




# Genetic differentiation and immunogenetics of two sympatric storm petrel species on the Azores

**A thesis submitted to Cardiff University for the degree of Doctor of Philosophy**

**By Alexandra McCubbin**

**April 2021**

# Summary

This thesis investigates genetic differentiation, mate choice and immunogenetics in two recently diverged species of storm petrels breeding on the Azores – the geographically widespread *H. castro* and the Azores endemic *H. monteiroi*. Previous work found the two species largely distinct for mitochondrial and nuclear loci, but the degree of reproductive isolation or breeding time switching between the two seasonally separated taxa remains unclear.

In chapter 2, a method for rapid and cost-effective screening for mitochondrial clade membership was developed, tested, and applied to both species. The method did not reveal any mismatches between presumed sample identity and mitochondrial clade, but future application may do so, allowing further investigation of the biology and genomics underlying mismatches. In chapter 3, a method for PCR amplification of Major Histocompatibility Complex (MHC) DAB loci was developed, demonstrating the retention of an ancient duplication of the MHC Class IIB region in both species. In chapter 4, this method was used for high-throughput sequencing of family trios from both species. A bioinformatic allele filtering pipeline was developed, revealing 65 DAB1 and 27 DAB2 alleles across both species (individuals for DAB1/DAB2: *H. castro* n=111/110; *H. monteiroi* n= 99/91). DAB1 was more variable than DAB2 within each species, and the 'vulnerable' Azores endemic *H. monteiroi* was found to exhibit similar immunogenetic variability as *H. castro*, suggesting that variability has been retained despite recent declines in *H. monteiroi.* Extensive signals of incomplete lineage sorting were found for DAB alleles of both species, likely reflecting balancing selection and their recent speciation. Nevertheless, DAB allele sharing between the two was limited, and multi-locus analyses found them to be clearly differentiated for MHC genotypes.

This thesis contributes valuable molecular insights into the ecology and evolutionary history of *H. castro* and *H. monteiroi*, including a comprehensive characterisation of their immunogenetic variability and differentiation.

# Table of Contents

# Table of Figures

# List of Tables

For my parents.

# Acknowledgements:

First and foremost, my gratitude goes towards my incredibly supportive supervisory team, who have been there from day one. I am sorry if I have caused a few grey hairs over the years, but I am truly grateful for your support. Frank – thank you for your knowledge, ideas and belief in me. I'm especially grateful for your help and understanding over the last 6 months of this PhD. Renata – thank you for introducing me to the haven that is Praia Islet and trusting me with your storm petrel 'babies'. It was an unforgettable experience (…those ants! Will we get home?! The long, LONG journeys either side…but most importantly, the sheer beauty of the place). Thank you also for the pep talks over a cup of coffee in Cathays community centre, when it was needed. Rob – ever a fountain of positivity, thank you for encouraging me to take on this PhD, and being there throughout; thank you for the wonderful writing retreats on Skokholm, I might even pretend to be writing something else just to return! Carsten – thank you for all your help in shaping this PhD, and your words of wisdom – 'don't let perfect get in the way of good' – when it was needed. To you all, thank you for your feedback, comments and encouragement at the final hurdle, when I needed it most. I couldn't have done this without you.

I'd like to thank colleagues in the Azores, notably Veronica Neves and Joel Bried, for their knowledge when it was needed, their guidance in sampling, collecting and sending SO many samples to me to analyse and their all-round kindness. I'd like to thank members of the Graciosa National Park for their help in getting to and from the islet, and their continued support in conservation and management of the islet. I'd also like to thank them for granting permits to carry out this work

Prior to my PhD, I was helped greatly by Professor Bill Symondson and Dr Jen Gale. Without their faith in me as a newly graduated undergrad and giving me a chance to work as a Research Assistant, I certainly would not be doing a PhD. Jen, you have been there whenever I have needed it and have always been one of my biggest supporters. I cannot thank you enough for all your help throughout the years and your never-ending faith. I am grateful to have had you as a mentor, and now also a friend.

I'd also like to thank Jenny Dunn, who's willingness to involve me in her research and include me in published papers, most certainly helped me grow as a researcher and get the experience I needed to advance.

Many PhD students would not be where they are without the help of undergraduates. They don't always get the recognition and thanks they deserve, and so I would like to extend my gratitude to all the students that have helped me along the way – Sean Hubbert, Sophie Barnett, Erin David, Nick Price and Jack Ford. All of these contributed in some way, through lab work or data analyses and I really appreciate the help.

I would like to give a special mention and gratitude to Lucy Rowley. Her contribution to the mitochondrial chapter is invaluable and I hope putting her into the author list will suffice as a huge thank you. Ever helpful and accommodating, I couldn't have asked for a nicer or hard-working assistant, and I am glad that I can now also call you a friend.

To The Fellowship of the PhD, thank you for being there for mutual ranting, meme sharing, and general comedic relief and support. I also would like to express endless gratitude for the wider Molecular Ecology lab group. It's always been such a welcoming, helpful, and supportive space, which I've felt lucky to be a part of. I can only hope my future colleagues are just as wonderful.

I'd like to give special thanks to Lorna and Jordan who have both been there from day 1 (and beyond for Jordan). Lorna, thank you endlessly for listening to my rants, being my KESSII buddy and always being there for me. Thank you for all the help you have given me over the years with various aspects of PhD life, I am so glad we were part of the same cohort and became friends. Jordan – worm boy – you have been there since the undergraduate days, and I am sad that our paths are now going to diverge. Thank you for your endless puns that provided me with the laughs when I needed them. Also, Becca – thank you for encouraging me to get out for a park walk and coffee to clear my head – they certainly helped keep me sane.

I most certainly need to thank Angela and Trudy in the Genome Hub at Cardiff University. Without Angela's knowledge and guidance, and Trudy's assistance, I would not have the data I do!

My Family. The support of my parents throughout my life has been endless, and I wouldn't be the person I am today without them. They have always believed in me and my sisters, always wanting the best for us. From them I have learned to be determined and work hard. I am forever grateful for their ongoing support, including letting me live rent free whilst I was a poor research student, which then led me onto this PhD. My sisters – thank you for keeping me grounded and never letting me forget what a nerd I am.

Nick, I don't want to be too soppy, but I honestly cannot thank you enough. I really do not think I would have carried on to the end of this, if it hadn't been for your endless support, encouragement, and love. You've been there for me on some of my darkest days with this PhD, ready to give me the hug and pep-talk I needed to keep at it. Thank you for supporting me financially in a year where I had no income, not letting me stress about it when I had this to worry about. Thank you for never complaining, even when I had to stay up late, had no time for normal, fun social things, and relied on you to cook and clean, whilst I wrote this thesis. It has meant so much to me - you have been my rock and I will be forever grateful.

Whilst I don't think they can read (…unless it's part of their master take-over-the-world-plan), I need to thank all the pets that have let me smother my face in them when I need some respite and cheering up. It's amazing what a stroke of a soft furry head, a comforting purr or a slobbery tennis ball in your lap, can do for the soul.

Having to get a job to help with finances was not planned but writing up during a pandemic had its challenges. To that end, I'd like to thank my colleagues in the Covid Testing Facility at Cardiff University, who helped keep me sane the last 6 months of my PhD. At a time when we were all had to stay indoors, being able to work with others for 3 days a week greatly improved my mental wellbeing, and it helps being part of such a great team. They also encouraged and supported me in finishing this PhD and deserve a place here.

A special mention goes out to dry shampoo, for keeping me looking mildly clean in the last few months of cave-dwelling, writing up period.

Finally, I would like to thank my funding body, KESSII, for funding this PhD project. I would also like to thank the Genetics Society and the British Ornithological Union for the grants they awarded me, at the beginning of this PhD. Together, all 3 of these funded vital field, lab work and sequencing, making this project possible.

# Chapter One: General Introduction

*Praia Islet in the Azores archipelago, home to the band-rumped and Monteiro's storm petrel*

**"Trust me, it's paradise… so never refuse an invitation, never resist the unfamiliar, never fail to be polite & never outstay the welcome. Just keep your mind open and suck in the experience"**

**- Alex Garland, The Beach**

# Chapter 1 - Introduction

## 1.1: Background

### 1.1.1: MHC and Immune Function

The vertebrate immune system works to fight off pathogens through two different pathways of defence – the innate immune response and the adaptive immune response. The innate immune response is non-specific, acting quickly to apply a more generalised response to all recognised pathogens (Alberts et al. 2002b) and it is often the first line of defence against pathogens. There are three main parts to the innate immune system – (1) the skin, which forms a physical barrier to pathogens, stopping them entering the body; (2) mucous membranes that prevent infection by pathogens through additional means such as enzymes, mucus and cilia moving pathogens before they can infect a host; and (3) immune system cells such as natural killer cells and macrophages which can destroy the pathogen (Riera Romo et al. 2016; Smith et al. 2019). If a pathogen evades these defences, the innate immune response can act to trigger the adaptive immune response.

The adaptive immune response is more sophisticated than the innate response, utilising B- or T-lymphocytes to provide antigen-specific responses to a variety of pathogens. In addition, the adaptive immune response can also contribute to long-term immunity from pathogens, as seen with measles and chickenpox (Alberts et al. 2002a). The adaptive immune response utilises two general lines of defence – (1) an antibody response where B-lymphocytes produce antibodies that bind to and inactivate pathogenic antigens; and (2) a cell-mediated response using T-lymphocytes that recognise host cells infected with a virus and kills the cell before it can replicate (Alberts et al. 2002a).

The major histocompatibility complex (MHC) forms part of the adaptive immune response, presenting pathogen peptides for recognition by T-lymphocytes on cell surfaces. MHC loci are highly specific, coding for cell surface proteins that have key functions in regulating the adaptive immune response for the recognition of foreign pathogens (Janeaway et al. 2001a). There are 3 classes of MHC molecules, with MHC Class I and Class II presenting antigens on cell

surfaces whilst Class III has a bigger role in alternative immune processes i.e. inflammation (Rivero-de Aguilar et al. 2016).

During an immune response, antigens produced by pathogens within infected cells are transported via proteasomes and TAP transported to MHC Class I and II molecules on the cell surface. Here, the antigens bind to the MHC molecules on the cell surface, consequently presenting them for recognition by T-cells, which in turn destroys the infected cell and the pathogen within (Rock et al. 2016). More specifically, Class I molecules are on the surface of all nucleated cells, usually responding to intracellular pathogens like viruses, through cytotoxic T cells. In contrast, Class II molecules are found specifically on lymphocytes and macrophages, and are mostly facilitated in the response to extracellular pathogens such as bacteria (Hess and Edwards 2002a; Wieczorek et al. 2017).

The MHC is able to respond to a wide variety of pathogens due to its polymorphic and polygenic properties, with the presence of both Class I and II proteins providing huge variation (Janeaway et al. 2001a). This is enhanced further by heterozygosity at these loci (Doherty and Zinkernagel 1975) which in turn allows the MHC to respond to and eliminate a wider range of pathogens, ('heterozygote advantage'; Penn et al. 2002). In some species, heterozygous genotypes at MHC loci have indeed shown to be associated with increased survival (Worley et al. 2010; Brambilla et al. 2018), providing a fitness advantage that can be selectively maintained, making MHC genes targets for selection (Hughes and Yeager 1998). Selection of MHC loci includes two broad forms – (1) disease (pathogen/parasite)-based natural selection, and (2) sexual selection (which itself can be in response to fitness benefits linked with (1)); when combined, this can generate high levels of MHC diversity (Jan Ejsmond et al. 2014). Disease-based selection implies that MHC diversity is influenced by balancing selection (Bernatchez and Landry 2003b). Co-evolution with a variety of pathogens leads to the maintenance of heterozygosity, allowing response to a wider range of pathogens (Doherty and Zinkernagel 1975). Conversely, disease-based selection can also lead to rare alleles being selected for, if pathogens have adapted to fight common immune defence alleles in host populations (Bernatchez and Landry 2003). In terms of (2) sexual selection, the maintenance of MHC diversity may influence the mating decisions of individuals, so that offspring inherit a fitness advantage (Neff and Pitcher 2005).

### 1.1.2: Mate Choice and MHC

Individuals have evolved to choose their reproductive mates, targeting certain traits that may be desirable or advantageous (Andersson 1994). In turn, these mate choice decisions can create selection pressure on a wide variety of traits, influencing their evolution within a population or species (Andersson, 1994; Edward, 2015). Mate choice has been defined by Edward (2015) as "*when traits expressed by one individual lead to a non-random mating event with a member of the opposite sex*". Mate choice is common and observed in multiple animal systems (Andersson, 1994); the process of its evolution is debated, with multiple different mechanisms proposed (summarised in Andersson and Simmons 2006). Individuals may choose mates based on phenotypic traits that are directly associated with increased ability of the mate for e.g., care for young, provision of food or protection. Choice of such mates would result in direct benefits to the fitness of the choosing individual and potential offspring (Neff and Pitcher 2005). Individuals may also choose mates based on preferences for certain alleles or genetic compatibility, i.e., targeting individuals with specific alleles that contribute to individual fitness, or a more general genetic compatibility with themselves, resulting in increased offspring fitness (Neff and Pitcher, 2005). These mechanisms are not mutually exclusive and are often considered in tandem, making mate choice a complex evolutionary process (Andersson and Simmons, 2006). The implications of mate choice can be diverse, with selection pressure altering the evolution of traits to improve the chances of being chosen. In some systems, males have evolved complex mating rituals to attract females and influence their mate choice (Borgia et al. 1987; Frith and Frith 1988); in others ornamental features have evolved to attract mates, as presumed signals of genetic fitness (Pardal et al. 2018). Where genetic quality is considered by individuals, mate choice can preferentially target certain genes, levels of heterozygosity or genetic compatibility of mates, to improve offspring fitness (Tregenza and Wedell 2000). With its role in immune defence, the MHC is one such region under selection pressure, commonly the target of genetic studies of mate choice.

MHC based mate choice was first described in mice 45 years ago (Yamazaki 1976). It is thought that MHC is under some form of balancing selection (Jan Ejsmond et al. 2014), whereby multiple alleles are maintained in a gene pool of a

population at frequencies higher than expected from genetic drift alone. Mate choice and sexual selection can act to maintain such alleles, with adults showing patterns of MHC-related mate choice more often than expected by chance alone (Forsberg et al. 2007). Comparison of MHC diversity in mated pairs, with overall MHC diversity of the population, can reveal if MHC diversity and selection are linked. An extensive amount of research into MHC-dependent mate choice has been carried out, displaying different patterns of mate choice across many animal taxa. Different animal systems may target different levels of MHC diversity, depending on selection pressures present in that system, i.e., inbreeding levels, MHC diversity or specific pathogens present in a population. Evolutionary and demographic history of populations may also influence MHC-dependent mate choice, and so differing patterns between taxa is not uncommon, and MHC-dependent mate choice can be species and population-specific (Kamiya et al. 2014).Many studies have focussed on female-mediated choice, but male responses can also be observed (Jeannerat et al. 2018). MHC-based mate choice includes 3 broad hypotheses; (1) the 'good genes' approach, where females choose males with genes that have a fitness advantage against common pathogens, and will pass them onto offspring (Hamilton & Zuk 1982); (2) mate choice for heterozygosity, where or (3) mate choice for MHC compatibility. Where mate choice for heterozygosity is concerned, individuals may mate disassortatively (i.e. choose mates dissimilar to themselves) to increase offspring MHC heterozygosity and overall fitness (Landry et al. 2001; Bernatchez and Landry 2003), or disassortative mating may occur to avoid more general inbreeding and its associated potential mutations or risks (Potts and Wakeland 1990; Bernatchez and Landry 2003). Conversely, in some systems, maximal heterozygosity is vetoed for a more intermediate level of heterozygosity, with sexual selection (1) targeting compatible MHC haplotypes that provide overall increased offspring fitness, or (2) mate choice targets specific MHC haplotypes related to resistance to a common pathogen (Hamilton and Zuk 1982, Eizaguirre et al. 2009b). Preferences for varying levels of MHC diversity can therefore influence how individuals choose their mates. In contrast, MHC-mediated mate choice is not observed in all animal systems, and sometimes there is no evidence for MHC selection in mate choice at all (e.g. in great tits *Parus major*; Sepil et al. 2015). However, the genomic complexity of

MHC loci can make mate choice detection difficult, with Class I and Class II sometimes exhibiting different signals of selection (i.e. individuals may target diversity at MHC Class I genes but not Class II genes, and vice versa), and it may be important to consider both classes when assessing mate choice (Strandh et al. 2011; Gaigher et al. 2018). Similarly, duplicated MHC genes and multiple alleles may provide enough diversity that random mating provides enough diversity without choosing mates based solely on MHC (Dearborn et al. 2016). To fully understand how MHC may shape sexual selection, and whether individuals are targeting particular genes or a level of heterozygosity, it is important to classify each individual's own MHC diversity. Individual diversity is shaped by how many different alleles an individual might have, and the sharing/non-sharing of these alleles has been shown to influence mate choice decisions (Santos et al. 2016; Santos et al. 2017).

For individuals to choose mates based on MHC, their genetic MHC profile must be detectable in some way. Whilst not definitively understood, it is thought that MHC is detectable by scent and is particularly susceptible to selection in species systems with good olfactory capabilities (Boehm and Zufall 2006).

### 1.1.3: MHC and Scent

Chemical signalling is a common form of communication in animal species, detection of which relies on a well-developed olfactory system to evaluate such signals (Boehm and Zufall 2006). Chemical signals play an important role in social interaction and reproduction, and the interplay of MHC and mate choice is part of such social signalling (Yamazaki 1976; Penn 2002; Ruff et al. 2012). There are several hypotheses on how MHC translates into social signals that are detectable by scent, with no single definitive mechanism defined (Boehm and Zufall 2006). MHC molecules are present in bodily fluids such as urine and sweat, and these fragments provide odorants in these fluids (Singh et al. 1987). However, MHC molecules alone are typically non-volatile and therefore undetectable by scent, and it is thought that when passing across the cell membrane and being transported into the urine, MHC molecules actually pick up odorants from serum (Pearse-Pratt et al. 1998). A second hypothesis predicts that peptides bound by MHC molecules may be responsible for volatile odorants (Penn and Potts 1998; Milinski et al. 2005); once MHC molecules and peptides

disassociate after shedding from cell surfaces, peptides are able to interact with other molecules, such as olfactory sensory neurons (Boehm and Zufall 2006). A third hypothesis proposes that MHC may influence an individual's own microflora, and the microflora in turn alter individual odour (Zomer et al. 2009). The fourth hypothesis stems from the MHC molecules association with bacteria – here, MHC molecules act as carriers of volatile bacteria, which are broken down and released as volatiles detectable to other individuals (Singh 1999). Finally, the peptide microbe hypothesis combines the peptide and microflora hypotheses, where MHC molecules determine which peptides are presented, and these are then made volatile by the action of an individual's microflora (Penn and Potts 1998; Zomer et al. 2009). Recent studies focus on peptides, with either the third or final hypotheses seeming most likely (Milinski et al. 2005; Boehm and Zufall 2006), however the true mechanism has not yet been confirmed.

Despite the mechanism by which MHC is detectable in scent being as-yet unconfirmed, fine-tuned detection of MHC molecules likely requires a sophisticated sense of smell, and that well-developed olfactory capabilities work alongside MHC expression (Santos et al. 2018).

### 1.1.4: Olfaction in Birds, with Focus on Seabirds from the Avian Order *Procellariiformes*

For a long time, it was incorrectly assumed that birds were anosmic, i.e., lacking olfactory ability, or that birds would at least rarely be using a sense of smell. It was thought that a lack of sniffing behaviours or scent marking, the physical shape and size of nostrils, and low levels of response to olfactory cues meant that birds overall lacked a sophisticated sense of smell (Roper 1999). In the 1960s, opinions began to change when evidence showed that, based on nasal physiology and chemoreceptor response to odours, some species of birds (turkey vultures *Cathartes aura* and some Procellariiform species) may be an exception and do indeed possess a sense of smell (Tucker 1965; Bang 1966). Advances in research have since shown that olfactory capability is not just limited to a small number of species, and instead most birds do indeed possess a sense of smell and a complex olfactory system (Bang and Cobb 1968; Wenzel 1992; Balthazart and Taziaux 2009), sophisticated enough to source food, assess

mates and pheromones and navigate home (Nevitt et al. 2008; Caro and Balthazart 2010; Gagliardo 2013).

Evolutionary studies focussing on olfactory receptor (OR) genes in birds have shown three distinct subfamilies exist (Driver et al. 2021)– the amniote alpha and gamma clade, and the γ-c clade that expanded duplicated and expanded around the radiation of extant bird lineages (Steiger et al. 2008). Since then, a multitude of studies have demonstrated the olfactory capabilities of many different bird orders and species (see Caro and Balthazart 2010, including numerous referenced studies therein). One example is the Procellariiformes, an order of tube-nosed seabirds that has been known since the 1960s to have olfactory capability, based on their well-developed and large olfactory bulbs (Bang 1966). Procellariiformes include species groups such as petrels, albatrosses, fulmars and shearwaters – all highly-pelagic species that spend most of their lives at sea, only using land to breed (Nevitt 2008). They forage at sea over large distances, and early studies found that this appeared to be olfaction-led (Verheyden and Jouventin 1994). Procellariiformes can detect natural scents such as dimethyl sulphide linked to sea upwellings, phytoplankton blooms and therefore large numbers of animal prey (Nevitt 2000). Only few studies of OR genes in Procellariiformes exist to date – Snow petrels (*Pagodroma nivea*) were found to have surprisingly few (relative to other, non-Procellariiform species) OR genes with 50% belonging to the γ-c clade (Steiger et al. 2008), whilst Northern fulmar (*Fulmarus glacialis*) also had average numbers of OR genes, with only 2 intact (Khan et al. 2015). A recent genomic study characterised OR genes in Cory's shearwater (*Calonectris borealis*), finding similar numbers of genes to other procellariiformes, more of which were intact than found in previous studies. Most of the OR genes characterised for Cory's shearwater belonged to the γ-c clade (85%), suggestive of recent gene expansion relating to scent of such odorants mentioned above. In addition positive selection for these γ-c clade genes was identified, and combined with a vast number of pseudogenes and polymorphisms, suggesting that adaptative evolution of OR genes is resulting from physiological or genomic links of olfaction with navigation, foraging and other behavioural aspects (Silva et al. 2020).

It is therefore unsurprising that Procellariiformes also use their well-developed olfactory sense to also recognise and assess individual odours of conspecifics (Bonadonna et al. 2007). Procellariiformes are notorious for having a detectable, musky scent that pervades their feathers, nests, eggs and offspring, aiding in nesting site location and – in the context of this thesis – mate choice (Bonadonna and Nevitt 2004).

A lot of research into Procellariform olfaction has focussed on use of scent to recognise nesting site. Olfaction is critical in homing process, with birds that have been made anosmic unable to find their nests (Bonadonna et al. 2001). Interestingly, it has been shown in choice-based behavioural studies that birds recognise the smell of self and non-self-nests, preferentially choosing the 'self' scented nest to that of a conspecific (Bonadonna et al. 2003). In contradiction to this, when presented directly with their own scent or that of a conspecific, petrels choose the conspecific, resulting in self-avoidance (Bonadonna and Nevitt 2004). Later studies solved this contradiction, including the scent of their mate into the choice studies. In these, it was shown that petrels will choose the scent of their mate as opposed to a conspecific. In conjunction with data on self-avoidance, it was proposed that when petrels choose their own nest over that of a conspecific, they are focussing on the individual-specific scent of the mate who also uses the nest (Bonadonna 2009), and when placed in a maze, petrels indeed consistently navigated towards the scent of their mate rather than their own scent (Mardon and Bonadonna 2009).

