from South-West Asia on one side of the network and more individuals from Central Asia on the other side of the network. We see similar trends in *HERC4* and *TGFBR1*, where the individuals from South-West Asia cluster together in the network and the individuals from Central Asia branch out in other directions (Figure 2.17). Contrasting to this pattern, in *CLOCK* individuals from Central Asia are more central in the network and individuals from South-West Asia are in the periphery of the network.

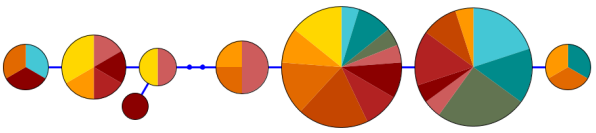
ADCY8



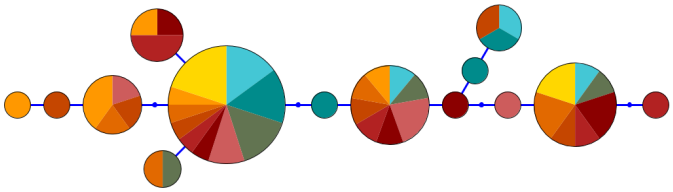
CLOCK



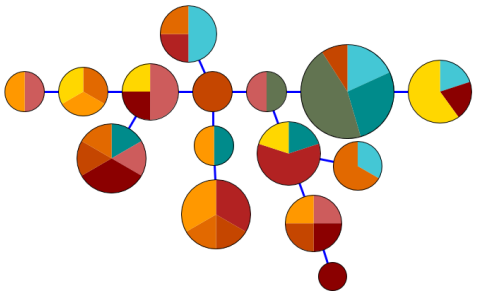
HERC4



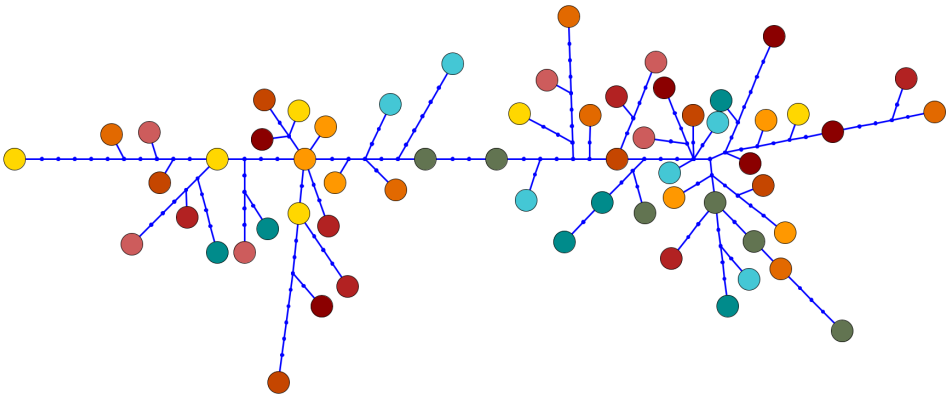
MLNR



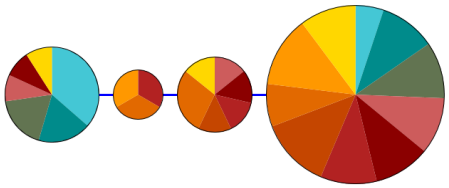
TGFBR1



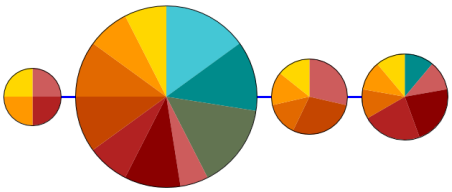
TOP2B



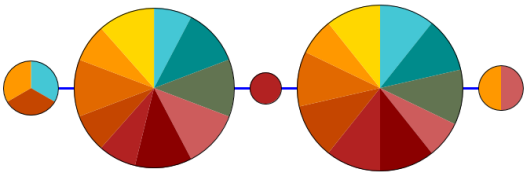
CRABP1



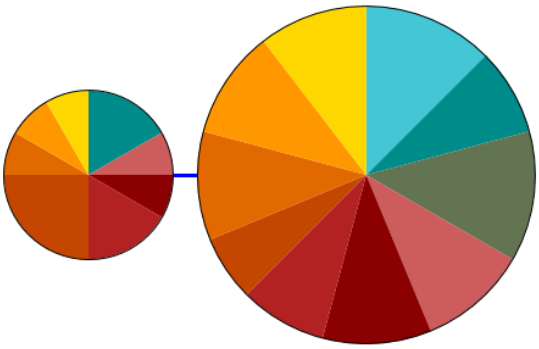
DIO2



DRD4



RAD51D



ST6GALNAC2

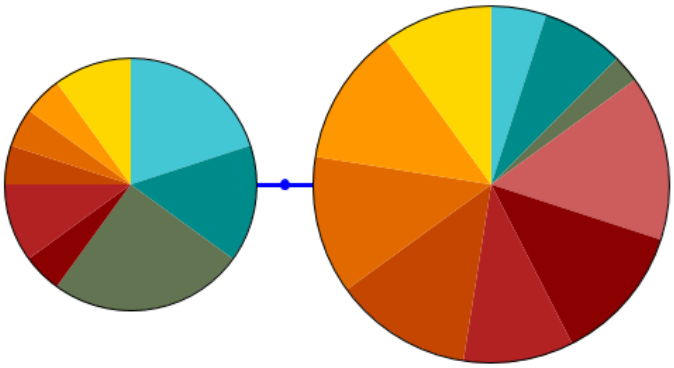


Figure 2.18: Amino acid haplotype networks of candidate genes. Residents and individuals from IRN2 are represented by cool colours while all other migrant individuals are represented by warm colours, as illustrated in Figure 2.11.

The networks from the amino acid sequences allow us to remove a lot of the variation as most of the polymorphisms in the sequences are synonymous. It can also give an indication of functional differences in the effect of these genes between resident and migrant individuals. Similar to the network made with the CDS, *DIO2* shows a major haplotype that is present in almost all of the resident individuals (Figure 2.18). The other haplotypes are almost exclusive to migrant individuals. CRABP1 shows four haplotypes of which two are exclusive to migrants (Figure 2.18). One haplotype is made of a large majority of resident individuals and the final, and largest, haplotype has an even distribution of individuals from all 10 locations.

ST6GALNAC2 only has two haplotypes, one with a majority of resident individuals and one with a majority of migrant individuals. *HERC4* shows a similar pattern in its two most frequent haplotypes, but it also has multiple less frequent haplotypes with various distributions of the locations. *TGFBR1* still has many haplotypes even when we use the amino acid sequence (Figure 2.18). Its most frequent haplotype is almost exclusively present in residents, with the other haplotypes having various distributions of migrants and residents.

We performed a principal component analysis with the complete genomes of our 30 selected individuals. When we used all of the individuals, the individuals from Yemen were clearly separated from the rest of the individuals (Figure 2.19A). As the differentiation with the Yemen population was overwhelming compared to the potential differences among the remaining populations, we also performed the analysis for the individuals excluding Yemen. The first principal component then explains 6.26% of the variance and splits the individuals into two groups (Figure 2.19B). This split is explained by differences linked to sex chromosomes (Figure 2.19C). When analysing the second and third biggest principal components, individuals are arranged according to geography (Figure 2.19D). A cline from the left of the plot to the right shows the geographical cline of the locations from MGL in the far east to IRN1 in the far south of the distribution range.