Whilst the link between MHC and scent has already been discussed above, the exact cause and origin of odorants in birds is not yet fully determined. The uropygial gland appears to play an important role, with males and females differing in their chemical composition of gland secretions, which may play a role in reproduction (Jacob et al. 1979). It was found that secretions changed seasonally between sexes, with the composition of male gland secretions staying consistent, whilst those of females change through the breeding season, only stabilising after a mating event (Jacob et al. 1979; Caro and Balthazart 2010). The link between secretions and oestrogen in females potentially explains these fluctuations in composition, used as a signal to males that females are ready to mate (Bohnet et al. 1991). Research into the uropygial gland has since revealed similar patterns in multiple species, with secretions regularly demonstrating

unique signatures that separate sexes or even individuals (Gabirot et al. 2018; Whittaker et al. 2018). It has also been demonstrated that uropygial gland secretions reveal differences between individuals' sex, age and even reproductive status, and profiles of gland secretions could even predict reproductive success (Whittaker et al. 2010; Amo et al. 2012; Whittaker et al. 2013). However, not all bird species possess a uropygial gland and so there is still some debate about its importance in individual recognition and sexual reproduction for all bird species (Salibian and Montalti 2009).

As mentioned above, it is thought that the binding site of MHC compounds produces peptides that create detectable odours, and that individuals can select mates based on MHC profile, but the link between these two factors is not fully understood. Using gas chromatography to assess uropygial gland secretions, it has been shown that the chemical composition of uropygial gland secretion was positively correlated with MHC similarity between black-legged kittiwakes (*Rissa tridactyla*). This suggests that uropygial gland secretions somehow translate information on MHC similarity (Leclaire et al. 2014; 2017). Uropygial gland secretions are used to preen feathers, and the compounds contained within are therefore spread onto feathers when birds groom themselves (Campagna et al. 2012). Whilst these are not volatile alone, it has been suggested that bacterial microbiota on feathers may be breaking them down to produce odorants, and that diversity of these microbiota reflects diversity of MHC profile (Leclaire et al. 2019). This link between MHC profile and feather microbiota is further complemented by the third and fifth hypotheses regarding MHC expression in odour (see 1.1.3, MHC and scent), and may provide the link between scent, MHC and mate choice in birds.

### 1.1.5: Study Species and Speciation

The Procellariiformes order of tube-nosed seabirds includes the storm petrels, a multi-family grouping which accounts for around 20 species of birds. Storm petrels fall into either the *Hydrobatidae* or the *Oceanitid*ae families and are typically small-bodied with dark-coloured plumage. Storm petrels, like all Procellariiformes, spend much of their life at sea and only come to land when breeding. They nest in burrows, laying a single egg which is cared for by both parents, taking turns to forage at sea. Storm petrels are long lived (up to 30

years) and form strong monogamous bonds with partners (Bolton et al. 2008). They also exhibit strong levels of philopatry, meaning they return to the same nesting site annually, usually the same site at which they were born (Carboneras 1992). This thesis focusses on two species of storm petrel, the band-rumped storm petrel (*Hydrobates castro*) and Montero's storm petrel (*Hydrobates monteiroi*), henceforth in this thesis collectively referred to as 'Azores storm petrels', with particular focus on the sympatric populations on this archipelago (Hoyo et al. 2014).

In 1996, storm petrels nesting on the Azores archipelago were discovered to comprise two seasonally distinct breeding populations – one breeding during the hot season, and one during the cold season (Monteiro et al. 1996). The two populations were similar in appearance and were originally thought to belong to only one species. The birds present in the two mating seasons utilise the same nesting sites on the archipelago, with only a small amount of temporal overlap between the two seasonal populations in August-October. Subsequent studies documented differences in morphometrics and diet (Monteiro and Furness 1998b). A comparison of burrow calls between the two breeding populations found that hot-season birds did not respond to calls of the cold-season conspecifics - suggesting some form of pre-mating isolation, as burrow calls are a fundamental part of acoustic signalling to locate mates, owing to their nocturnal existence (Harris 1969; Bolton 2007). Such acoustic communication is important in nocturnal vertebrates where visual cues not effective, especially for cryptic species that exist in sympatry. Similar to *H. castro* and *H. monteiroi*, cryptic species of mouse lemur also show evidence of call differences linking premating isolation and speciation (Braune et al. 2008). Later studies have indeed shown a correlation between genetic distance and acoustic differences; however this is not limited to mating calls and therefore may show importance for species recognition outside of mating (Hasiniaina et al. 2020). In some frog populations it has been suggested that sexual selection is in driving the divergence in mating calls of populations, which in turn can lead to behavioural isolation and speciation (Boul et al. 2007). Recent research suggests that diversification rates in populations is not linked to acoustic signals, when not considering the function of such signals, however diversification may be higher

where calls relate to mating functions, and more research is needed (Chen and Wiens 2020).

Genetic comparisons of mitochondrial DNA (mtDNA) demonstrated only very low levels of gene flow between the seasonal populations, and that these populations were, with few exceptions, mapping onto two distinct mitochondrial clades estimated to have separated 110,000-180,000 years ago (Friesen et al. 2007). This was concluded to result from generational changes in the timing of breeding, arising from nest site competition or exploitation of different food resources (Friesen et al. 2007; Smith et al. 2007; Bolton et al. 2008). Further ecological data on breeding and moult timing, morphometrics, dietary differences and vocalisations supported hypotheses that the seasonal populations were in fact separate sympatric species, and it was thus proposed by Bolton et al. (2008) that the two populations should be considered separate, sympatric species - Monteiro's storm-petrel (*H. monteiroi*; breeding in summer) and the band-rumped storm-petrel (*H. castro*; breeding in winter), Distribution maps (Figure 1.1 and Figure 1.2; taken from BirdLife International) show the different ranges of the two species, with the band-rumped storm petrel found across the Atlantic and Pacific, whilst Monteiro's storm petrel is endemic to the Azores, specifically three islets off Graciosa Island, and two islets off Flores Island (Birdlife International 2016). Monteiro's storm petrel is considered 'Vulnerable' according to the IUCN Red List, with a global estimate of 328-378 breeding pairs (Oliveira 2016). In contrast, the band-rumped storm petrel is listed as 'Least Concern', with an estimate of 13,100-13,700 mature individuals globally (BirdLife International 2016). In terms of migration it is thought that Monteiro's storm petrel remains within the vicinity of the breeding colony throughout the year, whilst band-rumped storm petrels migrate into the West Atlantic outside their breeding season (Bolton et al. 2008). As is the case with all storm petrels, both species only lay a single egg, which - coupled with competition with other seabirds and nest predation - results in low productivity (Bried et al. 2009; Neves et al. 2017). This is particularly problematic for the endemic and 'vulnerable' Monteiro's storm petrel, potentially limiting population growth. To mitigate this, the installation of artificial nesting chambers has proven to be successful, offering protection from predators, diminishing nest-site competition and increasing the number of successful

breeding attempts and chick fledglings (Bolton et al. 2004; Libois et al. 2012).


**Figure 1.1** Species distribution map for the band-rumped storm petrel, *Hydrobates castro*, taken from BirdLife International (2016). *H. castro* are found in colonies throughout the Atlantic and Pacific oceans, demonstrating a wide global distribution.


**Figure 1.2** Species distribution map for the Monteiro's storm petrel, *Hydrobates monteiroi*, taken from BirdLife International (2016). *H. monteiroi* are endemic to the Azores archipelago in the North Atlantic Ocean, breeding on just a few islets.

More recently, analysis of anonymous interspersed nuclear loci (Silva et al. 2016) and ddRAD sequencing (Taylor et al. 2019) has contributed further to resolving the phylogeography of the two species. Whilst some haplotype sharing between the two species has been found, *H. castro* and *H. monteiroi* form distinct genetic clusters (Silva et al. 2016; Taylor et al. 2019). Observed patterns of genetic differentiation and allele sharing between the two recently diverged species are attributable primarily to incomplete lineage sorting (Nichols 2001; Hailer et al. 2013), but with an unclear additional impact of possible (i) season

switching of some birds (i.e. breeding 'out of season'), (ii) interspecific hybridisation and (iii) long-term introgression.

At present, the two Azores storm petrel species are distinguished in-hand using morphometrics and breeding plumage, with genetic analysis confirming species if deemed necessary. Whilst this method is seemingly reliable, with low levels of misidentification (Bolton et al. 2008), development of a genetic screening method for species identification, similar to Dalén et al. (2004), could aid in efficient identification of samples of particular interest for further investigation (e.g., identification of individuals which switch between the two breeding seasons, potential hybrids, etc.). Whilst much is known about the breeding phenology of each species, there is a lack of knowledge surrounding the mate choice dynamics of both. Considering the strong monogamy exhibited, mate choice is a long-term investment which could potentially be linked to the MHC.

### 1.1.6: Evolution of the Avian MHC and its Characterisation in Azores Storm Petrels

Most genetic work on *H. castro* and *H. monteiroi* has focused on the mitochondrial control region, with further studies using anonymous nuclear loci and ddRAD sequencing (Friesen et al. 2007; Smith et al. 2007; Silva et al. 2016; Taylor et al. 2019), and so far, MHC diversity has only been characterised superficially for *H. castro* and *H. monteiroi.*

The MHC forms part of a multigene family, and as explained above, MHC genes encode for multiple different classes (Class I, II and III). Studies of avian MHC have considered both Class I and II (Minias et al. 2018), however as explained above, the role of Class II molecules in specialised immune cells makes it an interesting target for MHC-based mate choice studies; notably, a review of avian MHC has shown that non-passerines indeed show stronger selection in MHC Class II genes (Minias et al. 2018). This role in extra-cellular response has meant recent research on storm petrel MHC has focussed on the MHC Class IIb region (Burri et al. 2010; Dearborn et al. 2016), and more specifically targets exon 2, which encodes for the antigen binding site (Brown et al. 1993) of the MHC molecule. This makes it functionally important for fighting pathogens, and is prone to selection and high mutation rates as a result (Ohta 1998). Diversity at these loci is shown to be under balancing selection (Hedrick 1999), which itself

can be subject to sexual selection and a target for mate choice (Bernatchez and Landry 2003). Characterisation of MHC is further complicated in birds by the presence of an ancient duplication that predates the radiation of birds, and has resulted in the presence of two lineage-specific genes – DAB1 and DAB2 (Burri et al. 2008; Goebel et al. 2017). Presence of these two lineages is thought to have occurred through a 'birth and death' model of evolution, with the two genes created after a duplication event (Nei and Rooney 2005), and in some avian species one lineage may be lost through a deletion or gene conversion (Goebel et al. 2017). Recent studies on MHC Class IIb genes in Leach's storm petrel (*Hydrobates leucorhous*) have confirmed the presence of both DAB lineages, through the design and use of lineage-specific primers (Dearborn et al. 2014). Further to this, mate choice analyses using the MHC genotypes of mated pairs demonstrated that individuals do not mate disassortatively regarding MHC, as the duplicated lineages provide enough heterozygote advantage in offspring without individuals needing to specifically target mates based on MHC (Dearborn et al. 2016, but see Hoover et al. 2018). This demonstrates the importance of incorporation of both main avian DAB lineages and the full diversity of their allelic variation, to understand how mate choice and MHC genotype may interact in *H. castro* and *H. monteiroi*.

Studies so far have used general primers targeting exon 2 of the MHC Class IIb region, to suggest both species have at least 5 different MHC alleles, indicative of >3 loci, but do not confirm exact numbers (Burri et al. 2014). It has not yet been confirmed if *H, castro* and *H. monteiroi* still possess both DAB lineages, and it is possible that designing locus-specific primers for exon 2 of the MHC Class IIB region could confirm the presence of one or more lineage.

## 1.2: Project Aims:

### 1.2.1: Aims of this PhD Project

The two focal species of this thesis, *H. castro* and *H. monteiroi* are a recently diverged species pair for which only limited information exists about (a) the degree of reproductive isolation (e.g., the presence or absence of interspecific hybrids has not been investigated in detail), (b) the role of functional, putatively adaptive loci- such as MHC - in species divergence, (c) patterns of mate choice and how this translates from the individual to species-level genetic

differentiation. Research into these aspects is expected to add important knowledge about the patterns and mechanisms of reproductive isolation and speciation in storm petrels and will also add important information about the conservation management of the endemic and vulnerable *H. monteiroi*.

The central aim of this PhD project was therefore to develop and apply new methods suitable for species identification, MHC classification and sequencing, providing novel insights into genetic variability, differentiation, and mate choice behaviours in *H. castro* and *H. monteiroi*.

Specific aims of this thesis were to (1) develop a new assay for identification of *H. castro* and *H. monteiroi* individuals using mtDNA, to provide a rapid and effective method for screening of large numbers of samples for their mtDNA clade membership. This method will allow identification of putative 'outlier' individuals, aiming for these to be investigated for potential phenotypic and genome-wide signals of hybridisation or breeding season switching (**chapter 2**). Further aims of this thesis were to close knowledge gaps regarding adaptive genetic variation in MHC DAB lineages in *H. castro* and *H. monteiroi*, with (2) assessment of whether Azores storm petrels possess both or only one of the two main avian DAB lineages described by Goebel et al. (2017) (**chapter 3**), (3) the design of DAB lineage-specific primers for future studies of DAB lineage genetic variability in the two species (**chapter 3**). (4) Using these novel primers and high-throughput sequencing (HTS) approaches on population samples from both species, I then aimed to assess levels of genetic variability and patterns of selection between DAB1 and DAB2 loci and between the two Azores species, elucidating possible signals of divergent patterns of mate choice or natural selection, e.g., because of differing environmental conditions or parasite/pathogen pressures at different times of the year (**chapter 4**). Using data from mated pairs, this in-depth MHC characterisation was used to assess how individual mate choice translates into patterns of genetic differentiation within and between *H. castro* and *H. monteiroi* (**chapter 4**).

### 1.2.2: Chapter Structure

**Chapter 2** is the first data chapter of this thesis. A new clade-specific method to identify *H. castro* and *H. monteiroi* is described. Primers were designed around fixed base differences between the two species and, when used in a multiplexed

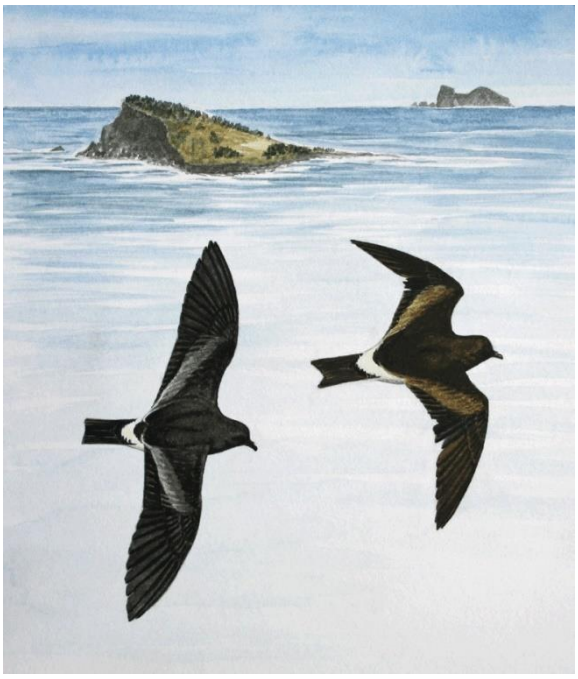
reaction, produce bands of different sizes on agarose gels. This is a novel method assay expected to prove useful in future studies on these species. I used the assay to screen population samples from both species, finding a 100% match between expected and observed mtDNA clade membership. Implications of this are discussed in the light of previous findings of mitochondrial 'mismatches' (birds with unexpected mtDNA clade membership), and their relevance for our understanding of reproductive isolation between the two recently diverged species.

**Chapter 3** details the novel discovery and sequencing of both DAB lineages in *H. castro* and *H. monteiroi*. New primers targeting exon 2 of the Class IIB region for both DAB lineages are presented, and results of their extensive testing are shown. Phylogenetic analysis confirms that the obtained sequences clearly separate into both two main avian DAB clades, as recently shown for the closely related Leach's storm petrel (Dearborn et al. 2014). The primers are DAB lineage-specific, yet general enough to capture extensive allelic diversity, known from a test panel of individuals from both species.

**Chapter 4** details the use of the primers designed in Chapter 3 for high-throughput Illumina sequencing. This allowed the determination of allele numbers for each DAB lineage in both *H. castro* and *H. monteiroi*. The chapter details the data analytical pipeline used to DAB genotypes for *H. castro* and *H. monteiroi*, respectively (data from mated pairs and their offspring). These data are used to assess patterns of Mendelian inheritance of MHC alleles, determine allelic variability at the individual and species level, sharing of alleles between species, and overall levels of MHC differentiation between the two species. Obtained results demonstrate that DAB1 tends to encompass higher genetic variability than DAB2, and that the two Azores species are clearly differentiated at DAB loci, consistent with the distinct taxonomic classification. Finally, the results show that *H. monteiroi* has retained comparatively high MHC variability despite its endemism and currently low population size.

Finally, in **chapter 5**, the findings from the PhD project are summarised, discussed, and contextualised, including a discussion of further implications and possible directions for future research. Conservation implications for *H. castro* and *H. monteiroi* are also discussed.

# Chapter Two: Novel screening method for mitochondrial clade identification reveals no signals of introgressive hybridization between allochronic species of *Hydrobates* storm petrels breeding on the Azores



*Hydrobates monteiroi (right) and Hydrobates castro (left), showing morphological differences apparent when both inhabit the islands in August. Taken from Bolton et al. (2008)*

**"It is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself."**

**- Charles Darwin**

# Chapter 2 - Novel screening method for mitochondrial clade identification reveals no signals of introgressive hybridization between allochronic species of *Hydrobates* storm petrels breeding on the Azores

## 2.1: Introduction

Recently diverged species offer evolutionary insights into processes that lead to and maintain reproductive isolation (Coyne and Orr 1998). Some level of gene flow between taxa can persist despite speciation, potentially triggered by behavioural, environmental or demographic changes (Nosil 2008). Studies of these cases can be urgent from a conservation perspective, especially if the species involved are rare and/or threatened.

One such case is provided by storm petrels (genus *Hydrobates*) on the Azores, where a rare endemic species, Monteiro's storm petrel (*Hydrobates monteiroi*), occurs in sympatry with the geographically widespread band-rumped storm petrel (*H. castro*). Together with the Cape Verde storm petrel *H. jabejabe*, these taxa form a species complex hereafter referred to as *H. castro (s.l.)*, or the 'band-rumped species complex'. In contrast to the endemism of *H. monteiroi*, *H. castro* has a breeding distribution spanning the Pacific and Atlantic oceans, including populations breeding on the Azores islands (BirdLife International 2020). The two species co-occur on the Azores, breeding in different seasons, one in the summer, 'hot season' (*H. monteiroi*), and one in the winter, 'cold season' (*H. castro*) (Monteiro et al. 1996), with both species exhibiting strong natal philopatry. Originally considered as two conspecific seasonal populations of *H. castro*, subsequent investigations into the two populations reported significant differences in morphometric data, including body weight/size, tail shape and egg size (Monteiro and Furness 1998a). Studies of mitochondrial DNA (mtDNA) control region sequence data in *H. castro* later provided further evidence for

species-level distinction of the two populations in the Azores. Smith et al. (2007) found distinct lineages for the hot- and cold season breeding populations, with a mtDNA control region sequence divergence between the clades of 1.7-2.1% (Smith and Friesen 2007). In 2008, the two seasonally distinct populations were recognised as two separate species, based on differences in morphometrics, vocalisations, diet, breeding phenology, and genetics (Bolton et al. 2008a). The summer-breeding population was thus recognised as a newly described species – Monteiro's storm petrel, (*Hydrobates monteiroi*). Friesen et al. (2007) estimated that the two species diverged 110,000-180,000 years ago, likely as a result of sympatric speciation by allochrony.

The role of allopatric isolation as a driver of speciation has long received attention in the literature (Mayr 2013). However, evolutionary processes governing genetic differentiation in sympatry are somewhat less well understood, and appear to be inherently complex (Friesen et al. 2007; Grant and Grant 2009). Besides *Hydrobates* storm petrels as a prominent case of ecology-driven speciation (Friesen et al 2007), several examples of recent ecology-driven population differentiation in sympatry exist: in such cases, genetic differentiation is low (e.g. crossbills: Edelaar et al. 2012; no genomic characterisation conducted) and can be restricted to only a few loci deemed to be involved in ecological adaptation to non-shared resources (seedeaters: Turdek et al 2021). In cases of deep divergence and secondary contact, the exchange of genes in sympatry is accompanied by non-monophyly of genetic markers (e.g. ducks: Peters et al. 2007). Furthermore, mitochondrial genetic differentiation despite absence of evidence for reproductive isolation has been described, e.g. for redstarts, possibly resulting from introgression from an unidentified extinct lineage, or mtDNA differentiation through stochastic genetic drift (Hogner et al. 2012). Despite this, nuclear and mtDNA differentiation of sympatric *Hydrobates* storm petrels remains relatively little studied.

Sequencing of 12 anonymous nuclear loci has further contributed to resolving the phylogeography of *H. castro* (Silva et al. 2016). The study also included both previously published, and novel mitochondrial data, complementing earlier research (Friesen et al. 2007c; Smith and Friesen 2007; Smith et al. 2007b), confirming previous results in terms of strong overall genetic differentiation between the two species. However, clustering of nuclear loci was far less

distinct, and haplotype-sharing between the two species was found to be extensive. Multi-locus clustering analysis suggested that four genetic clusters were most likely – (1) *H. monteiroi* in the Azores, (2) *H. castro* in the Azores/Madeira, (3) *H. castro* in Japan, and (4) *H. castro* in the Galapagos - reflecting those described by Smith et al. (2007). More recently, ddRAD data was used to investigate the extent of cryptic speciation within the *H. castro* species complex (Taylor et al. 2019). This study found evidence of at least six different geographically distinct clusters – (1) Cape Verde (recognised as a separate species in 2012 - Cape Verde Storm Petrel, *H. jabejabe*; Sangster et al. 2012), (2) Azores (North Atlantic), (3) hot season (*H. monteiroi*), (4) Galapagos, (5) South Atlantic, and (6) Japan/Hawaii (*H. castro*), reinforcing that *H. monteiroi* and *H. castro* in the Azores are two distinct species (Figure 2.2). Such divergence between species and populations is further enhanced by the strong natal philopatry of both *H. castro* and *H. monteiroi*, suggesting that dispersal between geographic areas is unlikely.

Despite the confirmed divergence of the two species in the Azores, instances of mtDNA clade sharing have been observed. In Silva et al. (2016), a shared mtDNA haplotype was observed between *H. monteiroi* and *H. castro* in the Azores, and previous work by Friesen et al. (2007) and Smith et al. (2007) documented a few individuals with haplotypes that grouped in clades not expected for that location/breeding phenology. Similarly, mtDNA haplotype networks by Taylor et al. (2019) showed a total of 17 shared haplotypes between the Azores hot season, Cape Verde and North Atlantic populations. Later in this chapter, phylogenetic analysis reveals that some individuals present as 'mismatches', sampled in one location yet they cluster within a clade dominated by samples from a different geographic location.

These 'mismatches' within the *H. castro* species complex could have various explanations; as suggested in Taylor et al. (2019), gene flow, breeding plasticity or incomplete lineage sorting could be occurring, or hybridisation between *H. castro* and *H. monteiroi*, suggested by Silva et al. (2016). Additional explanations could be genuine vagrancy, or individuals residing on site out of their breeding season. The two species diverged recently, and disparities between nuclear and mtDNA data could be explained by incomplete lineage sorting, as observed in Galapagos petrels *Pterodroma phaeopygia* and Hawaiian petrels *P. sandwichensis*

(Welch et al. 2011b). Introgressive hybridisation between two diverged species is also conceivable, which could even contribute to the evolution of reproductive isolation (e.g. in the Yellow-rumped warbler *Setophaga* spp. complex (Toews et al. 2016) and Mediterranean sparrows *Passer* spp. (Hermansen et al. 2011)). Hybridisation has been documented for other petrel species, for example among at least two species of *Pterodroma* petrels on Round Island, Mauritius (Brown et al. 2010). Whilst the potential occurrence of hybridisation is so far undocumented in the Azores, the presence of shared haplotypes and mixed ancestry (Friesen et al. 2007; Silva et al. 2016) could indicate the presence of hybrids.

Genetic research has been fundamental in differentiating between *H. castro* and *H. monteiroi*. Between August and September both species inhabit their breeding islets on the Azores at the same time, and during this overlap period, morphometrics, moult presence and breeding status are used to discern between the two species (Bolton et al. 2008). However, these criteria are more difficult to use away from the breeding colonies, especially where other populations of *H. castro* breed in the summer months (Bolton et al. 2008). It is also suspected that *H. monteiroi* remains in the vicinity of the breeding islets year-round, which could confuse matters if caught out of its breeding season (Bolton et al. 2008). In these cases, there is potential for species misidentification, which may explain the published, apparent mismatches between sampling location and phylogenetic clade – rather than introgression or incomplete lineage sorting.

Such mismatches appear to be rare - their characterisation has been limited and their evolutionary origins remain largely unexplored. To enable studies of these mismatches, development of a rapid, cheap method to screen Azores storm petrel samples and identify mismatches could prove useful. Singling out mismatched individuals for further research would enable more detailed phenotypic or genomic characterisation, which could in turn lead to explanations for mismatches. Furthermore, a method to rapidly diagnose individual clades could be beneficial during the overlap period between *H. castro* and *H. monteiroi*, when both species are present and accurate sampling based on morphology may need secondary genetic confirmation.

The use of species-specific PCR assays has been proven to aid in species identification (Palomares et al. 2002), and when designed to amplify different fragment sizes, primers can be used in a multiplexed reaction (e.g., Dalén et al. 2004). The resulting PCR products will present as different sized fragments depending on which species is being amplified, allowing for species to be determined without sequencing. Currently, such mtDNA sequencing is used to separate *H. castro* and *H. monteiroi*, which is expensive and time-consuming. In this study, a cheap, quick screening method to identify mismatches was developed, bypassing the need for sequencing. Following the approach of Dalén et al. (2004), a multiplexed PCR assay was designed to identify storm petrels breeding in the Azores, assigning individuals to a mtDNA clade and allowing identification of any mismatches between the observed and expected clade. Such identified mismatches could then be analysed further for other genotypic and phenotypic characteristics (e.g., morphometrics, vocalisations, behaviour, genomics).

## 2.2: Aims

The core aim of this chapter is to design a simple and affordable way to identify clade at the gel-electrophoresis level, through the design of clade-specific primers that utilise different fragment sizes to reveal mtDNA clade identity at the gel-electrophoresis stage. This would provide a simple and affordable way to screen for mismatches, where an individual's geographic sampling location does not match their assigned clade. Using clade-specific primers could reduce the need to sequence every sample, providing a comparatively low-cost and simple alternative to sequencing. The two designed, clade-specific primers were then aimed to be combined in a multiplexed PCR reaction, with resulting PCR products yielding a diagnostic banding pattern for each Azores species on agarose gels (presence/absence of each clade-specific band expected to yield an unambiguous result for each individual). By combining the primers in a multiplex PCR reaction, samples need to only be tested once, as opposed to being tested twice with two separate primer mixes, reducing the time and volume of reagents used.

When used on Azores storm petrel DNA samples, these primers would provide a suitable method for screening large numbers of storm petrels sampled on the

islets of Vila and Praia in the Azores. This method aims to discern which clade these similar species belong to, where inexperience of morphological differences could lead to misidentification. It is hoped that these primers will be used in future studies on these islets, and this method can be used to inspire similar clade-specific or species-specific diagnostic tools.

## 2.3: Methods

### 2.3.1: Sample Collection

Blood samples were collected from both species of storm petrels. Researchers working in the Azores have been collecting blood samples since the early 2000s, with the most recent samples presented here on Praia islet in June 2017 (Robert et al. 2014). Most samples were taken on the more easily accessible Praia islet (39.0555° N, 27.9425°W, 0.12 $km^2$, summit 59 m), 1km east off Graciosa Island, Azores (Bolton et al. 2004), with additional samples taken from Vila (*H. castro* only) and Baixo islets. In total, 536 blood samples, 224 of *H.* castro and 312 of *H. monteiroi*, were collected from mated pairs and their chicks between 2000 and 2017. On Praia islet, most of the birds sampled were breeding in artificial nesting chambers, with some utilising natural sites for nesting (Bolton et al. 2004). Currently, it is not known if genetic differentiation is observed between these islets. The close proximity of Baixo and Praia islet (both situated within 1 km of Graciosa islet) may preclude such differentiation, whilst Vila islet is around 300 km away from Baixo and Praia and may show different patterns.

Birds were temporarily kept in bird bags and checked for the presence of a metal ring on the leg containing a unique ID code – if already present, this was recorded, and the database checked to see if a blood sample had already been taken in previous years. If a blood sample had already been taken, then birds were weighed and measured before being returned to the nest; if no blood sample was recorded, a sample was taken along with measurements and weights. If no leg ring was present, a new ring was applied to the leg of the bird, a blood sample taken, and all measurements recorded.

Blood was collected from the brachial vein of the wing using a sterile 25G Whatman™ hypodermic needle (VWR / 16 mm) (Owen 2011). A small droplet of blood was left to pool on the skin surface after puncture and approximately 50 µL collected using a fresh Drummond Microcaps™ glass capillary tube (Sigma

Aldrich). Blood was expelled using a pipetting bulb, into a 2 mL screwcap Eppendorf containing 700 µL 1M Queen's Lysis buffer (pH 7.5), which was prepared following Seutin et al. (1991). To ensure that blood flow to the puncture wound ceased, pressure and cotton wool were applied. Also, an antibacterial wound powder was utilised to aid healing. Post-sampling, all birds were checked to guarantee bleeding had ceased, before being returned to the nesting chamber. Additional measurements taken included the weight and wing length of each bird. If eggs were present in the nest, the width and length of these were also recorded. In total, 42 new samples of blood were collected in 2007, amounting to 24 new pairs, inclusive of those sampled as chicks in previous years which had returned to the colony to breed. Samples were stored at room temperature until returning to the lab, where they were then stored in a refrigerator at approximately 4 °C.

The work was carried out with the permission of The Nature Park of Graciosa and with assistance from colleagues from the University of the Azores and SPEA (Sociedade Portuguesa para o Estudo das Aves) under licence numbers 33/2017/DRCT and No60/2017/DRA (Secretaria Regional da Energia, Ambiente e Turismo, Direcção Regional do Ambiente da Região Autónoma dos Açores).

All sampling was carried out with minimal disturbance, following precautions to maintain animal welfare and to ensure that birds were not adversely affected.

### 2.3.2: DNA extraction

DNA was extracted using a protocol based on (Bruford et al. 1998), provided by the NERC (Natural Environmental Research Council) Biomolecular Analysis Facility (NBAF) at Sheffield. All buffers and reagents listed were created according to the instructions provided in the protocol.

Volumes of 20 µL, 60 µL and 100 µL from the samples OMA5, OC149 and OC036 were used to determine the suitable amount of blood for the extraction, and DNA concentration of extracts was measured with an Invitrogen QuBit™. No relationship between blood volume and DNA concentration was observed and a volume of approximately 50 µL was used for all extractions.

Due to the viscosity of the mixture of blood with Queen's Lysis Buffer, 200 µL pipette tips were cut to aid suction. Consequently, blood volume values reflect

the setting of the pipette volume, and not the actual volume of blood. Blood samples (approx. 50 µL) in Queen's Lysis buffer were transferred to 1.5 mL Eppendorf tubes with 30 µL of proteinase K (Qiagen, Germany) and 250 µL DIGSOL buffer (created per the extraction protocol, provided by NBAF at Sheffield) were added. Samples were vortexed for 30 seconds and then placed on a shaking heater set at either 37°C overnight, or 55 °C for 3 hours. Once the blood sample had been fully digested, 300 µL of 4 M ammonium acetate was added to each tube, thoroughly vortexed and centrifuged at 8 °C for 30 minutes at 15,000 rpm. Approximately 450 µL of the supernatant was pipetted into a new, sterile 1.5 mL eppendorf, 1 mL of cold, 100 % ethanol was added and samples were briefly vortexed before being inverted 10-20 times. At this stage, DNA was usually visible in a precipitated state, consisting of thin white strands. The DNA was pelleted by centrifuging for 15 minutes at 6 °C and 15,000 rpm and the supernatant was carefully pipetted away. The pellet was washed twice with 900 µL of cold 70 % ethanol by inverting tubes gently 10-20 times. In the event of the pellet dislodging from the tube, it was again centrifuged at 6 °C and 15,000 rpm for 15 minutes. After washing, the pellet was dried in an Eppendorf™ Concentrator 5301 for 15 minutes at 45 °C. Once fully dried, 50-100 µL of T-low-E was added to the tube as depending on the size of the pellet. Pellets were resuspended by briefly shaking and placing tubes in a heat block at 56 °C for 30 minutes and. Extracts were stored at -20 °C. During each round of extractions, a negative sample containing no DNA was included as control for contamination.

### 2.3.3: Control region amplification

Initial lab methods focussed on amplifying the mitochondrial control region of both band-rumped storm petrels (*H. castro*) and Monteiro's storm petrels (*H. monteiroi*). To design clade-specific primers, the divergence of the two species at the control region had to be established. Smith et al. (2007) have already designed primers to amplify the control region in these two species , which have been used to great success in subsequent studies (Friesen et al. 2007; Silva et al. 2016; Taylor et al. 2019). To amplify the samples extracted in the lab, primers H521 (5'-ATGGCCCTGACATAGGAACCAGA-3') and OcL61 (5'-CAGTAGCGGGGCGGCTYTATGTAT-3') were ordered from Sigma Aldrich (USA).

All PCRs were run on an Applied Biosystems SimpliAmp thermal cycler (ThermoFisher Scientific), using a GoTaq® G2 Flexi DNA Polymerase kit. The PCR mix consisted of 1X GoTaq Green Buffer, 2.5 mM $MgCl_2$, 0.2 mM of each dNTPs (Promega), 0.3 µM of each primer (Sigma Aldrich), 0.03 U/µL of DNA Polymerase and 1 µL of genomic DNA. Each reaction was made up to a 15 µL reaction volume using 7.7 µL Affymetrix Ultrapure Water (ThermoFisher Scientific).

PCR was carried out using the following conditions: initial denaturation step of 95°C for 2 minutes, 35 cycles of denaturation at 95°C for 30 seconds, an annealing stage using a temperature gradient of 60°C, 62°C, 65°C, 68°C and 70°C for 45 seconds and an extension stage of 72°C for 1 minute 30 seconds. A final extension step of 72°C for 5 minutes concluded the PCR profile, and samples were kept at 15°C until removal from the PCR machine. A single sample of nuclease-free water was included in each PCR as negative (no template) control. PCR was repeated with new reagents if the negative control demonstrated a positive PCR result. Only a small number of such repetitions were necessary.

PCR products were visualised after agarose gel electrophoresis (1.5% agarose gel from 100 µL 0.5x TBE (Tris-borate-EDTA) buffer and 1.5 g Bioline agarose powder) under UV light after staining with 1 µL Invitrogen™ SYBRSafe.

An aliquot of 3 µL of PCR product was used in each well of the gel. For every gel run, the first well was filled with 2 µL of Promega ™ 100 BP ladder, to act as a reference for band sizes. Gel electrophoresis was run for 45 minutes at 120 V and 120 W. Gels were photographed using a GelDocIt UVP system and visualised using VisionWorks LS (Version 6.8). Results from the temperature gradient PCR showed 70 °C to be an optimum annealing temperature. All subsequent PCRs used 70 °C during the annealing stage, maintaining the same mix and general PCR profile.

All PCR products showing successful amplification on agarose gel were sent for Sanger sequencing with Eurofins Genomics (Eurofins MGW Operon, Ebersberg, Germany). PCR products were diluted 1.5 µL product in 13.5 µL ultrapure water, with 2 µL of forward or reverse primer. Forward and reverse sequences returned from Eurofins were aligned in Geneious 9.4 against a *H. castro* reference sequence for their species (DQ178708; *Oceanodroma castro* isolate G

control region, partial sequence, mitochondrial), originally isolated by Smith et al. (2007) and accessed via Genbank (Benson et al. 2005). Primer binding sites were located, and sequences were trimmed so that primers H521 and OcL61 were not included in the sequence. The aligned sequences were scanned for any instances whereby the sequencer had miscalled bases or assigned an 'N' where a clear peak was visible. In these cases, bases were corrected according to the visible peak, providing a more accurate sequence.

### 2.3.4: Phylogenetic tree construction using mtDNA control region data

The mitochondrial control region has already been investigated in these species (Friesen et al. 2007; Silva et al. 2016; Taylor et al. 2019), with sequences readily available on GenBank. In addition to the 94 consensus sequences created from Sanger data, 754 additional mtDNA control region sequences were retrieved using supplementary information from Taylor et al. (2019) and added to Geneious. To provide an outgroup, the Leach's storm petrel *Hydrobates leucorhous* sequence extracted in the lab, and Swinhoe's storm petrel from Genbank (*Hydrobates monorhis*; KR873325.1) were also added, creating an alignment of 848 sequences. Due to the control region amongst these samples being conserved, duplicate sequences were common. Using the 'Find Duplicates' feature in Geneious, sequences with identical residues were identified and unique sequences were extracted into a new alignment of 327 unique haplotypes. This reduced alignment was subsequently used to construct phylogenetic trees.

The reduced alignment was used in jMmodelTest2 (Guindon and Gascuel 2003; Darriba et al. 2012), to find the most likely phylogenetic model. JModelTest analysed 11 substitution schemes, with models that had unequal frequencies (+F), gamma variation (+G) and invariable sites (+I) included. The analysis was run using a BIONJ base tree to calculate likelihoods. A total of 88 models were analysed, and AIC, AICc, BIC and DT values were calculated for all. Using these values, a HKY+I+G phylogenetic model (Hasegawa et al. 1985) was selected, which was supported by AICc, BIC and DT calculations as being the optimum model to use.

To create a phylogenetic tree, the command-line programme IQ-TREE (version 1.6.12) was used. IQ-TREE uses a maximum-likelihood approach to construct a

phylogeny, comparable with other similar programmes such as RaxML (Nguyen et al. 2015b). Phylogenetic tree construction followed the HKY + I + G model suggested by JModelTest2, with 1000 replicates using the ultrafast bootstrapping (UFBoot) implemented in IQ-TREE (Minh et al. 2013; Hoang et al. 2018). Leach's storm petrel was set as an outgroup for the tree, as the least-closely related organism to the rest. The resulting consensus tree file was imported into FigTree 1.4.2 (Rambaut, 2009) for further annotation.

### 2.3.5: Design and Application of Clade Specific Primers

The published sharing of haplotypes between geographic locations (Friesen et al. 2007; Smith and Friesen 2007; Smith et al. 2007; Silva et al. 2016, Taylor et al. 2019) and subsequent phylogenetic clustering meant primers needed to be clade-specific, discerning between the North Atlantic *H. castro* and Azores-endemic *H. monteiroi.* Method development began in 2017, using sequences extracted in the lab (Section 2.2.3) (n=66) and all Azores Genbank sequences available at the time (n=58) (Friesen et al. 2007; Smith and Friesen 2007; Smith et al. 2007; Silva et al. 2016). These sequences were used to create an alignment in Geneious, along with two *H. leucorhous* sequences (also extracted in the lab) to act as an outgroup (total sequences n= 128). Next, a phylogenetic tree was constructed using IQTree (Minh et al. 2013; Hoang et al. 2018). The ModelFinder function determined a HKY+F+R3 model as most suitable using BIC criterion, which was carried out with 100 bootstraps. The clades presented by the phylogenetic tree were used to divide the sequences into two alignments – one representing North Atlantic, cold season *H. castro* and another for Azores, hot season *H. monteiroi.* Each GenBank sequence in the alignments were cross-referenced against their published sampling location to determine if sequences clustered as expected. The consensus sequences produced by the two different, clade-specific alignments were scanned for sites where a fixed base difference was present between the two Azorean clades. Three suitable sites with a fixed base difference between the clades were found at base pair 35, 193 and 266 (Figure 2.1). A total of 8 primers were manually designed, placing the fixed base difference in the 3' end of the primer (Table 2.1). This was to ensure that during elongation by polymerase the sequence continued to be clade-specific, based on this fixed difference. This also ensured maximum likelihood of primer binding at the clade-diagnostic site. Geneious was used to create consensus alignments for

each clade. Primers were tested *in silico* using the 'Test Primers' feature in Geneious, to ensure that the primers only bound to the sites and clades for which they were designed. Once confirmed to amplify the clades for which they were designed, primers were ordered from Sigma-Aldrich.

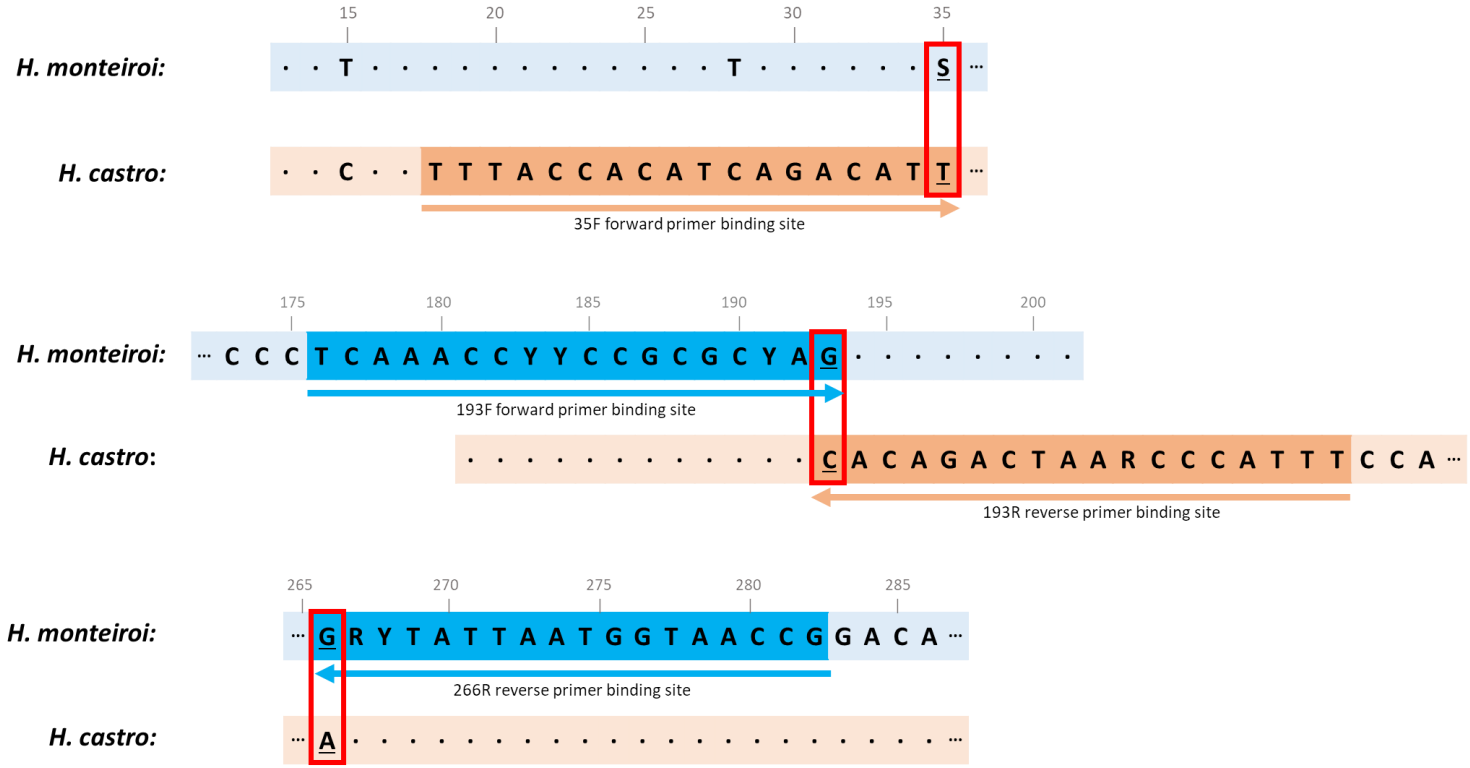



**Figure 2.1** Clade-specific primer design, focussing on fixed nucleotide differences between individuals from the *H. monteiroi* (Azores) and *H. castro* (North Atlantic) clades. A · indicates a base that is identical between the clades. Where a base-difference occurred, the appropriate IUPAC code letter is used for each clade. The fixed base differences around which primers were designed, are indicated with a red box displaying the exact base difference. A triple dot (…) indicates a continuation of sequence

Attempts to optimise all six possible primer pair combinations were made, aiming for a divide whereby they only bound to individuals for which they were designed. To test these primer pairs for clade-specificity, four individuals were selected for use in PCRs. These individuals were selected based on our sequencing with H521 and OcL61, whereby their clade had been confirmed in the phylogenetic tree. Six of the possible primer combinations proved to be non-specific and amplified all individuals non-discriminately. Two primer combinations were found that worked as desired – 35F CASTRO + 193R CASTRO and 193F MONT + 266R MONT - with each primer pair amplifying the one clade it was designed for.

**Table 2.1** Sequences of the primers designed to amplify in a clade-specific manner. Fixed base differences are highlighted in **bold underline**. These four primers (one pair per clade) were selected from a larger set of eight primers, for their ability to provide clear, specific bands.

| Primer Name | Sequence (5'-3') | Target clade and season |
|---|---|---|
| 35F CASTRO | TTT ACC ACA TYA GAC AT**T** | North Atlantic, cold season<br>*H. castro* |
| 193R CASTRO | RAA TGG GYT TAG TCT GT**G** | North Atlantic, cold season<br>*H. castro* |
| 193F MONT | TCA AAC CYY CCG CGC YA**G** | Azores, hot season<br>*H. monteiroi* |
| 266R MONT | CGG TTA CCA TTA ATA AY**C** | Azores, hot season<br>*H. monteiroi* |

The primers were initially optimised to work on their relevant clade in a single PCR reaction. Optimisation steps included altering concentrations of PCR reagents, cycle numbers, and annealing time and temperatures. PCRs were conducted using a GoTaq® G2 Flexi DNA Polymerase kit, consisting of 1X GoTaq Green Buffer, 2.5 mM $MgCl_2$, 0.2 mM of each dNTPs, 0.3 µM of each primer, 0.03 U/µL G2 Flexi Taq Polymerase, and 0.5 µL genomic DNA. The reaction volume was brought up to 15 µL with 7.7 µL Ultrapure water. A temperature gradient with differing annealing temperatures of 40 °C, 45 °C, 50 °C, 55 °C and 60 °C was tested to find an optimal temperature for use in a PCR profile. The final PCR profile used began with a 2-minute denaturation step at 95 °C, followed by 35

cycles of denaturation at 95 °C, annealing at 50 °C and extension at 72 °C for 30 seconds each. A final extension step of 72 °C for 5 minutes completed the PCR profile.