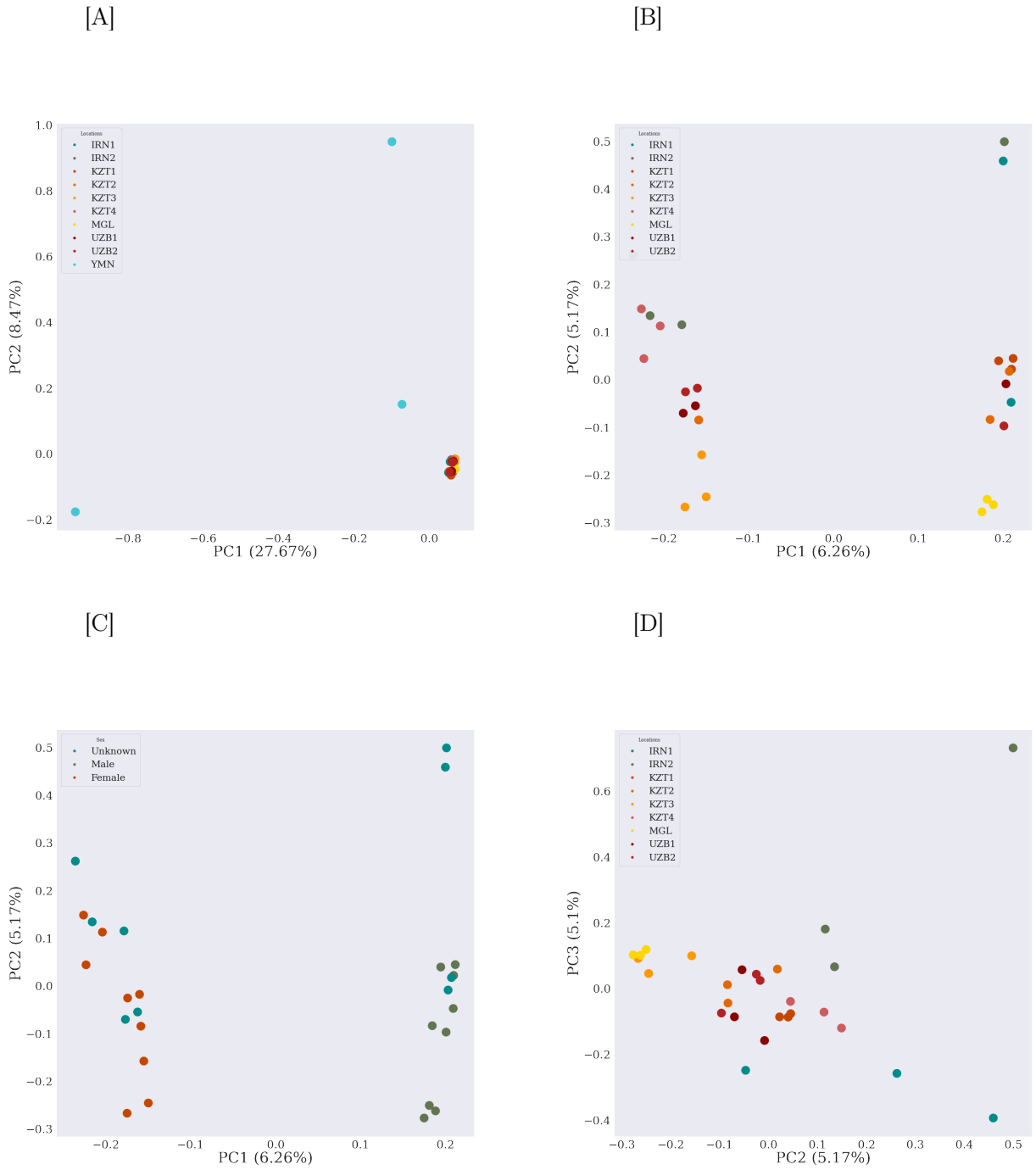


Figure 2.19: Principal component analysis using whole-genome data of selected individuals. [A]: All individuals, coloured by location, PC1 vs. PC2; [B]: Excluding YMN, coloured by location, PC1 vs. PC2; [C]: Excluding YMN, coloured by sex, PC1 vs. PC2; [D]: Excluding YMN, coloured by location, PC2 vs. PC3.

2.5 Discussion

2.5.1 A more contiguous reference genome assembly for Asian houbara

A highly contiguous and accurate genome assembly is the foundation of genomic analysis, allowing researchers to investigate regions of interest more efficiently and to make inferences with more confidence (Abdellah et al. 2004; Eilbeck et al. 2009; Zimin et al. 2009). The previous reference genome for Asian houbara (ASM69519v1) was generated by Zhang et al. (2014) using SOAPdenovo (Zhang et al. 2014; Li et al. 2010), which produces many short contigs due to the way it reduces computational complexity (Sohn and Nam 2018). The genome is from a wild-caught male houbara sourced from the Dubai Falcon Hospital, but the geographic origin of this individual is unknown.

The two assemblies, created for Asian houbara (REN_Cmacq_1.0 and REN_Cmacq_2.0) had high contiguity, but a large proportion of unassigned bases in the assemblies (Table 2.2), suggesting a potential technical artifact. After investigation, we found that the wrong insert size was used when the genomes were initially assembled. After correction by GenoScreen, the proportion of unassigned bases significantly dropped and the contiguity was still very high for both assemblies. Although the number of unassigned bases has dropped in the new assemblies, it still remains around 10 times higher than the genome available on NCBI (ASM69519v1). Compared to this first assembly, the new assemblies are more contiguous, with a 10-fold reduction in the number of scaffolds and a 100-fold decrease in the N50 value (Table 2.2; Figure 2.6). The completeness of the assemblies also increased by ~23% compared to the previous genome, with the presence of the near-complete set of core vertebrate genes (> 96% for REN_Cmacq_1.0 following the BUSCO protocol). Only a few of these core genes were fragmented in the new assemblies, which indicate that we would also be able to find a larger number of genes of interest through annotation. The new assemblies therefore represent an improved resource to mine candidate genes, which is one of the objectives of the present study.

2.5.2 Genome annotation of the new Asian houbara reference genome

Zhang et al. (2014) also annotated the protein-coding sequences of their genome assembly using a homology-based approach. They created a uniform set of genes using orthologous and non-orthologous genes from chicken (Hillier et al. 2004), zebra-finch (Warren et al. 2010) and human (Lander et al. 2001) genomes. These genes were found in the Ensembl database (Aken et al. 2016).

We increased the number of genes, proteins and expressed sequence tags (ESTs) in the annotation of the new reference assembly compared to the previous annotation (Table 2.4). The number of genes in the new assembly (17,992) is closer to that of the annotated chicken and zebra finch genomes than the previous assembly (13,996). The previous annotation was done without transcriptome data and only the protein-coding sequences were annotated. We obtained an improved annotation of the Asian houbara genome by doing multiple rounds of MAKER2 annotation with the new assembly (REN_Cmacq_1.0) and including RNA evidence.