PCR products were run on a 2 % agarose gel, made with 100 µL 0.5x TBE and 2 g agarose powder. The gel was stained with 1 µL SYBRSafe, and each well had a 3 µL aliquot of PCR product. Gel electrophoresis was run for 50 minutes at 110 V and 110 W, and results photographed using a GelDocIt UVP system and visualised using VisionWorks LS (Version 6.8).

In total, 10 samples were tested with this PCR profile, to check a pattern of clade-specific amplification. Once this was confirmed, the primers were tested in a multiplexed reaction. Here, the primers were made into two combined primer mixes – '*H. monteiroi* clade', consisting of 10 µM each of 193F MONT and 266R MONT, and '*H. castro* clade', consisting of 10 µM each of 25F CASTRO and 193R CASTRO. These primer mixes were then both used in the same PCR reaction. Initial PCRs were run without any PCR additives; however, weak band strength and non-specific banding necessitated the addition of DMSO and BSA (Farell and Alexandre 2012). Tests with these additives were carried out, comparing the use of them separately and together. It was found that using the additives in tandem significantly increased band strength, with minimal to absent non-specific banding (Appendix A2.1). The final PCR mix consisted of 1X GoTaq Green Buffer, 2.5 mM $MgCl_2$, 0.2 mM of each dNTPs, 0.3 µM of each primer, 6% DMSO, 1% BSA, 0.03 U/µL G2 Flexi Taq Polymerase, and 1 µL genomic DNA. The reaction volume was brought up to 15 µL with 6.05 µL Ultrapure water. Again, a temperature gradient with differing annealing temperatures of 48 °C, 50 °C, 52 °C and 54°C was initially tested. The final PCR profile used began with a 2-minute denaturation step at 95 °C, followed by 35 cycles of denaturation at 95 °C, annealing at 49 °C and extension at 72 °C for 30 seconds each. A final extension step of 72 °C for 5 minutes completed the PCR profile. PCR products were again viewed on an agarose gel, using the same specification as to when using the primers alone, with the *monteiroi* clade expected to yield PCR products at approximately 100 bp long, and the *castro* clade to yield products approximately 200 bp long.

In total, 242 samples from the Azores were screened using this method, consisting of 128 samples previously identified in the field as *H. castro,* and 114 samples previously identified as *H. monteiroi* (94 were sent for sequencing). In addition, 3 samples of *H. castro* from St Helena and 4 samples from Ascension Island were also tested. To ensure that the primers did not bind to other similar species, the method was tested on Leach's storm petrel (*Hydrobates leucorhous)* and European storm petrel (*Hydrobates pelagicus*). For both species, the primers did not anneal, and only very faint, non-specific banding was visible, confirming that the primers work as desired and only amplify the clades for which they were designed.

### 2.3.6: *in silico* testing of clade-specific primer binding

To determine the likely performance of the primers on existing sequence data from GenBank and on sequences obtained in the present study, the 'test with saved primers' function in Geneious was used. For this, a haplotype list using the full dataset of 848 individuals (see section 2.2.4) was created using DNAsp (Librado and Rozas 2009). Using the haplotype groups created in DNAsp, individuals that shared haplotypes were selected and clade-specific alignments were created in Geneious. Using the 'Test with saved primers' function in Geneious, the binding of the *castro*- and *monteiroi*- clade primers were tested on these clade-specific alignments. Each forward and reverse primer was selected in turn, targeting the whole fragment in each clade-specific alignment. Mismatches between primer and sequence were not permitted, and default settings were left for Tm calculation. This was done to assess whether the *castro* clade primers could be used across the geographic distribution of the species, or if using these primers outside of the Azores would prove unsuccessful or inconclusive.

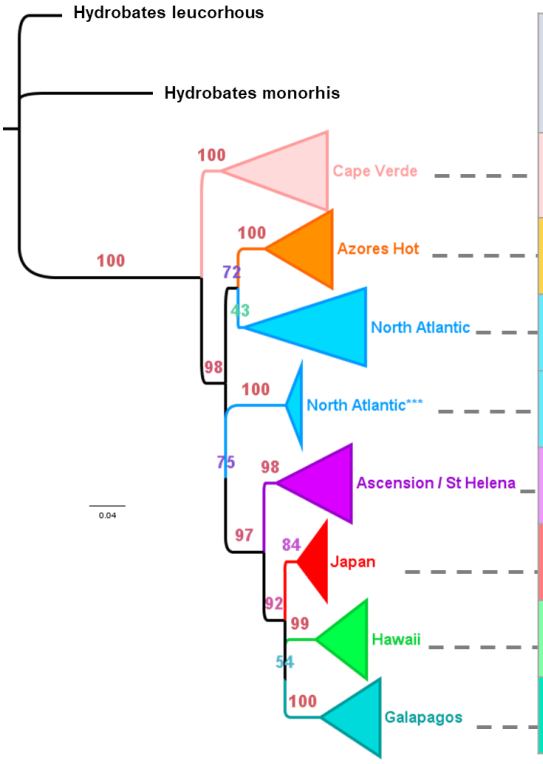
## 2.4: Results

### 2.4.1: Phylogenetic tree construction using MtDNA control region data

Once Sanger sequences were aligned, edited and trimmed, a consensus sequence of 450 bp for each individual was extracted. In total, 37 *H. castro* and 57 *H. monteiroi* were sequenced and analysed in this way (n=94 in total). When combining the new sequences with the extensive dataset by Taylor et al. (2019),

a total of 846 sequences for the two species (698 individuals for *H. castro* and 148 individuals for *H. monteiroi*), comprising 327 distinct haplotypes.

An alignment of the 848-individuals alignment revealed a degree of haplotype sharing between geographic locations. Several individuals from the North Atlantic, Azores Hot (*H. monteiroi*) and Cape Verde samples were found to cluster within mitochondrial clades that did not match expectations based on their sampling locations (Figure 2.2). Notably, none of the samples extracted and sequenced as part of the present study included mismatches, so all mismatches presented were based on previously published results.




| | Cape Verde (*H. jabejabe*) | Azores Hot (*H. monteiroi*) | North Atlantic (*H. castro*) | Ascension (*H. castro*) | Japan (*H. castro*) | Hawaii (*H. castro*) | Galapagos (*H. castro*) | TOTAL |
|---|---|---|---|---|---|---|---|---|
| Cape Verde | 190 | 0 | 0 | 1 | 0 | 0 | 0 | **191** |
| Azores Hot | 3 | 143 | 0 | 0 | 0 | 0 | 0 | **146** |
| North Atlantic | 4 | 5 | 288 | 1 | 4 | 0 | 0 | **302** |
| North Atlantic*** | 0 | 0 | 3 | 0 | 0 | 0 | 0 | **3** |
| Ascension / St Helena | 0 | 0 | 0 | 102 | 0 | 0 | 0 | **102** |
| Japan | 0 | 0 | 0 | 0 | 27 | 0 | 0 | **27** |
| Hawaii | 0 | 0 | 0 | 0 | 0 | 11 | 0 | **11** |
| Galapagos | 0 | 0 | 0 | 0 | 0 | 0 | 64 | **64** |

**Figure 2.2** Maximum Likelihood tree from IQ-TREE (Hoang et al. 2018) of mitochondrial control region data for *H. castro* and *H. monteiroi* (in total 327 haplotypes from 848 individuals, derived from Taylor et al. (2019) and sequences from the present study). Bootstrap support values are shown for the main clades. Next to the tree, the number of individuals from each geographic location contained in each clade is shown, demonstrating the sharing of some clades among locations/populations (including 'mismatched' individuals). The clade labelled 'North Atlantic***' corresponds to the three individuals from the Desertas hot season (see main text).

With only control region data (in comparison to the addition of COI data in Taylor et al. (2019)), branch support values were high for some well-resolved clades (Cape Verde, Azores Hot, Ascension / St Helena, Hawaii and Galapagos), but lower for other clades such as the 'North Atlantic' one. In contrast to Taylor et al. (2019), this tree included an additional, separate North Atlantic clade. This

clade consisted of 3 individuals from Desertas, Madeira, sampled during the hot season. The support value for this branch was only 75, and within a haplotype network (see A2.2) only few mutational steps (n=10 for *H. leucorhous* and n=17 for *H. monorhis*) separated these three individuals from the two outgroup species.

No 'mismatched' mtDNA haplotypes were found among the newly obtained sequence data. All *H. castro* samples (n=37) sequenced as part of this study clustered within the expected North Atlantic clade, and all *H. monteiroi* (n=57) clustered in the Azores Hot clade. This presents a 0% rate of mismatched haplotypes in 94 sequences. In comparison, sequences taken from Taylor's (2019) dataset, contribute to a total of 18 mismatched haplotypes spread out across the phylogenetic tree.

The Ascension Island and Pacific locations present no mismatches in their respective clades. The North Atlantic clade presents the highest degree of haplotype sharing, with a total of 14 unexpected individuals from Japan, Ascension, Cape Verde and the Azores Hot season clustering within that clade. In the Azores Hot clade, there are 3 unexpected Cape Verde individuals, and in Cape Verde, there is 1 unexpected Ascension individual. Amongst 848 individuals, 17 mismatched / shared haplotypes demonstrate that the occurrence of such mismatches is relatively low (1.6%).

### 2.4.2: PCR screening and sequencing of clade-specific primers

Initial screening with control region primers (H521 and OcL61 (Smith et al. 2007)) resulted in 94 samples (37 *H. monteiro* and 57 *H. castro*) being sent for Sanger sequencing, producing a 450 bp sequence once aligned and trimmed in Geneious. Once it was confirmed that sequences reflected the mitochondrial control region, clade-specific primers were designed for use in a multiplexed PCR reaction, which was used to screen samples and assign clade through gel electrophoresis.

In total, 128 *H. castro* and 114 *H. monteiroi* were screened using the multiplexed primer method. It was expected that when both primers were combined in a single PCR reaction, only *H. castro* samples should amplify with *castro*-clade primers, presenting a PCR band of approximately 200 bp. In contrast, it was anticipated that *H. monteiroi* samples should only amplify with *monteiroi-clade*

primers, resulting in a band of approximately 100 bp. In 100% of cases, the expected relationships were observed, and all samples amplified with their expected primer set (Figure 2.3). Of these individuals, 94 had already been screened with the control region primers (H521 and OcL61) and were of known origin and phylogenetic lineage confirmed to have been sampled only in the Azores. Only two *H. monteiroi* samples failed to amplify at all in PCR, producing no discernible band on agarose even when repeated. This is potentially due to poor sample quality or low DNA concentration. A lack of result is in practice preferable over a false result for such cases (Appendix 2.1).

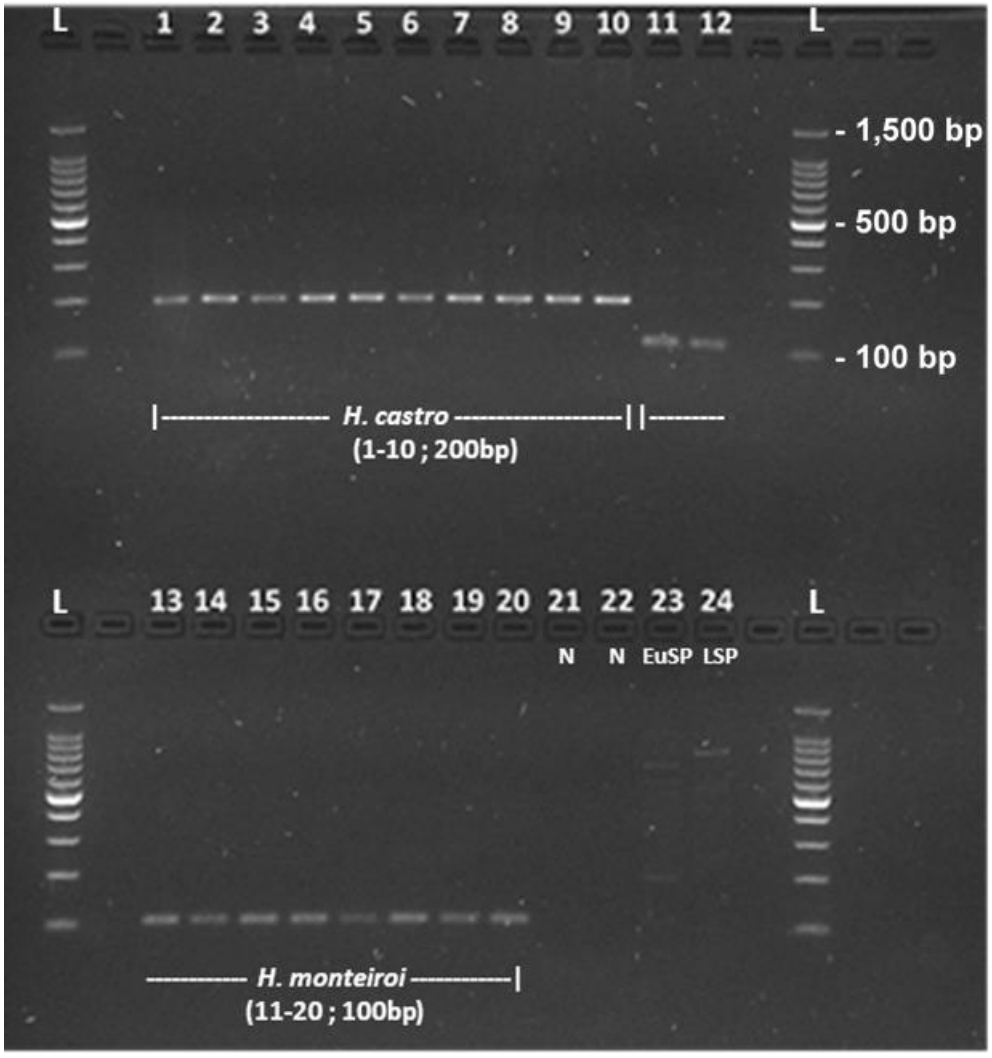



**Figure 2.3** Gel electrophoresis image displaying a clear difference in fragment length between *the H. monteiroi* and H. castro for the designed multiplex PCR assay. Samples 1 – 10: H. castro, 11 – 20: H. monteiroi. 21 (N): extraction negative, 22 (N): PCR negative, 23: *H. pelagicus* (European storm petrel), 24: *H. leucorhous* (Leach's storm petrel), L: 100bp ladder from Promega. All *H. castro* samples display a band at approximately 200 bp, while all *H. monteiroi* samples display a band at approximately 100 bp. A scale bar has been provided for the ladder, with markers for 100 bp, 500 bp and 1,500 bp. Each line represents an increment of 100 bp.

The use of the screening method on *H. castro* samples from St Helena and Ascension Island was less successful; whilst the multiplexing still worked, the PCR results were not consistent throughout all samples tested. Of four samples from Ascension Island, three amplified. In all three of these cases, the PCR band presented corresponded to that of *castro*-clade, which would be expected from a *H. castro* species. From St Helena only one of three samples amplified. In this single case, again the PCR band presented corresponded to amplification by the *castro*-clade primer set. Thus, whilst the correct amplification pattern was observed, the method appeared less suitable or efficient when used on Ascension and St Helena samples.

To confirm that the PCR bands corresponded to correct amplification of the targeted regions, PCR products from the multiplexed PCRs (two individuals of each species, *H. castro* and *H. monteiroi*) were sent for Sanger sequencing. Due to the short fragment size, sequence quality was poor - samples suspected to amplify with *H. castro* primers returned a sequence length of just 100-120 bp. By mapping the PCR sequence from a suspected *H. castro* sample to a reference sequence (GenBank Accession number DQ178708) it was confirmed that the primers amplified the expected fragment (35-193bp) of the full control region. Samples suspected to amplify with the *H. monteiroi* primers returned a sequence length of <50 bp and would not map to a reference genome. The sequence was compared to a reference sequence (Genbank Accession number DQ178704), looking for conserved regions between the two. Overall, the two sequences matched, and the *H. monteiroi* primers were amplifying correctly.

**Table 2.2** Screening of individuals with *monteiroi*-clade and *castro*-clade in a multiplexed PCR reaction. The number of individuals that were screened, their sampling location and subsequent amplification pattern are displayed. *H. castro* and *H. monteiroi* amplified with their target primers 100% as expected. The samples from Ascension and St Helena (*) amplified less successfully but patterns were as expected in PCRs that did amplify. H. leucorhous and *H. pelagicus* did not amplify with these primers, as expected.

| **Sample and number screened** | **Geographic Location** | **Amplified with *castro*-clade (Y/N)** | **Amplified with *monteiroi*-clade (Y/N)** | **Species / Clade inferred** | **Amplification Expected / Observed** |
|---|---|---|---|---|---|
| ***H. castro* (n=127)** | Azores, Praia Islet | Y | N | *H. castro* | 128 / 128 |
| ***H. castro* (n=3)** | St Helena | Y | N | *H. castro* | 1 / 3* |
| ***H. castro* (n=4)** | Ascension Island | Y | N | *H. castro* | 3 / 4* |
| ***H. monteiroi* (n = 114)** | Azores, Praia Islet | N | Y | *H. monteiroi* | 112 / 114 |
| ***H. pelagicus* (n=1)** | Portugal | N | N | N/A | 1 / 1 |
| ***H. leucorhous* (n=1)** | *N/A* | N | N | N/A | 1 / 1 |

The screening of these primers using a multiplexed reaction demonstrates a robust and quick method for discerning between the two Azores species, and when used in a multiplex reaction, primers only anneal to the clade for which they have been designed, resulting in two separate clade-specific PCR products. 127 *H. castro* and 112 *H. monteiroi* samples were successfully screened using this method, demonstrating a 98-100% success rate in amplifying with their expected primer sets and producing the expected band size.

### 2.4.3: *In* silico testing of clade-specific primers

In addition to *in vitro* testing of the primers on *H. castro* and *H. monteiroi* sampled in the Azores, the publication of an extensive mtDNA dataset by Taylor et al. (2019) allowed for limited *in silico* testing to check primer performance. Using alignments created in section 2.2.4, the 'Test With Saved Primers' feature in Geneious was used to test how primers would bind to samples in different clades, and if this method was applicable to other geographic locations throughout the *H. castro* species complex (North Atlantic, Azores Hot, Cape Verde, Ascension Island, Galapagos, Hawaii and Japan).

The North Atlantic *H. castro* clade contained 305 individuals, including 14 mismatches: four individuals from Cape Verde, five *H. monteiroi* individuals, a single individual from Ascension Island and three individuals sampled in Japan. When testing the *castro* clade-specific primers, it was suggested that all haplotypes would successfully amplify with these primers, aside from two individuals from Madeira that would not amplify with the 35F forward primer. In contrast, none of the samples amplified using the *monteiroi* clade-specific primers, even considering the five *H. monteiroi* individuals. Here the primer sets remained clade-specific, with only the *castro*-clade primers amplifying the individuals of a clade predominantly designated as *H. castro.*

Within the Azores Hot clade there were three *H. jabejabe* individuals sampled from Cape Verde, whilst the rest of the clade is represented by the Azores-endemic *H. monteiroi*. Here all *H. monteiroi* were judged to amplify with the *monteiroi* clade-specific primers, as expected, but results for the three Cape Verde birds were less clear: of the *monteiroi*-specific primers, these only amplified with the *H monteiroi* reverse primer (266R). However, Cape Verde birds also amplified with both forward and reverse *castro*-specific primers. Here it is possible that a double band may be observed for the Cape Verde samples, when viewed on agarose gel.

In the Ascension Island, *H. castro* clade, all but one individual amplified with the forward *H. castro* primer, whilst three failed to amplify with the reverse primer. Only one individual amplified with 193F Mont, whilst all individuals failed to amplify with 266R Mont.

The Cape Verde clade contained one individual from Ascension, however, all were described as *H. castro* and were expected to amplify with *castro*-clade primers. The reverse primer, 193R Castro bound successfully to all *H. castro* individuals. For 35F Castro, only three individuals were suitable for the primer to bind. In contrast, the *monteiroi*-clade primers did not bind to any individuals, and so this clade successfully amplified as *H. castro* only, albeit with only a reverse *H. castro* primer.

Within the Galapagos clade, all were sampled from Galapagos and there is no geographic mixing of individuals evident. As this clade is also described as *H. castro*, the 35F Castro primer bound to all but three individuals and 193R Castro bound to all but one individual. When testing with the *monteiroi*-clade primers, 193F Mont amplified one individual whilst three individuals amplified with 266R Mont.

Likewise, for the Japan clade all *H. castro* present were sampled from Japan. Here, 35F Castro amplified all individuals. For 193R Castro there is the same clade-specific base but there are additional bases in the individuals' sequence that are not present in the primer. No individuals amplified with 193F Mont and only one individual may have potentially amplified with 266R Mont, save for a single base-difference in the 3' end between primer and individual.

The Hawaii clade showed no geographic mixing of *H. castro* however, no samples amplified with the *castro*-specific pair, and primer binding of both forward and reverse primers appeared unlikely. Here, all individuals amplified with the forward *monteiroi*-specific primer (193F), whilst the matching *monteiroi* reverse primer (266R) only amplified one Hawaii sample. Here, the Test With Saved Primers feature demonstrated an unexpected result by amplifying with primers designed for the opposite clade.

## 2.5: Discussion

### 2.5.1: Methodological considerations

#### 2.5.1.1: Application of the developed method in the laboratory

There is a widespread scientific interest in the sympatric, allochronic speciation of the band-rumped storm petrel species-complex (Friesen et al. 2007; Smith et al. 2007; Silva et al. 2016; Taylor et al. 2019). This, along with conservation concerns for the vulnerable Azores endemic *H. monteiroi* (BirdLife International 2016) is spurring ongoing research on the genomic and phenotypic characteristics of Azores storm petrels. As a first step to investigate the potential occurrence of interspecific hybridization between the two species on the archipelago, the multiplex PCR assay developed in this chapter provides a cost- and time-effective alternative to DNA sequencing approaches applied in previous work on Azores storm petrels (Friesen et al. 2007; Smith et al. 2007; Silva et al. 2016; Taylor et al. 2019).

When applying the new multiplex method to screen a series of Azores storm petrel samples (collected 2002-2017), the assay was highly accurate and efficient, yielding correct results for 98% of *H. monteiroi* samples and 100% of the *H. castro* samples, that were sequenced and therefore of known clade. Of those samples not previously sequenced, the assay provided an unambiguous answer for >99% of unknown-clade Azores samples screened. Furthermore, all remaining screening samples yielded clear and unambiguous results, altogether demonstrating the applicability and robustness of the developed assay.

It is currently unclear as to why two *H. monteiroi* samples failed to amplify using the assay (OMA5 and OMA8, see Appendix 2.1), The samples were tested with the assay twice and did not produce a band on agarose gel. When trialled with sexing primers (Fridolfsson and Ellegren. 1999), both samples produced clear bands on agarose gel. This perhaps suggests the samples did not fail due to low DNA quality and may benefit from sequencing using the previously described control region primers (Smith et al. 2007) to investigate if failed amplification is due to primer incompatibility.

As expected from the primer design, PCR tests also confirmed that the primers do not bind to DNA extracted from *H. pelagicus* or *H. leucorhous* (European and

Leach's storm petrel, respectively) which both co-occur with the target species in the North Atlantic. A larger sample size of non-target species and of *H. castro* samples from Madeira and mainland Portugal (the species' closest breeding locations to the Azores) should be investigated to add further support to these findings.