We created transcriptome assemblies that had a combined completeness exceeding than 95% (following BUSCO protocol; Figure 2.8). The transcriptomes created from embryo samples were considerably more complete than those obtained from blood samples. Embryonic tissue contains many embryonic stem cells

which will continue to form various mature tissues. Some genes are thought to be uniquely or more abundantly expressed in these cells (Wobus and Boheler 2005), which could contribute to the higher completeness observed in the embryonic transcriptomes. In birds, red blood cells are nucleated and bird blood is generally a good source of DNA (Delmore and Liedvogel 2020), but considering the relatively low core gene completeness, it confirms that embryonic tissues are a better source of RNA. The new transcriptome assembly for Asian houbara represents a great resource to efficiently annotate genes of interest in our new and improved genome assembly (REN_Cmacq_1.0). It can also serve as RNA evidence in the annotation of other bustard species, none of which are currently annotated.

We did the first round of annotation on the assembly that was created with the wrong insert size. As expected, this assembly gave us the worst results based on our quality measures, which includes the AED scores. In comparison, the assembly with the correct insert size had better AED scores, but still had poor AED scores compared to the current reference (ASM69519v1). The AED score is a measure of how well the evidence given as a reference aligns with the predicted gene features from the annotation (Holt and Yandell 2011). As the previous reference annotation had much less evidence, there is a smaller chance for the predictions to not align with the evidence. AED is also closely related to splice complexity (Eilbeck et al. 2009). A larger number of predicted transcripts and splice variants increase the splice complexity of a gene model. The new reference annotation was done with RNA-evidence, whereas the previous annotation had no RNA-evidence. Without RNA-evidence, splice variants can't be predicted in the annotation, therefore the new annotation has greater splice complexity. AED increases as splice complexity increase, which could explain the higher AED scores in the new reference annotation. However, alternative explanations cannot be ruled out.

Nevertheless, we could retrieve 13 of the 14 candidate genes from our genome annotation of the REN_Cmacq_1.0 assembly (Table 2.7). We could not find *TSPO2* in the annotation with either its UniProt or Ensembl code. We could only identify seven candidate genes in the previous reference genome (ASM69519v1), as expected when considering its lower gene completeness. This result confirms that the new reference assembly is a better resource to search for our candidate genes. It will likely help improve the annotation of the houbara genome and identify other genes linked to phenotypic traits of interest in the future.

2.5.3 Analysis of candidate genes linked to migration in Asian houbara

Using a dataset that represents houbara across a large part of its distribution range, we were able to find a divide between migrants and residents in one candidate gene, *DIO2*. From AMOVA, we found that the component of variance due to differences among residents and migrants is greater than expected if there were no behavioural divide (Table 2.12). From haplotype analysis, we found that there is an indication of a geographical divide between residents and migrants from South-West Asia and migrants from Central Asia (Figure 2.16). In most of our candidate genes, linked sets of polymorphic sites occur more often in individuals from South-West Asia than Central Asian individuals. We also see that South-West Asian individuals share haplotypes in many genes, often sharing it with a minority of Central Asian individuals or no Central Asian individuals at all. An AMOVA organised according to these geographical groups show a significant component of variance due to difference among geographical groups in two genes, *CRABP1* and *DIO2*. According to the haplotype heatmaps, it is likely that *ADCY8*, *CLOCK*, *HERC4* and *TOP2B* share the same geographical divide as *CRABP1* and *DIO2*, but it could not be confirmed with the AMOVA as it is uninformative for these genes. It is also more difficult to analyze these genes

with haplotype networks, because each haplotype is unique. We also see no divide due to behaviour or geography in some genes, such as *FECH* and *RAD51D*. In these genes, haplotypes are evenly spread among behavioural and geographic groups.

Our haplotype analysis also gives some insight into the evolution of migration in Asian houbara. The strong divide between South-West Asian and Central Asian individuals suggests that migrants from northern Iran has a different genetic makeup than migrants from Central Asia. This points to the potential convergence of migratory behaviour in these different migrant groups which would support the idea that resident individuals could adopt migratory behaviour under certain conditions. The CDS haplotype networks of the genes for which we see a geographic divide, with the exception of *CLOCK,* show a consistent pattern among them. The South-West Asian individuals tend to cluster in the middle of the networks, while Central Asian individuals tend to be in the peripheral parts of the networks (Figure 2.17). This pattern could indicate that the South-West Asian individuals are closer to the ancestral population than Central Asian individuals, but it is difficult to confirm without information about ancestral haplotypes. Previous work on Asian houbara estimated a recent population expansion (around 34 kyr ago) which is concordant with historical separation and geographical distribution of Asian and African houbara bustards (*C. undulata*) (Pitra et al. 2004). Concerning the behavioural difference between African and Asian houbara, it is likely that resident Asian houbara would be more closely related to African houbara than their migrant conspecifics, as African houbara are resident.