#### 2.5.1.2: *In silico* testing of the developed primers to determine their potential suitability for storm petrels from non-Azores locations

The main mtDNA clades found in *H. castro (s.l.)* (see Taylor et al. (2019) and references therein) show a relatively shallow divergence, yet large intra-clade variability. This results in a challenging situation for design of clade-specific primers such as conducted in this work. The method presented here was therefore specifically developed to target two sympatric storm petrel clades found on the Azores. Nevertheless, the potential suitability of the method was tested *in silico* for published haplotypes from other locations in the *H. castro (s.l.)* range (i.e., Cape Verde, Ascension, Galapagos, Hawaii and Japan).

Of all clades tested, only the Hawaii clade demonstrated a consistent mismatch between expected and observed primer binding. Neither the forward nor reverse primer of the Azores *H. castro*-specific pair bound to the Hawaii haplotypes, as would be expected; instead, the forward primer of the *monteiroi* primers bound to all samples. Hawaii therefore presents a location where using this assay would likely lead to misclassification of mtDNA clades.

For all samples from Cape Verde and Japan, and a small number of samples/haplotypes from Galapagos and Ascension, the *castro*-clade primers showed nucleotide mismatches, potentially preventing efficient binding/amplification. However, this only involved either the forward or the reverse primer, not both at the same time. For the Galapagos and Ascension clades, the *in-silico* tests of the *monteiroi* clade primers revealed potential binding to a small number of samples/haplotypes, but again only for one of the two primers. For templates with only one primer binding, asymmetric PCR (Poddar 2000) might potentially lead to an apparent band on agarose gels, although likely not of a distinct length. Nevertheless, in such cases application of the PCR assay might lead to incorrect conclusions. For example, a sample

collected in Cape Verde, desired to amplify with the *castro*-clade primers, could also show some amplification with the *monteiroi* clade primers.

It is however worth emphasising that *in silico* predictions about primer binding are depending on the set algorithm, and the PCR protocol presented here includes two PCR additives (DMSO and BSA), which based on extensive testing led to clear improvements in primer specificity. The use of these additives could not be simulated in our *in-silico* tests. Therefore, before applying the method developed here to samples from non-Azores locations, a thorough validation using known-clade samples is strongly recommended. With a high success rate in birds sampled in the Azores, it is perhaps best that this method is used primarily on these samples for which it has been specifically designed.

### 2.5.2: Mismatches between Sampling Location and Phylogenetic Clade: insights into storm petrel evolutionary history and conservation

The phylogenetic analysis completed here used all previously published data (collated by Taylor et al. 2019) together with the sequence data newly generated in the present study (contributing a larger number of *H. castro* and *H. monteiroi* samples from the Azores than before). The addition of more samples did not alter the phylogenetic clustering of individuals, with the tree reflecting others produced in previous publications (Friesen et al. 2007; Smith et al. 2007; Taylor et al. 2019), displaying a clear divide between most geographic groups (Figure 2.2; Results). Notably, Cape Verde formed a monophyletic group that separates from all others (supporting the species-status of *H. jabejabe*), and this tree also demonstrates that the Azores hot population (*H. monteiroi*) forms a sister group to other North Atlantic populations. The same relationships between South Atlantic and Pacific populations are also represented here as in Taylor et al. (2019). In previously published data (Friesen et al. 2007; Silva et al. 2016; Taylor et al. 2019), the hot season breeders on Desertas island, Madeira Archipelago, cluster within a single North Atlantic clade, which is similarly demonstrated here, albeit in a clade outside of the main North Atlantic clade, and not within it as suggested by previous data.

Based on the screening of new samples (n=37 *H. castro*; n=57 *H. monteiroi*), the present study did not identify any mismatches between sampling location and expected mitochondrial clade. However, previous publications have reported the presence of such mismatches on the Azores (Friesen et al. 2007, Smith et al. 2007, Silva et al. 2016; Taylor et al. 2019). The phylogenetic analysis performed here demonstrated that out of 846 *H. castro* and *H. monteiroi* sequences, the current level of mismatches between sampling location and geographic clade is low.

For *H. monteiroi* and *H. castro* birds caught in the North Atlantic, a total of 5 'mismatched' individuals have so far been found among the total of 439 birds screened for this geographical region, corresponding to a mismatch frequency of ca. 1.1%. Of the reported mismatches, all were identified from sequences from previously published data, and are from presumed *H. monteiroi* birds (sampled in the Azores during the hot season) that cluster within the North Atlantic *H. castro* clade. *In silico* testing confirmed that all these mismatched samples would likely indeed have been diagnosed as *castro* clade with the newly developed multiplex assay.

For non-Azores locations that were not the focus of this study, the conducted joint analysis of previously published and newly generated mtDNA data revealed a few additional instances of 'mismatches' (Taylor et al. 2019; see also Figure 2.2), which may warrant further investigation. Among these are (1) the previously published finding of a low frequency of North Atlantic lineages (*H. monteiroi* and *H. castro* clades) on Cape Verde, (2) presence of Cape Verde and North Atlantic haplotypes in the South Atlantic (Ascension Island and St. Helena), possibly mirrored by a small proportion of cluster membership of South Atlantic animals in the Cape Verde cluster for ddRAD data by Taylor et al. (2019), and (3) presence of North Atlantic *castro* haplotypes in Japan. These findings are unexpected based on genome-wide RAD-sequencing analyses from Taylor et al. (2019), who found pronounced differentiation between storm petrels from these locations (e.g., placement in PCA, and Bayesian clustering results).

In summary, mtDNA analyses reveal a rare occurrence of mitochondrial mismatches among locations/breeding populations of *H. castro* and *H. monteiroi.*

New, additional sequences from *H.* castro and *H. monteiroi*, produced by this study do not present any of these mismatches, contrasting with findings from other publications (Friesen et al. 2007; Smith et al. 2007; Silva et al. 2016; Taylor et al. 2019). Next, potential biological and statistical explanations for these cases will be discussed.

#### 2.5.2.1: Phenological plasticity, introgression or hybridisation:

Haplotypes that are mismatched with regard to the expected mtDNA clade based on location and breeding time could result from ecological processes. For instance, individual birds showing breeding time plasticity and at least occasionally changing to breed in a different season would lead to such mismatches. Although the occurrence of such season switches has been discussed previously (Silva et al. 2016; Taylor et al. 2019), no clear data to support this have been published.

Introgressive hybridisation has been suggested as a potential explanation for haplotype sharing between *H. castro* and *H. monteiroi* on the Azores. For example, Silva et al. (2016) discussed introgression as an explanation for (i) mtDNA haplotype sharing between the species on the Azores, and (ii) their finding of a *H. monteiroi* individual with a 30% cluster membership of *H. castro* (based on sequencing of anonymous nuclear loci). Some degree of interspecific gene flow might thus still occur, although most data so far suggest relatively strong reproductive isolation (Friesen et al. 2007; Smith et al. 2007; Silva et al. 2016; Taylor et al. 2019). Without the occurrence of a mismatch in the conducted lab testing, it cannot be shown empirically how hybrids would present. With its focus on the non-recombining and maternally inherited mtDNA molecule, this assay will only identify the maternal lineage of the tested bird. Hence, the assay will identify first-generation 'migrants', or their backcrosses on the maternal side. Nuclear assays or genotyping approaches will be required for testing of less recent, admixed backcrosses.

While the two breeding populations on the Azores are seasonally distinct, there is an overlap of the two species' presence on the islets from August until October (Bolton et al. 2008). However, it has been recorded during playback experiments that *H. monteiroi* individuals do not respond to calls of *H. castro* birds, suggesting that interspecific mating may be unlikely. Perhaps the nocturnal behaviour of

storm petrels implies a heavy reliance on vocal rather than visual signals in mate recognition (Bretagnolle 1996), contributing to pre-mating isolation.

#### 2.5.2.2: Incomplete Lineage Sorting (ILS)

Coalescent theory predicts that, after a speciation event, the diverging species will show incomplete lineage sorting (ILS) over extended evolutionary time (on average 4 $N_e$ generations; Nichols 2001), even in the absence of gene flow (Hailer et al. 2013). Recently diverged species such as Azores storm petrels, whose mtDNA lineages have been estimated to have diverged ca. 70,000-150,000 years ago (Bolton et al. 2008), might thus contain shared ancestral haplotypes, simply resulting from ILS. Due to slower drift and lower mutation rates at autosomal loci compared with mtDNA (Zink and Barrowclough 2008) ILS is typically more prevalent at nuclear loci. Indeed, Silva et al. (2016) demonstrated that nuclear DNA markers showed much more extensive haplotype sharing than mitochondrial DNA in Azores storm petrels and their relatives. Previous studies have therefore referred to ILS as a potential mechanism to explain haplotype sharing in Azores storm petrels (Silva et al. 2016; Taylor et al. 2019).

#### 2.5.2.3: Human Error

Precise sampling conditions and methods of species assignation are not known for samples analysed in previous studies. In general, breeding population (and thus species) can be confidently assigned for Azores storm petrels when the bird in question is visibly confirmed to attend a local, active nest, or when the animal is firmly classified based on vocalisations or morphometrics (Bolton et al. 2008). However, for some cases, human error could have occurred, either during sample collection or at later laboratory stages: (1) mist-netting may catch birds that are not actually local breeders, (2) misidentification in the field based on morphometrics/calls, (3) lab/sample handling errors resulting in sample misclassification. These options will be discussed next.

*Mist-netting:* Sampling for previous studies and the present work was associated with a particular Azores breeding population's phenology (i.e., either in the *monteiroi* or *castro* breeding season, respectively). If used, mist-netting is carried out at night due to the nocturnal habits of *H. castro* and *H. monteiroi,* using tape lure recordings to attract the relevant species. Available data suggest that *H.*

*monteiroi* may remain around the Azores also during the non-breeding season (Bolton et al. 2008), which could result in mist-net captures. This could mean both species are liable to out-of-season mist-net captures and, combined with poor lighting in night conditions that is not conducive to morphometric scrutiny, could result in sample/species misclassification, if precautions are not taken (e.g. obtaining a reliable record photograph; see Bolton and Thomas 2001). Similarly, mist-netting can yield catches of individuals that are not part of the local breeding population (Monteiro and Furness 1998; Rowley et al. *in prep*). Finally, species classification based on sampling time alone would be unreliable on the Azores from ca. August until October, since this is known to be a brief period of overlap of the two species on the islets (Bolton et al. 2008). H. *castro* are known to visit breeding chambers/burrows towards the tail end of the *H. monteiroi* breeding season. Hence, any mismatches resulting from samples that were classified based on capture time and mist-netting location should be regarded as inconclusive, unless further phenotypic or genomic characterisation is conducted to firmly determine the species.

*Sample classification in the field:* In the Azores, *H. castro* and *H. monteiroi* are distinguished in the field based on their calls or using morphometrics (Bolton et al. 2008). A difference in size, tail fork and moult stage (post- versus pre-breeding state) is usually used to discern between species during this time. Whilst generally reliable, these require some knowledge and understanding of a bird's breeding and moult strategy. Morphometric classification could be unreliable when based on measurements collected by inexperienced individuals. In the Azores, only one case of seasonal switching, and potential misidentification, has been documented (Bolton et al. 2008). Further, one recorded case of an out-of-season *H. castro* caught in June (Monteiro and Furness 1998a), and two cases whereby a single bird has possibly switched between different breeding seasons (Monteiro and Furness 1998; Bolton et. al 2008) show that species classification based on sampling time could sometimes be misleading. The screening method developed here could be used to routinely screen samples from such cases.

### 2.5.3: Future applications

Whilst taking small blood samples generally does not adversely affect the bird when recommended guidelines are being followed (Owen 2011), this extra step increases handling time (Harvey et al. 2006). Less invasive methods for collecting samples such as the taking of a single breast feather (Segelbacher 2002), or collection of faecal samples would therefore be useful to trial for this newly developed multiplex assay. On the other hand, the quality/quantity of DNA obtained from a blood sample is larger and more reliable than that from a feather sample (McDonald and Griffith 2011).

Analogous screening methods could be developed for other species complexes. For example, throughout the range of the band-rumped petrel species complex, presence of several cryptic species has been suggested (Taylor et al. 2019), e.g. the 'Azores hot season' population (*H. monteiroi*) and more recently also the Cape Verde population (*Hydrobates jajabe;* Sangster et al. 2012). Furthermore, seasonal populations of *H. castro* on the Galapagos show some genetic and phenotypic differentiation (Friesen et al. 2007; Smith et al. 2007). As discussed by Taylor et al. (2019), the possible taxonomic classification of such populations as separate species requires robust genetic and non-genetic data, for which PCR assays such as this one would prove useful.

If mismatches identified using this assay occur frequently enough on the Azores, geolocators could be deployed on "mismatched" individuals, to assess their movement behaviour within and between the breeding seasons of the two candidate taxa, *H. castro* and *H. monteiroi*. Geolocators and GPS tags have already been deployed for these two species in the Azores to investigate isotopic niche segregation (Paiva et al. 2017), and the use of GPS tracking devices is ongoing (Hereward et al. *unpubl*). Knowing more about the two species' year-round movement could shed light on the likelihood of vagrant birds being captured out of season, or if breeding time of the two taxa is changing; for example, breeding time could potentially shift in response to climate change, changing breeding seasons out of allochronic synchronisation. Continuous and regular location data for mismatched individuals may give a more precise indication of breeding season, especially when compared to a single retrap occurrence using current methods.

In addition, this assay could prove useful when sampling young birds prior to their first mating, when typical morphometric identifiers such as moult and brood patch would not be present to aid identification of breeding adults. It is estimated that *H. castro* and *H. monteiroi* do not breed until they are at least 2 years old (Bried and Bolton 2005), and this assay could aid the classification of fledgling and first-year birds that may not be so easy to identify using morphometrics.

The assay developed here is a novel method that has proved successful in prescribing clade at the gel electrophoresis stage of sampling. The potential for its use in future applications is yet to be determined, but the scope and use of such an assay remains promising.

## 2.6: Acknowledgements

Thank you to Renata for organising field work on Praia islet and accompanying me there. I can really see why you love it so much and I am grateful I got to experience it. Thank you to Veronica Neves for your assistance on Praia, your endless knowledge and guidance in navigating the islet and sampling blood. Between you and Renata I couldn't have asked for better company on the islet. Thank you to Joel Bried for all your knowledge on the historical records of these storm petrels, and for sending so many samples for use in this project. Thank you to the entire Graciosa National Park team, for granting of permits to visit and sample on the islet, and for getting us there and back again safely. Thank you to the storm petrel team on Graciosa for being so welcoming and kind, including me in conversations despite my lack of Portuguese. Thank you to the Genetics Society for granting me the Heredity Field Work Grant which funded the trip to sample *H. monteiroi*. Thank you to Sophie Barnett, Nick Price, Jack Ford and Sean Hubbert for help with mtDNA extractions and PCRs. Huge thanks to Lucy Rowley for her help in screening samples using this new method.

# Chapter Three: Characterisation and primer development for two distinct DAB lineages in two sympatric species of *Hydrobates* storm petrels on the Azores

*Renata Medeiros Mirra (right), and I sampling Hydrobates monteiroi on Praia Islet*

**"Sometimes I'll start a sentence and I don't even know where it's going. I just hope I find it along the way." - Michael Scott, The Office US**

# Chapter 3 Characterisation and primer development for two distinct DAB lineages in two sympatric species of *Hydrobates* storm petrels on the Azores

## 3.1: Introduction

The Major Histocompatibility Complex (MHC) forms part of the adaptive immune system in vertebrates, coding for highly specific cell-surface proteins that elicit an immune response against pathogens (Janeaway et al. 2001a). The avian MHC is polygenic, presenting multiple loci that derive from ancestral duplications (Burri et al. 2008), with typically numerous alleles present at these loci that are susceptible targets for balancing or diversifying selection (Piertney and Oliver 2006). This results in individuals with highly specific and varied MHC profiles, capable of fighting off multiple pathogens (Janeaway et al. 2001a). The genetic diversity present at MHC loci has been characterised for a wide range of animal species, demonstrating that this diversity is maintained and conserved through a broad range of taxonomic groups (Reche and Reinherz 2003; Richardson and Westerdahl 2003; Wegner et al. 2006; Minias et al. 2018), with increased diversity providing the ability to elucidate an immune response to a wider range of pathogens. Maintaining this MHC diversity demonstrates a fitness advantage for individuals, making the MHC gene region a candidate for selection to maintain this advantage (Piertney and Oliver 2006). In many model systems, this selection has been shown to influence mate choice, with individuals choosing mates based on MHC genotype (Jordan and Bruford 1998), and the MHC has since become the target in multiple immunity and mate-choice studies (Eizaguirre et al. 2009; Juola and Dearborn 2012; Dearborn et al. 2016; Jeannerat et al. 2018).

MHC genes consist of two different 'classes' of molecules, MHC Class I and MHC Class II. MHC Class I molecules are present on most cells, largely associated with intracellular defence against viruses (Janeaway et al. 2001b). In contrast, MHC

class II molecules are extracellular, present on specialised immune cells that are responsible for fighting off parasites and pathogens (Piertney and Oliver 2006). Many published studies of mate choice have targeted the MHC Class IIB gene region, as its importance in extra-cellular immune response makes it a target for natural or sexual selection (Burri et al. 2010; Dearborn et al. 2016a). More specifically, exon 2 of this gene has been a focus of much research, since part of this region codes for the antigen binding site (Brown et al. 1993; Sommer 2005a) – a functionally important region involved in fighting pathogens and prone to strong selection and higher substitution rates (Ohta 1998). Diversity of class IIB sequences is frequently elevated due to balancing selection (Hedrick 1999; Sommer 2005b). This balancing selection can be influenced by (1) natural selection (e.g. spatio-temporally varying selection as a response to pathogens and parasites) (Biedrzycka et al. 2018), which itself can drive the process of (2) sexual selection and mate choice (Bernatchez and Landry 2003a), where individuals can select mates based on MHC profile, targeting two different mechanisms – the 'good genes' hypothesis, or heterozygote advantage. The 'good genes' hypothesis implies that specific haplotypes are selected for by individuals, based on the haplotype's ability to fight off specific pathogens or parasites that regularly present in a population (Eizaguirre et al. 2009b). In contrast, originally proposed by Doherty and Zinkernagel (1975), heterozygote advantage implies that individuals with heterozygous MHC genes are better equipped to respond to and fight off a wider array of pathogens; individuals therefore choose mates that will result in heterozygous offspring (Penn et al. 2002). Maintaining heterozygosity in offspring can be achieved through individuals choosing mates that are dissimilar to themselves (disassortative mating), a trait that has shown to be detectable by scent (Boehm and Zufall 2006) and demonstrated in many animal systems (Kamiya et al. 2014). To investigate MHC-based signals of natural and sexual selection in a species, a method to specifically detect and sequence the highly polymorphic MHC loci is required.

Previous research on MHC in different seabird systems has demonstrated both presence and absence of disassortative mating (Forsberg et al. 2007; Juola and Dearborn 2012; Strandh et al. 2012; Dearborn et al. 2016; Hoover et al. 2018). However, these studies can be complicated by the presence of an ancient gene duplication of the MHC that pre-dates the radiation of all extant birds, resulting

in two MHC lineages, DAB1 and DAB2 (Burri et al. 2008a; Goebel et al. 2017). This was found after phylogenetic analysis revealed that MHC sequences do not cluster by species, but by locus (Nei and Rooney 2005). This is especially evident from phylogenies of exon 3 sequences, which show clustering by locus (i.e., DAB1 and DAB2 on separate branches, and avian phylogeny recapitulated separately within each DAB lineage), reflecting a retained signal of the ancestral DAB duplication. In contrast, exon 2 sequences that code for the peptide-binding region do not cluster by DAB lineage but frequently by species, likely a result of balancing and directional selection as well as gene conversion (Edwards et al 1999, Burri et al 2008). It is thought that the MHC family is evolving through a 'birth and death model', where gene copies are created through duplication events, counteracted by loss from some genomes through deletion events. Furthermore, gene conversion has assimilated the two lineages (Goebel et al. 2017). Hence, many bird species contain two MHC class IIB lineages in their genomes, allowing considerable allelic diversity to be passed on to offspring regardless of allelic match to the mate – by-passing the need for disassortative mating (Dearborn et al. 2015; but see also Hoover et al. 2018). For instance, presence of both DAB lineages have been documented for genomes of owls *Strigiformes* (Burri et al. 2008), Leach's storm petrel *Hydrobates leucorhous* (Dearborn et al. 2014), herons *Ardea* (Wang et al. 2013), and egrets *Egretta* (Wang et al. 2013).

Understanding the mechanisms and consequences of mate choice is a central aspect of understanding the drivers of speciation in species-complexes such as the band-rumped storm-petrels *Hydrobates castro* (s.l.). Currently, little work has been carried out on MHC structure and genetic diversity in the sympatric but allochronically-breeding "sibling-species" such as *H. castro* and *H. monteiroi* on the Azores. PCR, cloning and Sanger sequencing of MHC Class IIB loci (spanning a region between the end of exon 1 and the beginning of exon 3) in single individuals from a wide variety of bird species including *H. castro* and *H. monteiroi* revealed each species to contain at least 5 MHC DAB alleles, indicative of >3 loci (Burri et al. 2014). The study by Burri et al. used primers applicable to a wide range of avian taxa, and due to small sample size (n=1 individual per species) could only provide preliminary estimates of MHC diversity in terms of alleles and loci. In addition, the utilised primers might only have amplified a

subset of the actual allelic variation present in the individual studied by Burri et al. (2014). Finally, the study did not address how the discovered alleles were related to two main DAB lineages (Goebel et al. 2017) observed in other seabird species (Burri et al. 2008a; Dearborn et al. 2015a).

In summary, the presence or absence of DAB1 and DAB2 in *H. castro* and *H. monteiroi* remains unknown, and DAB diversity remains to be thoroughly characterised in these species.

## 3.2: Aims:

This chapter aims to expand on the current characterisation of *H. castro* and *H. monteiroi* MHC Class IIB region, and to develop methods for large-scale population screening of this locus. Specifically, the aims are to: (1) determine whether both main DAB lineages that derive from an early avian genomic duplication (Goebel et al. 2017) exist in these species, or whether Azores storm petrels might have lost either of these lineages; (2) conduct a preliminary assessment based on Sanger sequencing of the MHC diversity of *H. monteiroi* and *H. castro* breeding under sympatric conditions on the Azores. (3) Based on this assessment of allelic diversity, develop novel primers and establish a PCR protocol to amplify any encountered DAB loci (exon 2 alleles) in both species, with amplicon size being suitable for high-throughput Illumina sequencing approaches (see Chapter 4). Ideally, these primers should amplify both species at the same time.

## 3.3: Methods

Samples from *H. castro* and *H. monteiroi* were collected and extracted as detailed in Chapter 2.2.1: Sample Collection and Chapter 2.2.2: DNA Extraction. In short, 56 samples were collected from breeding *H. monteiroi* adults in 2017, which were added to a larger sample set for both species provided by colleagues in the Azores (Nava et al. 2017). DNA from these samples was then extracted using an ethanol extraction method provided by the NERC NBAF facility in Sheffield, based on a protocol by Bruford et al. (1998). Details of the individuals used can be found in A4.6.