In the search for positive selection differentially affecting the individuals from the different behavioural and geographic groups, *CRABP1* and *DIO2* could be considered the best candidates. Based on the literature, both of these two genes have been involved in the differential selection of migrants in the European Swainson's thrush (*Catharus ustulatus*) (Johnston et al. 2016). These genes are thought to be associated with changes in neural plasticity. The product of a pathway catalyzed by *DIO2,* thyroid hormone T3, was linked to neural plasticity and gonadal growth and regression in Japanese quail (*Coturnix japonica*) (Yoshimura et al. 2003; Yamamura et al. 2004). *DIO2* has also been linked to seasonal changes in body weight in the Siberian hamster (*Phodopus sungorus*) (Murphy and Ebling 2011) and adaptive thermogenesis in mice (Jesus et al. 2001). As body mass changes and seasonal gonadal size changes are both linked to migration in birds (Newton 2007), *DIO2* is also potentially involved in these mechanisms during migration. Thyroid hormones and retinoic acid share the same nuclear hormone receptor family, retinoic X receptor (RXR) (Helfer et al. 2012) and *DIO2* and *CRABP1* are involved in retinoic acid signalling, which has been linked to seasonal neural plasticity in various organisms (Johnston et al. 2016).

The divide between South-West Asian and Central Asian individuals do not group individuals from these groups together exclusively. The CDS haplotype networks of *DIO2* and *CRABP1* show South-West Asian individuals grouping together in one or a few haplotypes, with the exception of one individual from southern Iran (IRN1) (Ring number: N2875) (Figure 2.17).

2.5.4 Using candidate genes to assign Asian houbara individuals to Central Asian or South-West Asian groups

Assigning a migratory behaviour to an individual remains a challenge, especially if metadata and tracking data is not available or limited. We chose to analyse individuals based on both tracking data and other metadata, such as collection date. For our genomic analysis, we chose only individuals whose collection

date coincides with their breeding season. One wild-caught individual from IRN1 (N2875) was sampled in the beginning of May, which coincides with the breeding season of migrants (Madon et al. 2015), and was only tracked for a few weeks before it died. We included this individual in the analysis as a resident from southern Iran, according to its sampling location. Two other wild-caught individuals from central Kazakhstan were sampled in southern Iran in the beginning of April, which is close to spring departure dates for migrants from central and East Kazakhstan (Madon et al. 2015). These individuals were not part of our analysed set.

From our results, we have reason to believe that the individual (N2875) sampled in IRN1 is also a migrant from central or east Kazakhstan. In our haplotype analysis, N2785 groups more regularly with Central-Asian migrants than with South-West Asian individuals (Figure 2.17). This is most evident in *DIO2* and *CRABP1,* which show the most likely link to migratory behaviour and a geographical divide. When excluding N2785, all South–West Asian individuals are found in only two out of 31 haplotypes for *DIO2* and four out of 13 for *CRABP1.* N2785 is an outlier in the networks of *DIO2* and *CRABP1*, indicating that it could be possible to use these genes to assign a migratory status or geographic origin to an individual if there is uncertainty around the behaviour or population of the individual. Considering that it is a challenge to confidently assign migratory status with the current dataset, further investigation of *DIO2* and *CRABP1* is needed and could improve the assignment.

2.5.5 Geography shapes population genomics of Asian houbara

Using the coding sequences of the candidate genes for selected samples, we found that the majority of the genetic variance was explained by differences among locations. These findings are illustrated by both the AMOVA tests (Table 2.11 and 2.12) and the geographic distribution of haplotypes (Figure 2.16). This component of variance was also significantly greater than expected when there is no population structure present. The gene flow among locations is therefore too limited to erode differences across the distribution range.

Applying the principal component analysis of the complete genomes of the 30 Asian houbara individuals, we can observe a pattern that suggests genetic isolation-by-distance. There is no strong genetic clustering of individuals according to their geographical origin, but it is possible to observe some groupings of the individuals based on their location. Moreover, there is a steady cline ranging from Mongolian individuals on the left-hand side of the PCA to the south of Iran on the right-hand side of the PCA, illustrating that geographically close locations that are closer on the graph. This suggests that, across the genome, the largest contribution to the genetic differences between individuals is from their geographic separation.

2.6 Conclusion

In this study, we were able to evaluate and improve the genomic resource for the Asian houbara bustard by confirming that the new genome (REN_Cmacq_1.0) is more contiguous and by providing a more robust annotation. The number of unassigned bases in the assembly is still a concern, as it creates uncertainty in the correctness of certain parts of the assembly, and will require further investigation. Incorporating long-read sequencing to the new assembly can contribute to making it even more contiguous and will help in removing a large fraction of unassigned bases. The AED scores from the annotation also reflect

the need for further improvement of the genome annotation. Finally, manual curation of the annotated genome is a crucial next step to a robust genome annotation for this species. We were also able to use our new genomic resource, together with whole-genome re-sequencing, to investigate 12 candidate genes of interest. Further analysis of the most promising candidate genes, CRABP1 and DIO2, could produce even greater insight into the role of these genes in long-distance migration. Investigations into the protein structure of these genes could be useful to understand the functional differences between different variations of the genes. A larger dataset could provide a better resolution of differences in behaviour and it would be beneficial to add samples from the large resident population in Israel. On a population scale, understanding the distribution of alleles at these gene regions in Asian houbara populations could also contribute to a possible marker which could be used to assign migratory behaviour to an individual without additional information. Hybrid and translocation experiments could also provide insights into the effect of the alleles in individuals that are released into specific locations. Variations in alleles other than that of the coding sequences, such as allele length which is common in the *CLOCK* gene region, are also worth investigating. A whole-genome approach where differentiation is calculated over sliding windows of the genome would be a good next step into investigating signals of selection in Asian houbara. Tackling the questions of selection and adaptation linked to seasonal migration would support conservation efforts in Asian houbara by allowing better management of this species with highly diverse migratory behaviour. Genetic information could provide relevant information to ongoing management efforts to help further increase survival rates of released individuals and maintain genetic diversity in the species, combating the declines in numbers and diversity due to hunting and habitat loss.