Throughout this chapter, all primer combinations that were designed were tested extensively in PCR. Numerous optimisation attempts were made for all

primer pairs, including adjusting PCR reagent concentrations, adjusting PCR times, temperatures and cycles, testing PCR additives (i.e., Dimethyl sulfoxide (DMSO), Qiagen Q-solution and Bovine Serum Albumin (BSA)), trialling different Taq polymerases (Qiagen Multiplex, HotStar Taq), altering template DNA amount, and dilution or purification of PCR products (ExoSAP-IT™, ThermoFisher Scientific, or the QIAquick PCR purification kit, Qiagen). After successful optimisation, PCRs were carried out in an Applied Biosystems SimpliAmp thermal cycler (ThermoFisher Scientific), using GoTaq® G2 Flexi DNA Polymerase. Most PCR products suitable for Sanger sequencing (sent to Eurofins, Germany) were either diluted ca. 1:15 with water, or purified using EXOsap_IT (ThermoFisher). Sequences were visualised, edited and aligned in Geneious (version 9.0; https://www.geneious.com).

### 3.3.1: Testing of Published Storm Petrel MHC Primers

Initial attempts to amplify the MHC Class IIB region in *H. castro* and *H. monteiroi* utilised previously published primers from Dearborn et al. (2015) and Burri et al. (2014).

First, the nested PCR protocol used by Dearborn et al. (2015) was trialled, which targeted DAB1 and DAB2 of the MHC Class IIB gene region in Leach's storm petrels. Primers were ordered from Sigma Aldrich and initially tested following the PCR programme used by Dearborn et al. (2015). Later, various optimisation steps were taken (altering concentrations of reagents and changing PCR protocol timings and temperatures) aiming to achieve specific and strong PCR amplification for *H. castro* and *H. monteiroi*. PCR products for both the first 'outer' reaction and the secondary 'inner' reaction were run on 1.5% agarose gels. Amplification using the outer DAB1 and DAB2 primers (Locus 1 and 2, Outer F and R) was successful, presenting bands of approximately 850 bp (DAB1 primers) and 750 bp (DAB2 primers). Whist the outer PCR reactions for both DAB1 and DAB2 were successful, the inner PCR reaction remained unsuccessful, presenting multiple non-specific bands, despite attempts to optimise the PCR. Additional trials using different polymerase kits, such as HotStar Taq and Multiplex (Qiagen, Germany), were also unsuccessful in optimising the inner PCR reaction of the nested PCR.

Primers from Burri et al. (2014) (AvesEx1-F1, AvesEx1-F2 and AvesEx1-R1) also failed, despite numerous optimisation attempts. Following the PCR protocol used in Burri et al. (2014), PCRs resulted in non-specific bands when viewed on an agarose gel. For both primer sets (Burri et al. 2014; Dearborn et al. 2015) optimisation steps included altering PCR temperatures, reagent concentrations and PCR stage times.

In summary, none of the previously published primers produced a clear, single band for *H. castro* and *H. monteiroi*.

### 3.3.2: Exploring Genetic Variability of DAB Exon 2 and its Flanking Regions

To isolate large fragments of the MHC Class IIB for both DAB1 and DAB2 in *H. castro* and *H. monteiroi* and to explore the genetic variability of exon 2 and its surrounding flanking regions, long fragments of exon 2 along with its surrounding flanking regions were characterised in a range of different individuals for both species. This information was used to detect conserved regions, suitable for the design of exon specific primers, ideally suitable for both *H. castro* and *H. monteiroi.*

For investigation of the DAB1 lineage in *H. castro* and *H. monteiroi*, new primers were developed. To this end, sequence data from Burri et al. (2014) was downloaded from GenBank (accession numbers KJ162507 – KJ162517), yielding a 1,289 bp alignment of 11 sequences (6 samples from *H. castro,* and 5 from *H. monteiroi*). Sanger sequencing is best suited to amplicons of <1,000 bp. Thus, two sets of primers were designed yielding fragments that overlapped in the middle of this region (Fig. 3.1). For this, the 'Design New Primers' feature in Geneious was used, which is based on the PRIMER3 algorithm (version 2.3.4; Untergasser et al. 2012). The first overlapping fragment aimed to cover approximately the first 630bp, and the second aimed to cover 580bp to 1270bp. Primer size was constrained to 18-24 bp, with an optimal size of 20 bp. Melting temperature ($T_m$) was set to 55-67°C, with an optimal value of 60°C. The MHC region is typically GC-rich, and so %GC settings were generous, ranging between 40-80%, and the optimal amount set to 50%. Typically, %GC should range between 40-60% to avoid primer dimer formation and a too high $T_m$. Here, a larger %GC would avoid restricting our options for primers in a GC-rich region.

Max Tm difference allowed within the primer pair was set to 5°C. A 1 bp GC-clamp was included, to improve specificity of the primer binding.

Each primer pair was manually inspected – rejecting primers that were located far into the fragment or did not yield enough overlap. Primers in acceptable binding regions were further inspected, avoiding high $T_m$ for any hairpin, self-dimer or pair-dimer structures. A final set of 11 primers met the following criteria:

- A PCR product that fell within 1-650 bp for fragment further 5', and 580-1270 bp for the fragment further downstream (Fig. 3.1).
- A primer length of approximately 18-20 bp.
- A low, or zero, chance of self-, primer- or pair-dimer formation.
- Low or zero chance of hairpin formation.
- Primer %GC as low as possible. GC content varied, with primers ranging from 50-67%. This fell outside of the generally advised percentage of 40-60%, but still acceptable in this study, especially considering the GC-rich MHC region.
- A between-primer $T_m$ difference of <5°C. Primers were preferentially chosen where this difference was smallest.

These 11 primers were ordered from Sigma-Aldrich and the 7 combinations they provided were subjected to lab testing and optimisation. (Table 3.1, Figure 3.2).

**Table 3.1** Primers and combinations designed and used to amplify the DAB1 lineage of the MHC Class IIB gene region, in two overlapping segments. Combinations in bold indicate the final, successful primers that were used

| Primer Name and Combination | Primer Sequence (5'-3') |
|---|---|
| 9F<br>627R | CACCTGGCTCGTGGGGAG<br>CTTGTGCCCGTCCCTCAGG |
| 18F<br>627R | YGTGGGGAGGAGACCTCAG<br>CTTGTGCCCGTCCCTCAGG |
| 16F<br>619R | CTCGTGGGGAGGAGACCT<br>CGTCCCTCAGGGCAATGTT |
| **25F**<br>**624R** | **AGGAGACCTCAGGTGAGCTC**<br>**CTTGTGCCCGTCCCTCAG** |
| **25F**<br>**616R** | **AGGAGACCTCAGGTGAGCTC**<br>**CGTCCCTCAGGGCAATGTTC** |
| **589F**<br>**1265R** | **AGTGCTTGGCAGAACATTGC**<br>**ACCCCCGTGAGCATCTTG** |
| **597F**<br>**1265R** | **GCAGAACATTGCCCTGAGG**<br>**ACCCCCGTGAGCATCTTG** |

Primers were initially tested using the standard PCR mix and cycle profile. On agarose gels, all primer pairs initially showed at least one additional, unexpected band. Where additional bands were equally as bright as the expected amplicon length, it was anticipated that it would be impossible to optimise these primers, and these were rejected. The remaining primers (25F/624R, 25F/616R, 589F/1265R and 597F/1265R) represented two potential primer combinations for each half of the MHC Class IIB, DAB1 fragment, and were later optimised, using the previously mentioned optimisation steps, to produce a single band on agarose gels. Between rounds of optimisation PCR products were sent for Sanger sequencing, and quality checked using FinchTV (Geospiza, Inc., USA) and Geneious to check that PCRs were producing clean, strong and readable data, aiding decisions regarding which optimisation steps had been successful.

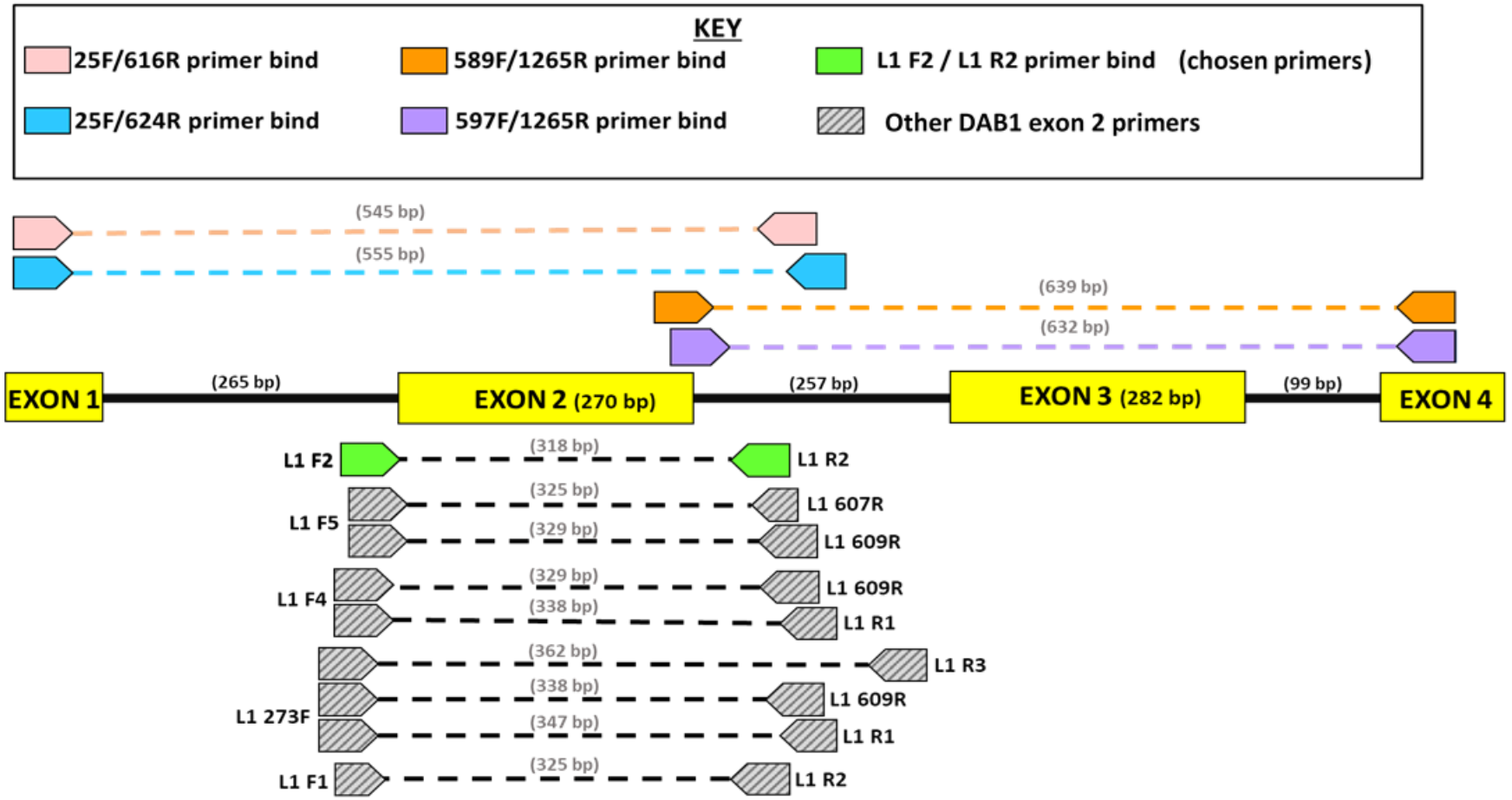



**Figure 3.1** Primer binding sites for DAB1. Grey: primers used for exploratory work; green: primers ultimately deemed suitable for future high-throughput sequencing. Intron and exon lengths are based on GenBank sequences from Burri et al. (2014), and amplicon lengths are taken from Sanger sequencing results

The addition of DMSO to the PCR mix was also tested. DMSO is often used in amplifying GC-rich regions, making the DNA conformation more sensitive to temperature increase, and thus reducing secondary structure formation (Hardjasa et al. 2010).

After lab optimisation, the PCR reactions and protocols for each primer pair varied, according to the optimum conditions achieved in lab testing. Every PCR contained the same two stages – beginning with a 2-minute denaturation step at 95°C, and ending with a 5-minute extension step at 72°C. Only the cycling stage changed between primer pairs. The following pair-specific conditions for the primers applied:

i. **25F/616R** – The PCR reaction consisted of 1X GoTaq Green Flexi buffer, 1.5 mM MgCDAB2, 0.16 mM of each dNTP, 0.2 µM of each primer, 5% DMSO, 0.3 U/µL Taq polymerase and 1 µL genomic DNA, or water for PCR negatives. This was raised to a 15 µL reaction volume with 7.65 µL of ultrapure water (Affymetrix, ThermoFisher Scientific). The cycling stage of this PCR lasted for 32 cycles, beginning with denaturation at 95°C for 30 seconds, annealing at 64°C for 30 seconds and extension at 72°C for 30 seconds.

ii. **25F/624R** – This PCR reaction included 1X GoTaq Green Flexi buffer, 2.5 mM MgCDAB2, 0.16 mM of each dNTP, 0.3 µM of each primer, 0.025 U/µL Taq polymerase, and 1 µL genomic DNA or water. 7.425 µL of ultrapure water was added, to increase the reaction volume to 15 µL. In this PCR, there were 34 cycles of 35 seconds at 95°C, 40 seconds at 68°C and 1 minute 10 seconds at 72°C.

iii. **589F/1265R** – This PCR mix contained 1X GoTaq Green Flexi buffer, 2.5 mM MgCDAB2, 0.16 mM of each dNTP, 0.3 µM of each primer, 5% DMSO and 0.03 U/µL Taq polymerase. Each tube was topped up to 15 µL with 6.65 µL ultrapure water and contained 1 µL of genomic DNA

or water. The cycling conditions here consisted of 35 cycles of a 30-second denaturation step at 95°C, annealing for 45 seconds at 63°C and an extension of 45 seconds at 72°C.

iv. **597F/1265R** – Here, the mix consisted of 1X GoTaq Green Flexi buffer, 1.5 mM MgCDAB2, 0.16 mM of each dNTP, 0.2 µM of each primer, 5% DMSO and 0.03 U/µL of Taq polymerase. Again, 1 µL of genomic DNA or water was added and the PCR mix topped up to a 15 µL reaction volume, using 7.65 µL ultrapure water. The middle stage of this PCR ran for 35 cycles – 30 seconds at 95°C, 45 seconds at 67°C and 1 minute at 72°C.

Successful PCR products showed a clear single band at the expected lengths (Figure 3.2) and were sent for Sanger sequencing in both directions (forward and reverse). These sequencing results provided a characterisation of exon 2 flanking region variability of DAB1 and DAB2, allowing subsequent primer design to target exon 2 of the MHC Class IIB.

To investigate variability in and around exon 2 in the DAB2 lineage, 'outer reaction' primers from Dearborn et al. (2015) were used to amplify four samples (two *H. castro,* two *H. monteiroi*). The PCR mix consisted of 1X Green GoTaq Flexi buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 µM of each primer, 0.03 U/µL Taq polymerase and 1 µL DNA extract. The total reaction volume of 15 µL, following addition of 8.2 µL of ultrapure water. PCR reactions consisted of denaturation for 2 minutes at 95°C, followed by 25 cycles of denaturation for 30 seconds at 95°C, annealing for 45 seconds at 60°C, extension for 1 minute 30 seconds at 72°C, followed by a final extension of 72°C for 5 minutes. Use of Locus 2 Outer F and Outer R was successful, and PCR products were sequenced. Sequences were aligned in Geneious, along with DAB2 sequences extracted from Dearborn et al. (2015; supplementary material).

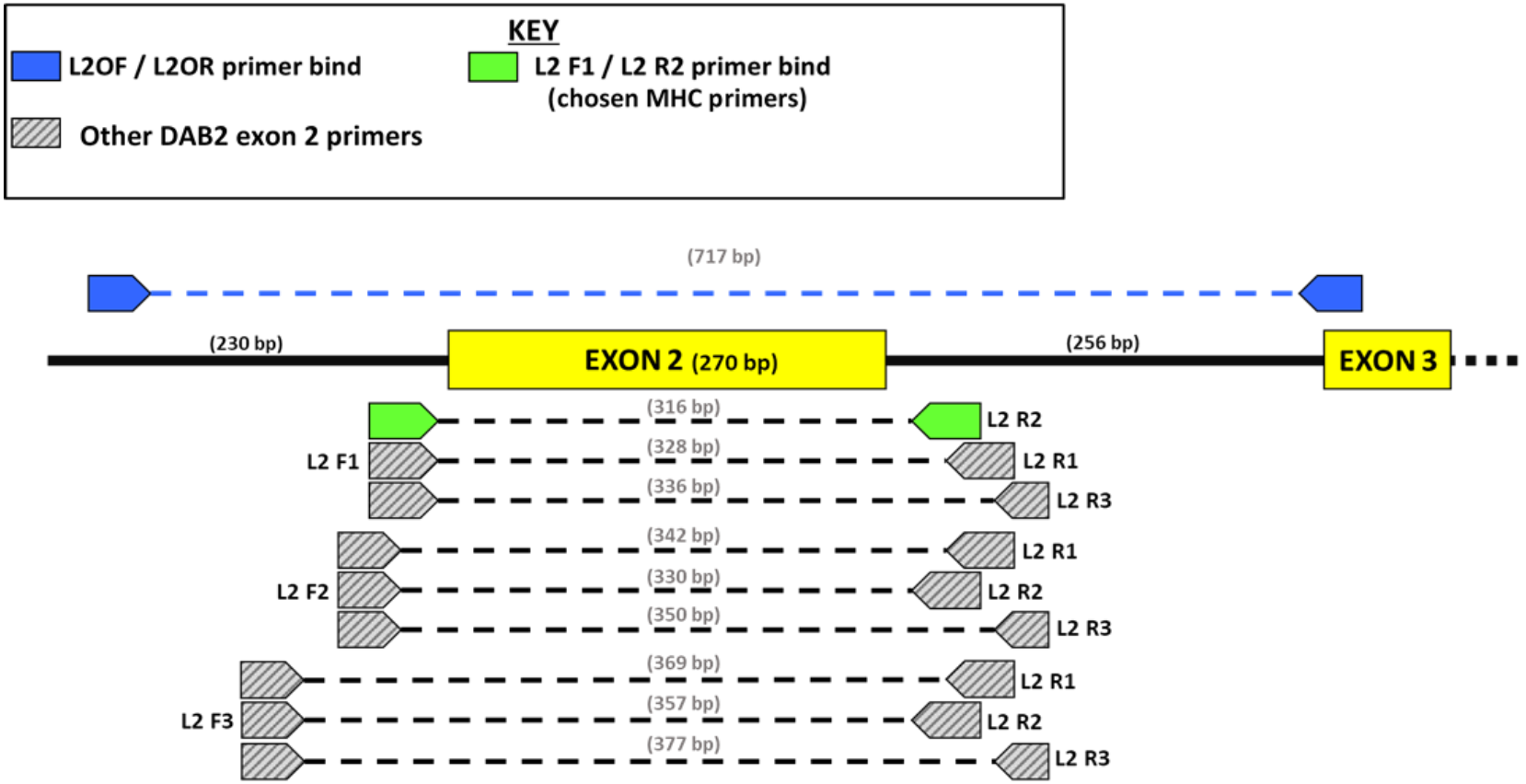



**Figure 3.2** Primer binding locations for DAB2 primers. Primers from Dearborn et al. (2015) are displayed above the MHC fragment (blue); primers designed as part of this study are displayed below (grey: primers used for exploratory work; green: primers ultimately deemed suitable for future high-throughput sequencing). Expected amplicon lengths for each primer set are shown. Intron and exon lengths are based on sequences from Dearborn et al. (2015)

### 3.3.3: Design of lineage-specific primers targeting exon 2

For both DAB1 and DAB2, Sanger sequences obtained as described in section 3.2.1 were first quality checked by viewing the raw data in FinchTV. Separate alignments for DAB1 and DAB2 were created in Geneious, aligned to reference sequences from Burri et al. (2014) (Genbank accession number KJ162507) and Dearborn et al. (2015) (taken from supplementary information), respectively. Primer binding regions for both DAB1 and DAB2 were mapped and trimmed away, and sequences of low quality were discarded. Using the chromatograms for each sequence, and comparing between forward and reverse sequences, base calls were manually checked and corrected if necessary. In cases where a peak was unclear, an ambiguous IUPAC code was inserted. Each DAB lineage was then used to design primers targeting exon 2. Using previously published data (Dearborn et al. 2014), the location of MHC introns and exons in each DAB lineage were mapped. Exon-targeting primers were designed by eye for both DAB1 and DAB2, to ensure the binding sites were close to exon 2. Heterozygous sites close to the exon 2 boundaries were targeted for designing primers in both DAB lineages, as these were most likely to represent the detection of >1 allele.

For DAB1 a total of 6 forward primers and 10 reverse primers were initially designed and subsequently tested *in silico* using the 'Test with Saved Primers' feature in Geneious. This demonstrated which primer combinations would most likely be compatible, based on characteristics such as pair-dimer and pair $T_m$ difference. For these primers, the same considerations were taken as before (Section 3.2.2); primer length was targeted at 19-23 bp, GC content <70%, $T_m$ <70°, and the chance of self-dimer, primer-dimer and hairpin formation was minimised. After *in silico* testing, 5 forward primers and 7 reverse primers were ordered and tested in 13 different combinations (Table 3.2).

**Table 3.2** Primers targeting exon 2 of the DAB1 lineage. Amplicon size was taken from samples sent for Sanger sequencing Primers in bold were chosen to create MID-tag primers

| Primer names | Primer Sequences ('5-3') | Amplicon Size |
|---|---|---|
| DAB1 273F<br>DAB1 609R | GCACAGCCCTGACCTCTCCATG<br>GTCCCTCAGGGCAATGTTCTGC | 338 bp |
| DAB1 273F<br>DAB1 R1 | GCACAGCCCTGACCTCTCCATG<br>CTTGTGCCYGTCCCTCAGGGC | 347 bp |
| DAB1 273F<br>DAB1 R3 | GCACAGCCCTGACCTCTCCATG<br>ARMCCGGGGGCTTGGCTTGT | 362 bp |
| DAB1 F4<br>DAB1 609R | TGACCTCTCCATGTCTGCACGA<br>GTCCCTCAGGGCAATGTTCTGC | 329 bp |
| DAB1 F4<br>DAB1 R1 | TGACCTCTCCATGTCTGCACGA<br>CTTGTGCCYGTCCCTCAGGGC | 338 bp |
| DAB1 F5<br>DAB1 607R | ACCTCTCCATGTCTGCACGAG<br>CCCTCAGGGCAATGTTCTG | 325 bp |
| DAB1 F5<br>DAB1 609R | ACCTCTCCATGTCTGCACGAG<br>GTCCCTCAGGGCAATGTTCTGC | 329 bp |
| DAB1 F1<br>DAB1 R1 | CAGCCCTGACCTCTCCATGTYT<br>CTCTCCCCTCCTGCGARMC | -<br>(Failed PCR) |
| DAB1 F3<br>DAB1 R3 | CAGCCCTGACCTCTCCATGTYT<br>CGRGGGCTTGGCTTGKGC | -<br>(Failed PCR) |
| DAB1 F1<br>DAB1 R3 | CAGCCCTGACCTCTCCATGTYT<br>CGRGGGCTTGGCTTGKGC | -<br>(Failed PCR) |
| DAB1 F2<br>DAB1 R1 | GACCTCTCCATGTYTGCACGAR<br>CTCTCCCCTCCTGCGARMC | -<br>(Failed PCR) |
| **DAB1 F2<br>DAB1/DAB2 R2** | **GACCTCTCCATGTYTGCACGAR<br>GCAATGTTCTGCCMAGCACT** | **318 bp** |
| **DAB1 F1<br>DAB1/DAB2 R2** | CAGCCCTGACCTCTCCATGTYT<br>GCAATGTTCTGCCMAGCACT | 325 bp |

Each primer combination was first tested using a temperature gradient PCR, to find a suitable annealing temperature, using the same PCR mix as above. A total of 9 primer pairs could be optimised, whilst 4 pairs were rejected after failing to amplify clearly.

For DAB2, a total of 3 forward and 3 reverse primers were designed to target exon 2, tested in 9 different combinations (Table 3.3). When designing the reverse primers, a conserved region between DAB1 and DAB2 was found in the intron after exon 2 (see also Dearborn et al. 2014). This meant the R2 primer (5'-GCAATGTTCTGCCMAGCACT-'3), originally designed for DAB1, could also be tested for DAB2, with lineage specificity provided by the forward primers.

**Table 3.3** Primer pairs and combinations designed for amplifying DAB2. Amplicon lengths are shown. The most suitable primer pair (DAB2F1/R2 NEW) is indicated in **bold**.

| Primer names | Primer Sequences (5'-3') | Amplicon length |
|---|---|---|
| DAB2F1<br>DAB2R1 | GACCTGCCTCCCTGCACAAACA<br>CCGTCCCTCAGGGCAATGTTC | 328 bp |
| **DAB2F1<br>DAB1/DAB2 R2** | **GACCTGCCTCCCTGCACAAACA<br>GCAATGTTCTGCCMAGCACT** | **316 bp** |
| DAB2F1<br>DAB2R3 | GACCTGCCTCCCTGCACAAACA<br>GCTTGTGCCCGTCCCTCAG | 336 bp |
| DAB2F2<br>DAB2R1 | GAGCTGGCCACCGTGACCTG<br>CCGTCCCTCAGGGCAATGTTC | 342 bp |
| DAB2F2<br>DAB1/DAB2 R2 | GAGCTGGCCACCGTGACCTG<br>GCAATGTTCTGCCMAGCACT | 330 bp |
| DAB2F2<br>DAB2R3 | GAGCTGGCCACCGTGACCTG<br>GCTTGTGCCCGTCCCTCAG | 350 bp |
| DAB2F3<br>DAB2R1 | GCTGAGAGCACCTTTTGGS<br>CCGTCCCTCAGGGCAATGTTC | 369 bp |
| DAB2F3<br>DAB1/DAB2 R2 | GCTGAGAGCACCTTTTGGS<br>GCAATGTTCTGCCMAGCACT | 357 bp |
| DAB2F3<br>DAB2R3 | GCTGAGAGCACCTTTTGGS<br>GCTTGTGCCCGTCCCTCAG | 377 bp |

Primers were optimised following the methods described above. Three primer pairs failed and did not provide any visible results, whilst the remaining 6 produced clear bands on agarose gels.

For both DAB lineages, the aim was to identify allelic variation within and between individuals, and to check for concordance of individual results among different utilised primers (following recommendations by Burri et al. 2014). High quality signal was required to ensure that heterozygous sites were true and not the result of PCR artifacts. For both DAB1 and DAB2, resulting alignments included four individuals (two *H. castro* and two *H. monteiroi*). To compare DAB1, sequences using the previously utilised flanking-region primers (25F/624R, 25F/616R, 589F/1265R and 597F/1265R), and the exon-specific primers (Table 3.2) were compared in an alignment. For DAB1, a total of 60 different sequences were compared across all tested primers. For DAB2, the corresponding alignment consisted of 32 sequences using the primers from Dearborn et al. (2015), and sequences from exon-specific primer testing (Table 3.3).

The two DAB primer pairs (DAB1F2 + DAB1/DAB2 R2 and DAB2F2 + DAB1/DAB2 R2) chosen as candidates for further high-throughput sequencing (**Chapter 4**) were then used to PCR amplify a set of 10 samples, to evaluate consistency in primer performance.

### 3.3.4: Confirming Amplification of DAB1 and DAB2 Using Phylogenetics

To verify whether both the DAB1 and DAB2 lineages had been amplified in storm petrels from the Azores, a phylogenetic tree of sequences was created.

Sanger sequencing data of DAB exon 2 in storm petrels from the Azores was combined with homologous data obtained from Genbank, yielding an alignment of 76 sequences, including sequences from owls, herons, storm petrels, chicken *Gallus gallus* (Table 3.4), and using spectacled caiman, *Caiman crocodilus.* as an outgroup (which diverged prior to the radiation of modern birds; Jarvis et al. 2014). Sequences were trimmed to Exon II of the MHC, and the alignment was exported as a phylip file. Using IQTree (Nguyen et al. 2015b) the most suitable substitution model was identified using the 'ModelFinder' feature (Kalyaanamoorthy et al. 2017). A total of 280 different models were tested, and assessed according to BIC, AIC and AICc values. For all three criteria, a Kimura-2-

Parameter model with four gamma categories (K2P+G4) was recovered as the best-fit model and was chosen for further analyses. IQTree was run with 200 bootstrap replicates to assess clustering reliability, and a consensus tree was created and edited in FigTree and Inkscape. Analysis scripts for IQTree can be found in A3.1.

**Table 3.4** Details of all samples included in the phylogenetic analysis shown in Fig. 3.6. Genbank accession codes and DAB lineage are specified where known.

| Common / Species Name | Accession / Sample Name | Locus | Reference |
|---|---|---|---|
| Spectacled caiman / *Caiman crocodilus* | AF277661.1 | Unspecified | (Miller et al. 2005) |
| Red junglefowl / *Gallus gallus* | AM489767.1 | Unspecified | (Worley et al. 2008) |
| | AM489776.1 | | |
| Barn owl / *Tyto alba* | EU442606.2 | DAB1 | (Burri et al. 2008b) |
| | EU442607.1 | DAB2 | |
| Spotted eagle owl / *Bubo africanus* | EF641246.1 | DAB1 | (Burri et al. 2008a) |
| | EF641244.1 | DAB2 | |
| Tawny owl / *Strix aluco* | EF641256.1 | DAB1 | |
| | EF641254.1 | DAB2 | |
| Tawny owl / *Strix aluco* | KJ162541.1 | DAB1 | (Burri et al. 2014a) |
| | KJ162544.1 | DAB2 | |
| Burrowing owl / *Athene cunicularia* | XM_026850938.1 | Unspecified | (Mueller et al. 2018) |
| | XM_026867015.1 | | |
| Little egret / *Egretta garzetta* | XM_009647145 | Unspecified | (Zhang et al. 2014) |
| Crested ibis / *Nipponia nipon* | NW_008999372.1 | Unspecified | |
| Chinese egret / *Egretta eulophotes* | KC282848.1 | DAB1 | (Wang et al. 2013) |
| | KC282849.1 | DAB2 | |
| Adelie penguin / *Pygoscelis adeliae* | KDAB225323.1 | Unspecified | (Zhang et al. 2014) |
| Rock dove / *Columba livia* | AKCR00000000.2 | Unspecified | (Holt et al. 2018) |
| Ivory gull / *Pagophila eburnea* | KJ162440.1 | Unspecified | (Burri et al. 2014a) |
| | KJ162441.1 | | |
| Bulwer's storm petrel / *Bulweria bulwerii* | KJ162487.1 | Unspecified | |
| | KJ162488.1 | | |
| Monteiro's storm petrel / *Hydrobates monteiroi* | KJ162513 - KJ162516 | Unspecified | |
| Band-rumped storm petrel / *Hydrobates castro* | KJ162507 - KJ162509 | Unspecified | |
| Leach's storm petrel / *Hydrobates leucorhous* | KP090142 - KP090150 | DAB1 | (Dearborn et al. 2014) |
| | KP090151 - KP090162 | DAB2 | |
| Monteiro's storm petrel / *Hydrobates monteiroi* | OC311 (x3) | DAB1 | N/A (Sanger sequences from primer testing) |
| | OC311 (x4), OC350 | DAB2 | |
| Band-rumped storm petrel / *Hydrobates castro* | OC029 (x3), OC036 (x5), OC015, OC149 | DAB1 | |
| | OC036 (x5), OC029, OC015 | DAB2 | |

### 3.3.5: Characterisation of DNA polymorphism of DAB1 and DAB2 in Azores storm petrels

A subset of samples (six samples, three per species) of the individuals analysed previously in this study were analysed for all primer combinations. Using the same individuals for each primer pair enabled detection of any differences in the recovered allelic variation among primer pairs.

Once sequences were trimmed and quality-checked, alignments for each DAB lineage were created, and mapped to a reference sequence (KJ162507 for DAB1 and a sequence extracted from supplementary information of Dearborn et al. (2014) for DAB2). The consensus sequence from the aligned sequences was used to determine total number of polymorphic sites present across all sequences, whilst each individual was manually investigated to determine the range of polymorphic sites between individuals. To determine the average number of polymorphic sites per individual, the total number of sites present in all individuals was divided by the number of individuals.

## 3.4: Results

### 3.4.1: Exploring genetic variability of exon 2 and its flanking regions

For DAB1, previously published primers did not yield clear and reliable PCR/sequencing results, so new primers were designed. Among the 11 designed primers (6 forward, 5 reverse; tested in 7 combinations; Fig. 3.2) that covered the majority of the MHC Class IIB DAB1 lineage in two overlapping fragments, four primer pairs were successfully optimised for use in PCR (Table 3.5). Trialling PCR additives demonstrated that the addition of DMSO to PCRs improved results for all primer pairs except 25F/624R.

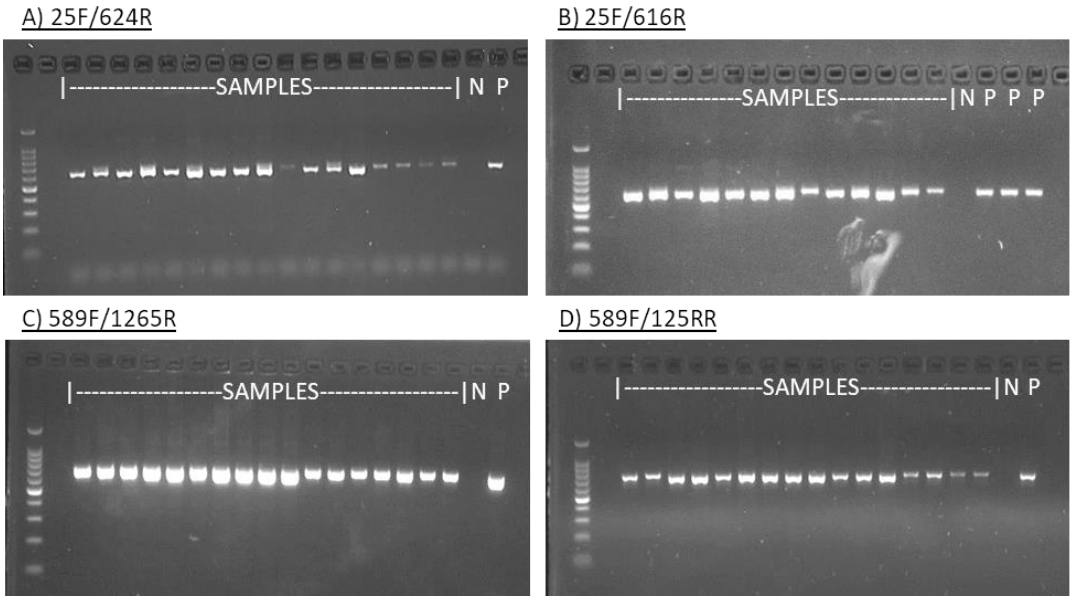



**Figure 3.3** Gel electrophoresis images showing the successful amplification of MHC fragments for DAB1, using four new primer pairs. Each PCR contained a negative control (water) and a positive control (a PCR product already confirmed as storm petrel MHC, from Sanger sequencing). All primer pairs showed positive results at their expected fragment lengths

**Table 3.5** Results of primer design, testing and optimisation when characterising exon 2 and its flanking regions in the DAB1 lineage.

| Primer Name and Combination | Success / Fail in PCR |
|---|---|
| 9F<br>627R | **REJECTED**: multiple additional bands on agarose gel, same brightness as expected amplicon. Not possible to optimise. |
| 18F<br>627R | **REJECTED**: multiple additional bands on agarose gel, same brightness as expected amplicon. Not possible to optimise. |
| 16F<br>619R | **FAILED** to amplify any DNA. |
| **25F**<br>**624R** | **SUCCESS**: Amplified a single fragment successfully, after optimisation. |
| **25F**<br>**616R** | **SUCCESS**: Amplified a single fragment successfully, after optimisation. |
| **589F**<br>**1265R** | **SUCCESS**: Amplified a single fragment successfully, after optimisation. |
| **597F**<br>**1265R** | **SUCCESS**: Amplified a single fragment successfully, after optimisation. |

Optimisation resulted in the four successful primer pairs yielding bright, clear, single bands on agarose (Figure 3.3). Sanger sequencing and a BLAST search confirmed a close match to published *H. castro* and *H. monteiroi* sequences MHC sequences on Genbank from Burri et al. (2016), demonstrating that the primers amplified both the correct species and the correct gene region. Sequences produced by these novel primers were subsequently used to design primers that targeted exon 2 of the MHC Class IIB, DAB1 lineage.

For DAB2, primers *Locus 2 Outer F* and *Locus 2 Outer R* from Dearborn et al. (2015) yielded clear sequencing results from two *H. castro* and two *H. monteiroi* individuals, including conserved regions suitable for primer design targeting exon 2 of the DAB2 lineage.

### 3.4.2: Design of lineage-specific primers targeting exon 2

For DAB1 a total of 6 forward primers and 10 reverse primers were initially designed. *In silico* testing using Geneious resulted in 1 forward and 3 reverse primers being rejected, and the remaining primers were ordered and subsequently tested in 13 different combinations, across 3 different individuals for each species (Table 3.6). PCR testing of primers resulted in 4 combinations being rejected. For DAB2, 3 forward and 3 reverse primers were designed, ordered and tested in PCR, using 9 different combinations. When testing DAB2 primers, 3 of these primer combinations failed to amplify a clear, single band on agarose gel, and were rejected.

**Table 3.6** Results of primer testing, targeting exon 2 of the DAB1 lineage. Whether primers were a success or fail is shown. Primers in bold were chosen to create MID-tag primers

| Primer names | Success / Fail |
|---|---|
| DAB1 273F<br>DAB1 609R | Worked in initial optimisation, sent for Sanger sequencing and rejected for displaying no heterozygous sites. Previous Sanger sequencing using primers for the longer fragment (25F, 616R, 624R, 589F, 597F and 1265R) picked up more heterozygous sites in the same samples. Failure to amplify the same sites with these exon-targeting primers demonstrates that allelic variation is being missed by these primers. Primers were therefore rejected. |
| DAB1 273F<br>DAB1 R1 | |
| DAB1 273F<br>DAB1 R3 | |
| DAB1 F4<br>DAB1 609R | |
| DAB1 F4<br>DAB1 R1 | |
| DAB1 F5<br>DAB1 607R | |
| DAB1 F5<br>DAB1 609R | |
| DAB1 F1<br>DAB1 R1 | Primer pairs failed to amplify regardless of optimisation steps taken. |
| DAB1 F3<br>DAB1 R3 | |
| DAB1 F1<br>DAB1 R3 | |
| DAB1 F2<br>DAB1 R1 | |
| **DAB1 F2<br>DAB1 R2** | **Primer pair successful, displayed the most heterozygous sites and picked for MID-tag development** |
| DAB1 F1<br>DAB1 R2 | Successful – sent for Sanger sequencing. Heterozygous sites picked up, but fewer than with chosen primer pair. |

Sanger sequences were compared, to find a primer pair for each DAB lineage that produced good quality sequence, whilst also detecting all previously observed heterozygous sites (Figure 3.4).

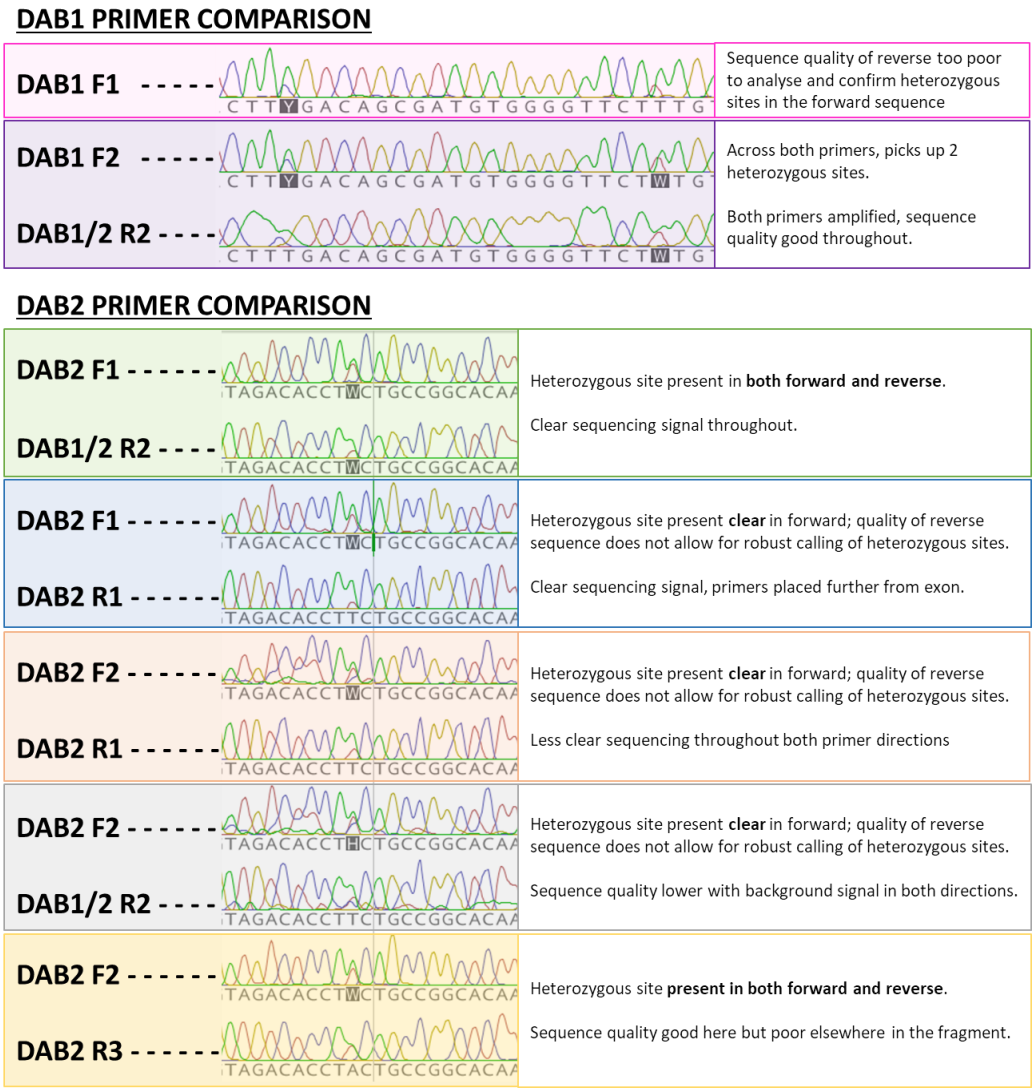



**Figure 3.4** Visual representation of how primers were assessed for suitability. Bases highlighted grey represent a heterozygous site. Primers yielding higher sequence quality and higher number of heterozygous sites were chosen.

For DAB1, seven primer pairs were rejected for containing no heterozygous sites, missing those detected when using primers for the flanking regions of exon 2. Of the remaining DAB1 primers, combination 'DAB1F2 + DAB1/DAB2 R2' amplified all previously detected heterozygous sites and some additional SNPs only seen in amplicons from this primer pair, so this pair was chosen for amplifying Exon II of the MHC Class IIB DAB1 gene, and subsequent development of MID-tag primers. For DAB2, 'DAB2F1 + DAB1/DAB2 R2' amplified all previously observed heterozygous sites, and provided the clearest

sequencing signal, with clean and readable peaks. In addition to amplifying well, the primers 'DAB2F1 + DAB1/DAB2 R2' were also placed more closely to the targeted exon II than all other primer pairs. Both DAB lineages used the same reverse primer (DAB1/DAB2 R2), as it fell in a region conserved between the two DAB lineages (also seen in Dearborn et al. 2014)

The chosen DAB1 and DAB2 primers (DAB1: F2 and DAB1/2 R2; DAB2: F1 & DAB1/2 F2) were then evaluated for performance in PCR amplification of 9 additional individuals. Both primer pairs produced distinct and bright band on agarose, at the expected length of approximately 300-310 bp (Figure 3.5).

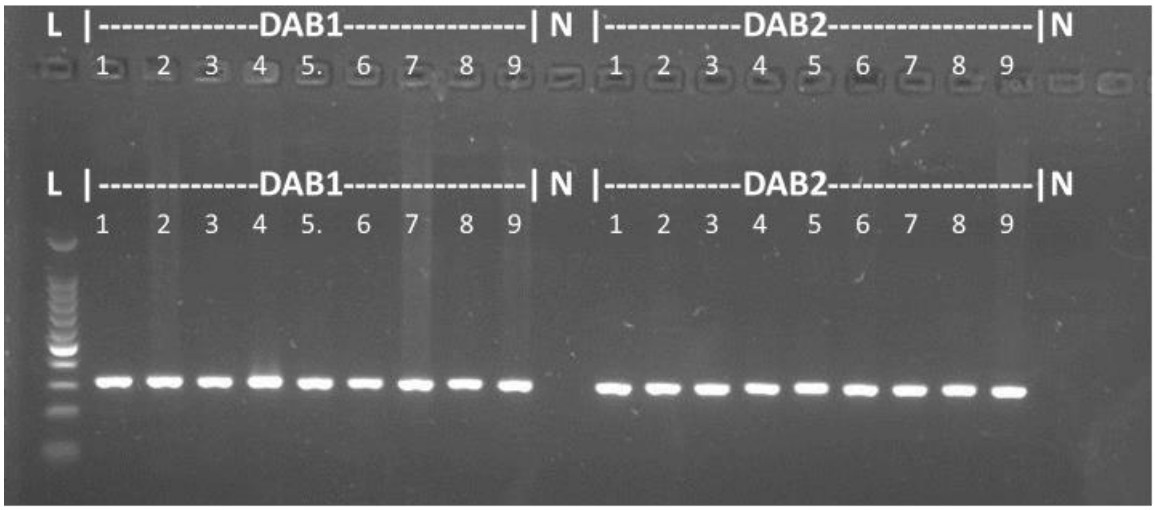



**Figure 3.5** Gel image showing successful PCR amplification with the chosen exon 2 primers (DAB1: DAB1 F2 & DAB1/DAB2 R2; and DAB2: DAB2 F1 & DAB1/DAB2 R2), tested on the same 9 samples (*H. castro*: 1-4; *H. monteiroi*: 5-9. L represents a 100 bp ladder (Promega); N are negative (no template) PCR controls.

### 3.4.3: Confirming amplification of both DAB Lineages Using Phylogenetics

Phylogenetic analysis showed that both DAB lineages are found in *H. castro* and *H. monteiroi*. The phylogenetic tree produced using 76 MHC sequences (Figure 3.6) displayed distinct groupings, largely defined by species. Exon 2 being a short fragment that is subject to balancing selection (Hedrick 1999), bootstrap values were typically low and some sequences clustered unexpectedly, e.g. a barn owl MHC sequence clustering near rock pigeon and ivory gull, as opposed to within the well-supported owl clade. However, there was a clear, distinct clustering of DAB1 and DAB2, each lineage showing recent common ancestry among Leach's storm petrel and the two band-rumped storm petrel taxa breeding on the Azores. Here, a phylogenetic separation according to DAB lineage rather than species was observed, consistent with the duplication of the two lineages prior to the divergence of these species.