2.7 References

Abdellah Z., A. Ahmadi, S. Ahmed, M. Aimable, R. Ainscough, et al., 2004 Finishing the euchromatic sequence of the human genome. Nature 431: 931-945. https://doi.org/10.1038/nature03001

Aken B. L., S. Ayling, D. Barrell, L. Clarke, V. Curwen, et al., 2016 The Ensembl gene annotation system. Database 2016: 1-19. https://doi.org/10.1093/database/baw093

Altschul S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews S., 2010 FastQC: A Quality Control tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bairoch A., and R. Apweiler, 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research 28: 45–48. https://doi.org/10.1093/nar/28.1.45

Baker M., 2012 De novo genome assembly: what every biologist should know. Nature Methods 9: 333–337. https://doi.org/10.1038/nmeth.1935

Bazzi G., R. Ambrosini, M. Caprioli, A. Costanzo, F. Liechti, et al., 2015 Clock gene polymorphism and scheduling of migration: a geolocator study of the barn swallow *Hirundo rustica*. Scientific Reports 5: 12443.
https://doi.org/10.1038/srep12443

Bazzi G., J. G. Cecere, M. Caprioli, E. Gatti, L. Gianfranceschi, et al., 2016 Clock gene polymorphism, migratory behaviour and geographic distribution: a comparative study of trans-Saharan migratory birds. Molecular Ecology 25: 6077–6091. https://doi.org/10.1111/mec.13913

Bird Life International, 2014 Review of the global conservation status of the Asian Houbara Bustard *Chlamydotis macqueenii*. BirdLife International.

Bourret A., and D. Garant, 2015 Candidate gene–environment interactions and their relationships with timing of breeding in a wild bird population. Ecology and Evolution 5: 3628–3641. https://doi.org/10.1002/ece3.1630

Brůna T., A. Lomsadze, and M. Borodovsky, 2020 GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics and Bioinformatics 2: 2. https://doi.org/10.1093/nargab/lqaa026

Brůna T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics 3: 1. https://doi.org/10.1093/nargab/lqaa108

Burnside R. J., N. J. Collar, and P. M. Dolman, 2017 Comparative migration strategies of wild and captive-bred Asian Houbara *Chlamydotis macqueenii*. Ibis 159: 374–389. https://doi.org/10.1111/ibi.12462

Cantarel B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, et al., 2008 MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. 18: 188–196. https://doi.org/10.1101/gr.6743907

Caprioli M., R. Ambrosini, G. Boncoraglio, E. Gatti, A. Romano, et al., 2012 Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. PLOS ONE 7: e35140. https://doi.org/10.1371/journal.pone.0035140

Combreau O., F. Launay, M. A. Bowardi, and B. Gubin, 1999 Outward migration of Houbara Bustards from two breeding areas in Kazakhstan. The Condor 101: 159–164. https://doi-org.uplib.idm.oclc.org/10.2307/1370458

Combreau O., F. Launay, and M. Lawrence, 2001 An assessment of annual mortality rates in adult-sized migrant houbara bustards (*Chlamydotis [undulata] macqueenii*). Animal Conservation 4: 133–141. https://doi.org/10.1017/S1367943001001160

Combreau O., S. Riou, J. Judas, M. Lawrence, and F. Launay, 2011 Migratory pathways and connectivity in Asian Houbara Bustards: evidence from 15 Years of satellite tracking. PLOS ONE 6: e20570. http://dx.doi.org.uplib.idm.oclc.org/10.1371/journal.pone.0020570

Contina A., E. S. Bridge, J. D. Ross, J. R. Shipley, and J. F. Kelly, 2018 Examination of Clock and Adcyap1 gene variation in a neotropical migratory passerine. PLOS ONE 13: e0190859. https://doi.org/10.1371/journal.pone.0190859

Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, et al., 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, et al., 2021 Twelve years of SAMtools and BCFtools. GigaScience 10: 1-4. https://doi.org/10.1093/gigascience/giab008

Delmore K. E., and M. Liedvogel, 2020 Avian Population Genomics Taking Off: Latest Findings and Future Prospects. In *Statistical Population Genomics Methods in Molecular Biology*, ed. Dutheil J. Y., 413–433. Springer US, New York, NY.

Eilbeck K., B. Moore, C. Holt, and M. Yandell, 2009 Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 10: 67. https://doi.org/10.1186/1471-2105-10-67

Ejigu G. F., and J. Jung, 2020 Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology 9: 295. https://doi.org/10.3390/biology9090295

Fidler A. E., K. van Oers, P. J. Drent, S. Kuhn, J. C. Mueller, et al., 2007 Drd4 Gene polymorphisms are associated with personality variation in a passerine bird. Proceedings: Biological Sciences 274: 1685–1691.

Franchini P., I. Irisarri, A. Fudickar, A. Schmidt, A. Meyer, et al., 2017 Animal tracking meets migration genomics: transcriptomic analysis of a partially migratory bird species. Molecular Ecology 26: 3204–3216. https://doi.org/10.1111/mec.14108

Grabherr M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, et al., 2011 Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 29: 644–652. https://doi.org/10.1038/nbt.1883

Gu Z., S. Pan, Z. Lin, L. Hu, X. Dai, et al., 2021 Climate-driven flyway changes and memory-based long-distance migration. Nature 591: 259–264. https://doi.org/10.1038/s41586-021-03265-0

Gurevich A., V. Saveliev, N. Vyahhi, and G. Tesler, 2013 QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haghani A., M. Aliabadian, J. Sarhangzadeh, and A. Setoodeh, 2018 Evaluation of genetic diversity and population structure of Macqueen's Bustard *Chlamydotis macqueenii* in Iran. Bird Study 65: 108–113. https://doi.org/10.1080/00063657.2017.1414770

Helfer G., A. W. Ross, L. Russell, L. M. Thomson, K. D. Shearer, et al., 2012 Photoperiod Regulates Vitamin A and Wnt/β-Catenin Signaling in F344 Rats. Endocrinology 153: 815–824. https://doi.org/10.1210/en.2011-1792

Hillier L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, et al., 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716. https://doi.org/10.1038/nature03154

Holt C., and M. Yandell, 2011 MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491. https://doi.org/10.1186/1471-2105-12-491

Hunter J. D., 2007 Matplotlib: A 2D Graphics Environment. Computing in Science Engineering 9: 90–95. https://doi.org/10.1109/MCSE.2007.55

IFHC, 2020 International Fund for Houbara Conservation Annual Report 2019-2020

Jesus L. A. de, S. D. Carvalho, M. O. Ribeiro, M. Schneider, S.-W. Kim, et al., 2001 The type 2 iodothyronine deiodinase is essential for adaptive thermogenesis in brown adipose tissue. The Journal of Clinical Investigation 108: 1379–1385. https://doi.org/10.1172/JCI13803

Johnsen A., A. E. Fidler, S. Kuhn, K. L. Carter, A. Hoffmann, et al., 2007 Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. Molecular Ecology 16: 4867–4880. https://doi.org/10.1111/j.1365-294X.2007.03552.x

Johnston R. A., K. L. Paxton, F. R. Moore, R. K. Wayne, and T. B. Smith, 2016 Seasonal gene expression in a migratory songbird. Molecular Ecology 25: 5680–5691. https://doi.org/10.1111/mec.13879

Jones P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, et al., 2014 InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Judas J., O. Combreau, M. Lawrence, M. Saleh, F. Launay, et al., 2006 Migration and range use of Asian Houbara Bustard *Chlamydotis macqueenii* breeding in the Gobi Desert, China, revealed by satellite tracking. Ibis 148: 343–351. https://doi.org/10.1111/j.1474-919X.2006.00546.x

Kamvar Z. N., J. F. Tabima, and N. J. Grünwald, 2014 Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2: e281. https://doi.org/10.7717/peerj.281

Kamvar Z. N., J. C. Brooks, and N. J. Grünwald, 2015 Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. Frontiers in Genetics 6:208. https://doi: 10.3389/fgene.2015.00208

Kim D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37: 907–915. https://doi.org/10.1038/s41587-019-0201-4

Knief U., and W. Forstmeier, 2016 Mapping centromeres of microchromosomes in the zebra finch (*Taeniopygia guttata*) using half-tetrad analysis. Chromosoma 125: 757–768. https://doi.org/10.1007/s00412-015-0560-7

Kuhl H., C. Frankl-Vilches, A. Bakker, G. Mayr, G. Nikolaus, et al., 2021 An unbiased molecular approach using 3´-UTRs resolves the avian family-level tree of life. Molecular Biology and Evolution 38: 108–127. https://doi.org/10.1093/molbev/msaa191

Lander E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al., 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921. https://doi.org/10.1038/35057062

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Lee E., G. A. Helt, J. T. Reese, M. C. Munoz-Torres, C. P. Childers, et al., 2013 Web Apollo: a web-based genomic annotation editing platform. Genome Biology 14: R93. https://doi.org/10.1186/gb-2013-14-8-r93

Leigh J. W., and D. Bryant, 2015 Popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 6: 1110–1116. https://doi.org/10.1111/2041-210X.12410

Li H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li R., H. Zhu, J. Ruan, W. Qian, X. Fang, et al., 2010 De novo assembly of human genomes with massively parallel short read sequencing. Genome Research. 20: 265–272. https://doi.org/10.1101/gr.097261.109

Madon B., E. L. Nuz, C. Ferlat, and Y. Hingrat, 2015 Insights into the phenology of migration and survival of a long migrant land bird. bioRxiv 028597. https://doi.org/10.1101/028597

Martin M., M. Patterson, S. Garg, S. O Fischer, N. Pisanti, et al., 2016 WhatsHap: fast and accurate read-based phasing. bioRxiv 085050. doi: https://doi.org/10.1101/085050

McKenna A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, et al., 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303. https://doi.org/10.1101/gr.107524.110

Mueller J. C., F. Pulido, and B. Kempenaers, 2011 Identification of a gene associated with avian migratory behaviour. Proceedings: Biological Sciences 278: 2848–2856. https://doi-org.uplib.idm.oclc.org/10.1098/rspb.2010.2567

Newton I., 2007 *The Migration Ecology of Birds*. Elsevier.

Murphy M., and F. J. P. Ebling, 2011 The role of hypothalamic tri-iodothyronine availability in seasonal regulation of energy balance and body weight. Journal of Thyroid Research 2011: e387562. https://doi.org/10.4061/2011/387562

Paradis E., 2010 PEGAS: an R package for population genetics with an integrated–modular approach. Bioinformatics 26: 419–20. https://doi.org/10.1093/bioinformatics/btp696

Pitra C., M.-A. D'Aloia, D. Lieckfeldt, and O. Combreau, 2004 Genetic variation across the current range of the Asian houbara bustard (*Chlamydotis undulata macqueenii*). Conservation Genetics 5: 205–215. https://doi.org/10.1023/B:COGE.0000030004.51398.28

Purcell S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, et al., 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–575. https://doi: 10.1086/519795

Riou S., J. Judas, M. Lawrence, S. Pole, and O. Combreau, 2011 A 10-year assessment of Asian Houbara Bustard populations: trends in Kazakhstan reveal important regional differences. Bird Conservation International 21: 134–141. https://doi.org/10.1017/S0959270910000377

Riou S., O. Combreau, J. Judas, M. Lawrence, M. S. Al Baidani, et al., 2012 Genetic differentiation among migrant and resident populations of the threatened Asian houbara bustard. Journal of Heredity 103: 64–70. https://doi.org/10.1093/jhered/esr113

Saino N., G. Bazzi, E. Gatti, M. Caprioli, J. G. Cecere, et al., 2015 Polymorphism at the Clock gene predicts phenology of long-distance migration in birds. Molecular Ecology 24: 1758–1773. https://doi.org/10.1111/mec.13159

Simão F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sohn J., and J.-W. Nam, 2018 The present and future of de novo whole-genome assembly. Briefings in Bioinformatics 19: 23–40. https://doi.org/10.1093/bib/bbw096

Stamatakis A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stelzer G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, et al., 2016 The GeneCards suite: from gene data mining to disease genome sequence analyses. Current Protocols in Bioinformatics 54: 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

The UniProt Consortium, 2021 UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49: D480–D489. https://doi.org/10.1093/nar/gkaa1100

Tourenq C., O. Combreau, S. B. Pole, M. Lawrence, V. S. Ageyev, et al., 2004 Monitoring of Asian houbara bustard *Chlamydotis macqueenii* populations in Kazakhstan reveals dramatic decline. Oryx; Cambridge 38: 62–67. https://doi.org/10.1017/S0030605304000109

Tourenq C., O. Combreau, M. Lawrence, S. B. Pole, A. Spalton, et al., 2005 Alarming houbara bustard population trends in Asia. Biological Conservation 121: 1–8. https://doi.org/10.1016/j.biocon.2004.03.031

Warren W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, et al., 2010 The genome of a songbird. Nature 464: 757–762. https://doi.org/10.1038/nature08819

Yamamura T., K. Hirunagi, S. Ebihara, and T. Yoshimura, 2004 Seasonal morphological changes in the neuro-glial interaction between gonadotropin-releasing hormone nerve terminals and glial endfeet in Japanese quail. Endocrinology 145: 4264–4267. https://doi.org/10.1210/en.2004-0366

Yoshimura T., S. Yasuo, M. Watanabe, M. Iigo, T. Yamamura, et al., 2003 Light-induced hormone conversion of T4 to T3 regulates photoperiodic response of gonads in birds. Nature 426: 178–181. https://doi.org/10.1038/nature02117

Zdobnov E. M., and R. Apweiler, 2001 InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zhang G., C. Li, Q. Li, B. Li, D. M. Larkin, et al., 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346: 1311–1320. https://doi.org/10.1126/science.1251385

Zimin A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, et al., 2009 A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biology 10: R42. https://doi.org/10.1186/gb-2009-10-4-r42

Zimin A. V., G. Marcais, D. Puiu, M. Roberts, S. L. Salzberg, et al., 2013 The MaSuRCA genome assembler. Bioinformatics 29: 2669–2677. https://doi.org/10.1093/bioinformatics/btt476