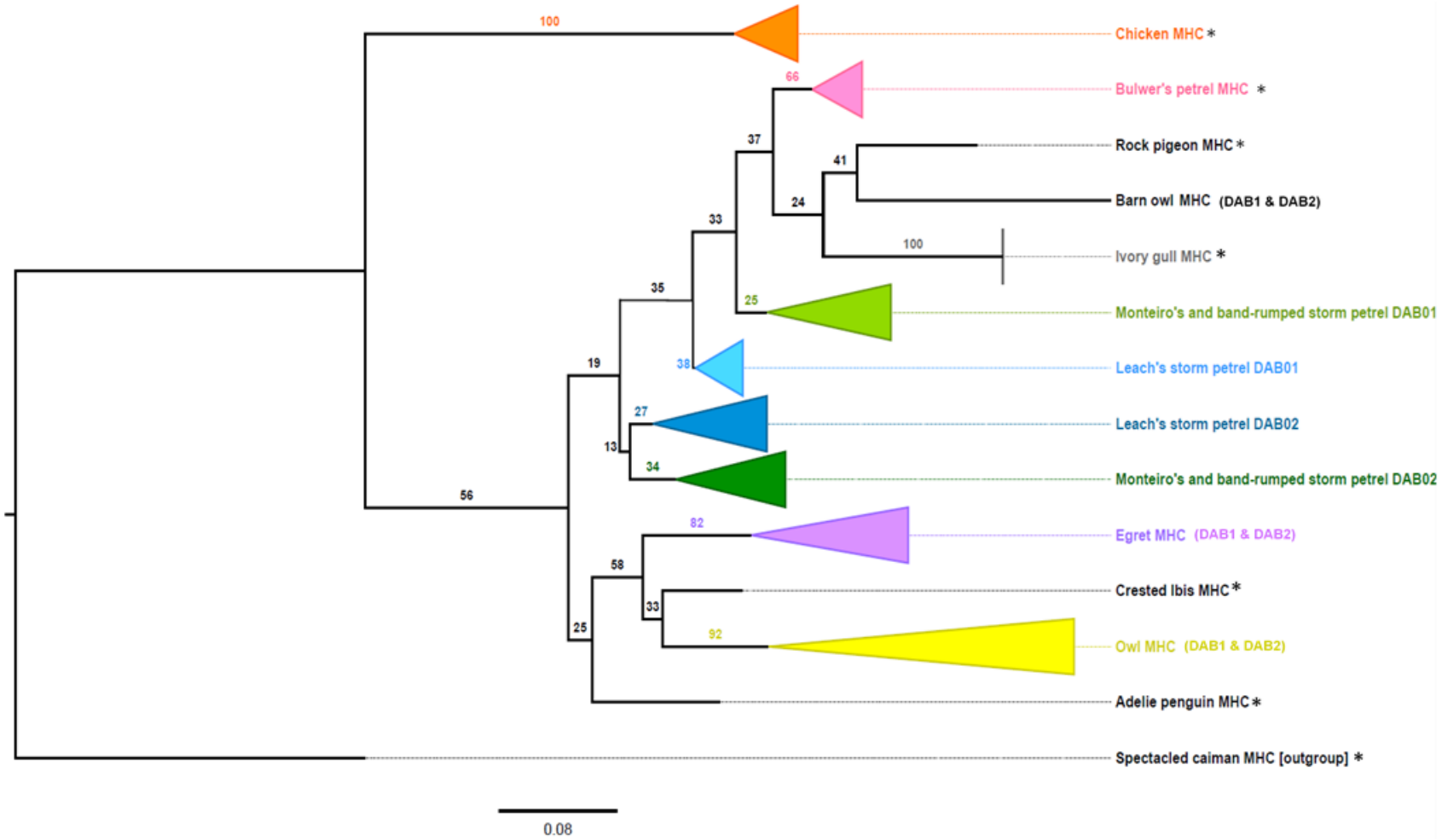



**Figure 3.6** Maximum likelihood phylogenetic tree of MHC Class IIB exon 2 sequences in multiple bird species, obtained from IQTree (Nguyen et al. 2015). Numbers on branches denote bootstrap support. The tree contains Class IIB exon 2 sequences from both DAB lineages in some species (see Table 3.4 for details). The tree and clade arrangements demonstrate that both DAB lineages are present in Monteiro's and band-rumped storm petrel and have been successfully amplified using exon-specific MHC primers developed here. An asterisk (*) denotes an unspecified DAB lineage

### 3.4.4: Characterisation of DNA polymorphism of DAB1 and DAB2 in storm petrels breeding on the Azores

Determining number of SNPs between DAB lineages and species showed that overall, DAB1 demonstrates more polymorphism than DAB2. The total number of polymorphic sites was higher overall in DAB1 than DAB2, with individual sequences also demonstrating higher numbers of SNPs. In terms of species-specific differences, the two DAB lineages differed somewhat. In DAB1, *H. monteiroi* demonstrated substantially more SNPs than *H. castro* (40 total site compared to 25, respectively), with a higher number of polymorphic sites both in total and between individuals. In contrast, DAB2 showed that *H. castro* was the species with a higher number of SNPs, with around twice as many total polymorphic sites. In DAB2, the between-individual numbers of SNPs did not differ as greatly between species.

**Table 3.7** Polymorphic sites present in exon 2, DAB-specific sequences. Polymorphic sites were determined using consensus sequences from the alignments created, whilst range was counted individually, and average was determined by dividing total sites across individuals by the total number of sites

| Lineage | Species | Number of individuals | Average number of polymorphic Sites | Total number of Polymorphic Sites | Range of polymorphic sites within individuals |
|---|---|---|---|---|---|
| **DAB1** | *H. castro* | 9 Primer combos; 3 diff indivs. | 2.68 if div by total<br>7.4 if indivs | 25 | 0-13 |
| | *H. monteiroi* | 4 combos<br>2 indivs | 1.5 if total<br>15.75 if indivs | 40 | 11-23 |
| **DAB2** | *H. castro* | 7 combos<br>3 indivs | 0.6 if total sites<br>2.7 if indivs | 32 | 0-10 |
| | *H. monteiroi* | 5 combos 2 indivs | 2.5 if all sites<br>7.6 if indivs | 15 | 4-9 |

The testing of multiple primers, and subsequent comparison of polymorphisms demonstrated that primer pairs differed in their ability to pick up on multiple alleles. Some primers were too specific, with no polymorphic sites detected despite other primer pairs showing the contrary, for the same individuals. This assessment allowed for primers to be scrutinised and rejected if they proved to amplify only one allele. Using this data, the final exon 2 primers were selected

based on both sequence quality and number of polymorphic sites determined. The primer pair that detected the most polymorphic sites, whilst also providing good-quality sequence data was chosen, to be used in downstream analysis such as HTS (DAB1: DAB1 F2 & DAB1/DAB2 R2; and DAB2: DAB2 F1 & DAB1/DAB2 R2).

## 3.5: Discussion

### 3.5.1. Maintenance of both lineages of the ancestral avian DAB duplication in storm petrels breeding in the Azores

The present study is the first to generate DAB2 sequences for band-rumped storm petrels breeding in the Azores, and the first to demonstrate the presence of both DAB1 and DAB2 in these species. Previous research on avian MHC DNA sequences demonstrated the presence of two deeply divergent DAB lineages, resulting from an ancient duplication event that pre-dates the radiation of extant avian lineages (Burri et al. 2008; Goebel et al. 2017). Prior to the present study, only the DAB1 lineage of the MHC Class IIB gene region had been sequenced for *H. castro* and *H. monteiroi* (Burri et al. 2014) – and its DAB lineage affiliation had not yet been identified. The affiliation of the DAB sequences from *H. castro* and *H. monteiroi* generated by Burri et al. (2014) is shown to be with DAB1. The findings of the present study show that these storm petrels possess both DAB lineages. Bootstrap support for the divergence of these lineages is limited (Figure 3.6), however this may be due to the use of exon 2 sequences, which typically show closer relationships between species (Burri et al. 2008). These results mirror previous findings for Leach's storm petrels by Dearborn et al. (2015), suggesting that in all three storm petrel species, signals of an ancient DAB duplication have been retained, despite the homogenising force of gene conversion that has assimilated DAB lineages in other branches of the avian tree (Burri et al. 2010). In other avian species, only one DAB lineage remains or is detectable (Nei et al. 1997; Burri et al. 2010), believed to result from a birth-death scenario where lineages arise through duplication, and are subsequently lost or modified through deletions and/or gene conversion, with the latter leading to secondary assimilation of the originally divergent DAB lineages (Goebel et al. 2017). Based on the time required for lineage sorting according to coalescent theory (e.g. Nichols 2001), the lack of reciprocal monophyly of *H.*

*castro* and *H. monteiroi* sequences for DAB1 as well as for DAB2 is likely the consequence of incomplete lineage sorting among this species pair that is believed to have diverged ca. 110,000-180,000 years ago (Friesen et al. 2007).

A main reason for the focus of many evolutionary genetic studies on MHC sequences is the importance of this gene family in individual fitness and mate choice (Yamazaki 1976b; Piertney and Oliver 2006; Juola and Dearborn 2012; Gasparini et al. 2015). The present finding of both DAB lineages in *H. castro* and *H. monteiroi* therefore necessitates the utilisation of methods that encompass this broad allelic and locus spectrum - both DAB lineages represent multiple alleles, which jointly contribute to a varied MHC profile in individuals.

To allow for comprehensive characterisation of MHC Class IIB diversity, and its potential effects on behaviours such as mate choice in *H. castro* and *H. monteiroi*, this study reports novel primers for amplification of both DAB lineages. The primers yield a short fragment (<280 bp), suitable for HTS using shorts reads (e.g., on Illumina instruments). The primers are targeting exon 2 and the encompassed peptide binding site for each DAB lineage (Brown et al. 1993; Sommer 2005a). The same DAB-specific primer sets for exon 2 were suitable for use on both species of storm petrels breeding on the Azores. Hence, the optimised primers were not required to be species-specific (instead being suitable for both species), but lineage-specific design was required given the absence of suitable, conserved regions near exon 2, and the requirements on fragment length for HTS using short-read technology such as Illumina. The presented tests suggest that the developed primers are reliable and robust, producing clear and strong bands on agarose gels across an array of tested individuals, and yielding good-quality Sanger-sequencing data. The presence of clear heterozygous sites further demonstrated that multiple alleles were being detected by the primers. Future investigations utilising these primers for HTS are expected to enable precise characterisation of the allelic repertoire of DAB1 and DAB2 in these two taxa of band-rumped storm petrels.

### 3.5.2. Exon 2 variability of DAB loci in Storm Petrels breeding in the Azores

Sanger sequencing demonstrated that both DAB lineages are highly variable and possess multiple alleles in *H. castro* and *H. monteiroi*. Between primers

amplifying a larger portion of the MHC (Section 3.2.1 and 3.2.2) and those amplifying exon 2 (Section 3.2.3), no clear difference in levels of variability was observed. For both DAB lineages, there was clear evidence that >1 nucleotide was being amplified at several sites, indicative of two or more alleles being amplified.

DAB1 contained a higher number of variable sites than DAB2, unlike in Leach's storm petrel, where DAB2 had more alleles than DAB1 (Dearborn et al. 2014). Regardless, results from these three storm petrel species suggest that polymorphism within DAB lineages may be common in this taxonomic group.

Limitations of the conducted direct Sanger sequencing (i.e., without prior cloning) imply that determination of allele numbers is difficult or impossible. While algorithms commonly used to deduce SNP phasing from direct sequencing data (e.g. PHASE; Stephens et al. 2001) are effective for diploid sequences, this approach is not applicable for this data, given the likely presence of three or more DAB alleles at least in some individuals. A further limitation of the chosen approach is that direct sequencing may miss some variability, especially if alleles within individuals amplify unequally. Similarly, sequencing signal quality is typically not sufficient to allow robust characterisation of sites with more than two nucleotides, as is common for DAB exon 2 data (e.g., Dearborn et al. 2016): true allelic variability can be missed, or 'false' heterozygosity can be recorded due to PCR or sequencing artifacts. To mitigate these issues, chromatograms were analysed as stringently as possible and numerous repeats were conducted. A strategy involving cloning was previously used by Burri et al. (2014) to detect multiple MHC alleles in birds. While this has advantages over direct Sanger sequencing, interpretation of cloning-derived sequencing data is complicated by artefacts arising from mutations arising during PCR or bacterial replication (due to their lack of polymerase proof reading capability), also producing false SNPs (Hoseini and Sauer 2015). Ultimately, HTS approaches – facilitated by the primers developed in this study – are likely to provide much more accurate estimates of DAB variability in storm petrels.

### 3.5.3. Conclusions and Future Directions

The demonstration that both DAB lineages exist in *H. castro* and *H. monteiroi* is a novel discovery for these species, and the design of primers targeting exon 2 in each lineage provides a novel method for characterising variability in this region. Phylogenetic analysis proved to be a useful tool in confirming the presence of both lineages and demonstrates that the divergence of the two DAB lineages precedes speciation in *H. castro* and *H. monteiroi*.

An MHC method development strategy that followed the recommendations by Burri et al. (2014) was found to be effective; designing multiple primer pairs and trialling them in different PCR mixes and conditions proved useful in determining whether the primers picked up on heterozygosity or not.

For future studies, the primers designed can be used as MID-tagged primers, whereby 8-bp unique identifiers for each sample are added to the forward and reverse sequences of the chosen primers (Binladen et al. 2007). Here, the fact that the developed primers target DAB1 and DAB2 with a lineage-specific forward primer, but a shared reverse primer, significantly reduces costs when ordering MID-tagged primers for Illumina High-Throughput Sequencing (HTS).

In summary, the novel developed primers appear suitable for HTS subsequent detailed characterisation of DAB lineages in *H. castro* and *H. monteiroi* (**Chapter 4**).

## 3.6 Acknowledgements

Thank you to Lucy Rowley for help with lab work, as always. Thank you to fellow lab members for their advice when I got stuck with this method development, and their encouragement and company during long lab hours. Thank you to Frank for advice, encouragement, and knowledge, when the development of these primers and methods seemed endless. Thank you to Rob, Carsten, and Renata for your ideas on how to formulate this into a full chapter, and not just a boring lab methods section.

# Chapter Four: High-throughput Sequencing of MHC and Divergence of Azores storm petrels

*An adult Hydrobates monteiroi, on Praia islet*

**"It's better to enjoy life - selection is a temporary thing"**
*An adult Hydrobates monteiroi, on Praia islet*

**- Shreyas Iyer**

# Chapter 4 High-throughput Sequencing of MHC and Divergence of Azores Storm Petrels

## 4.1: Introduction

The recent speciation of *H. castro* and *H. monteiroi* has been described in **Chapter 2**. Briefly, previous genetic studies of *H. castro* and *H. monteiroi* have demonstrated that these recently diverged species show a reasonably pronounced genetic differentiation. Based on analysis of mtDNA sequences, *H. monteiroi* samples generally fall within a clade that clusters separately from North Atlantic *H. castro*. A low incidence of clade-sharing is however observed, with some *H. monteiroi* individuals clustered amongst *H. castro*, and vice-versa (Friesen et al. 2007; Smith and Friesen 2007; Smith et al. 2007; Silva et al. 2016; **Chapter 2**). From analysis of 12 anonymous nuclear loci, Silva et al. (2016) found that genetic distinctiveness at individual loci was far less pronounced than for mtDNA, with high levels of haplotype sharing and absence of any distinct species-specific clusters in haplotype networks. When combined in multi-locus analyses, the nuclear loci by Silva et al. (2016) and the ddRAD sequencing data from Taylor et al. (2019) revealed overall clear genetic differentiation between *H. castro* and *H. monteiroi*.

Discordance between haplotype sharing patterns between mitochondrial and nuclear DNA is common, and theoretically predicted due to the larger effective population size and thus longer coalescence time of nuclear than mitochondrial loci (Zink and Barrowclough 2008; Toews and Brelsford 2012). This discordance is commonly observed in large populations of recently diverged demes, with mtDNA reaching reciprocal monophyly faster than nuclear DNA. The latter therefore often displays patterns of incomplete lineage sorting (ILS; Welch et al. 2011; Gangloff et al. 2013). For these reasons, mitochondrial DNA is usually considered a 'leading' indicator of population differentiation, and nuclear DNA a more 'lagging' indicator (Zink and Barrowclough 2008).

Encoded by nuclear DNA, MHC loci have the potential to display similar lagging signals of differentiation, however MHC-encoded traits under positive selection could cause MHC genes to diverge faster than other nuclear loci (Zink and

Barrowclough 2008). As explained earlier (**Chapters 1 and 3**), MHC loci can be subject to strong balancing and/or directional selection (Hedrick 1999), reflecting their role in immune response and the resulting large selective advantage of certain alleles (Piertney and Oliver 2006), and e.g. exposure to common pathogens can shape MHC diversity (Bernatchez and Landry 2003). High allelic variation at the MHC can also be maintained in populations through sexual selection and mate choice (Jordan and Bruford 1998). Individuals may choose mates based on MHC heterozygosity, complementarity, or perhaps 'good genes' to provide a fitness advantage in offspring (Eizaguirre et al. 2009b; Huchard et al. 2010). Given its adaptive value in terms of selection, the MHC could provide particularly interesting insights into evolutionary pressures acting on recently diverged species. Balancing selection and incomplete lineage sorting could be occurring at MHC loci, showing a lack of differentiation between species in MHC sequences (Welch et al. 2011). Conversely, strong environmental or sexual selection, along with reproductive isolation, could be causing rapid divergence of MHC loci between *H. castro* and *H. monteiroi* (Dean et al. 2021). However, if both species are under similar patterns of sexual or environmental selection (i.e., targeting the same levels of MHC diversity, or subject to the same pathogens), MHC loci may be indistinguishable between species. Species and individual recognition for many species also links back to the MHC. Individuals may using MHC-encoded information to identify conspecifics (Caro and Balthazart 2010), which has been suggested to contribute to pre-mating isolation between species (Eizaguirre et al. 2011). Investigating MHC divergence in *H. castro* and *H. monteiroi* could therefore shed more light on the selection pressures involved, and whether an interplay between MHC, selection and mate choice is contributing to genetic differentiation between the two species.

Comprehensive characterisation of MHC divergence between species in birds is complicated by an ancient duplication of the MHC Class II that has resulted in two distinct DAB lineages (Burri et al. 2008). In some avian species, one lineage has been lost by deletion, or become assimilated to the other DAB lineage due to gene conversion (Goebel et al. 2017). In other bird species, the ancient duplication has been maintained, e.g. in Leach's storm petrels (*H. leucorhous*; Dearborn et al. 2015). In **Chapter 3** of this thesis, it was confirmed that both *H. castro* and *H. monteiroi* have retained both these MHC lineages. MHC genes are

highly polymorphic, and individuals can have multiple alleles per gene (Janeaway et al. 2001a). Coupled with the ancient duplication present in *H. castro* and *H. monteiroi*, determining individuals' MHC profiles can be challenging. Direct Sanger sequencing alone is not suited for multi-allelic and/or multi-copy loci due to analytical challenges to extract individual haplotypes from the combined Sanger electropherograms (Griffin et al. 2011). This challenge can be overcome with cloning (e.g. Burri et al. 2014), however this is time-consuming, costly, and can be prone to PCR and bacterial replication errors that result in false substitutions. In the recent decade, high-throughput sequencing (HTS) approaches allow rapid, highly parallelised and accurate sequencing of millions of bases, enabling characterisation of multiple alleles independently due to the single-molecule nature of many such approaches (Margulies et al. 2005; Sikkema-Raddatz et al. 2013; Behjati and Tarpey 2013). As HTS costs keep decreasing (Glenn 2011), numerous MHC genotyping studies have utilised HTS systems such as 454, Ion Torrent and Illumina (Lighten et al. 2014; Dearborn et al. 2015; Rekdal et al. 2018). The Illumina MiSeq platform provides short-read fragments, relatively high accuracy, and is relatively cheap for small to medium-sized projects in comparison to other methods (see Glenn 2011). At approximately 270bp long, the MHC Class IIB exon 2 fragment, isolated in Chapter 3, is suited to Illumina MiSeq sequencing, as proven in other studies that have used Illumina for such MHC genotyping (Lighten et al. 2014; Dearborn et al. 2015b; Gaigher et al. 2016; Biedrzycka et al. 2017b). When sequencing with HTS, samples can be pooled and sequenced in parallel, by adding a unique multiplex identifier (MID) tag to each individual (Glenn 2011; Zavodna et al. 2013). These are added to the primers used in PCR amplification, and in paired-end sequencing dual-indexed primers can be used to generate unique MID combinations (Rekdal et al. 2018).

## 4.2: Aims:

The primers developed in **Chapter 3** were used in conjunction with Illumina MiSeq sequencing to provide the first comprehensive characterisation of the DAB1 and DAB2 lineage of the MHC Class IIB gene region in *H. castro* and *H. monteiroi*. The main aims of this chapter were: (1) Develop a rigorous bioinformatic data filtering pipeline, following PCR amplification and sequencing, to identify the true alleles found for each individual. Subsequently,

the obtained data were used to determine (2) the minimum number of distinct DAB1 and DAB2 locus copies within each species, and characterise and compare levels of MHC genetic variability between (3) the DAB loci and (4) the two storm petrel species. (5) Data from mated pairs and their offspring was used to verify whether the alleles show Mendelian inheritance. (6) Finally, the obtained data were used to investigate levels of genetic differentiation for these adaptively relevant MHC loci between the recently diverged *H. castro* and *H. monteiroi*, informing how individual mate choice might translate to population/species-level differentiation.

## 4.3: Methods

### 4.3.1: Sample Selection for Illumina Sequencing

The approach when amplifying exon II of the MHC Class IIB was to allow for eventual comparison of mated pairs, rather than taking a randomised sampling approach. For both species, long-term breeding data on mated pairs and their offspring (224 samples for *H. castro* and 312 records for *H. monteiroi)* have been collected since 2002 by a team of researchers including Joël Bried, Veronica Neves, Renata Medeiros-Mirra, Hannah Hereward and Mark Bolton. Due to the long-lived nature of the birds and their strong philopatry (Bolton et al. 2008a), repeat records are common across years. Birds that had mated more than once and those that had changed partner across years were targeted for selection, to allow future research to perform comparisons of MHC across different pairings. Birds that were recorded with a chick or an egg were also targeted, as the presence of these confirm a mating occurred, and therefore a 'true' mated pair. All *H. castro* samples were collected in 2002, while the dataset for *H. monteiroi* was dispersed over a larger range of years (mostly 2002 – 2005, but with some *H. monteiroi* samples from 2006, 2007 and 2017, if these records were from breeders or chicks sampled in the earlier years). The inclusion of these later samples will allow future investigations of drivers of mate choice and mate fidelity in the species. Overall, both studied storm petrel species are long lived, breed generally annually, and exhibit strong site fidelity and natal philopatry. Long term studies on the islets using recapture techniques suggest the same birds come back annually to breed (Joel Bried, personal communication), so the sampled birds from each species can reasonably be considered one population

data set per species, and individuals sampled are highly likely to be breeding in multiple years across the sampling period, even if not sampled that year (similar to the approach taken by Dearborn et al 2016, in their mate-choice analyses). Despite a mean generation time of 12 years (Smith et al 2007), it was decided that sampling the species in different years would be comparable, as it was not expected that genetic differentiation would be strongly affected between sampling years.

Fifty storm petrel families were ultimately selected for analysis with the MHC primers. For both species, 12 families included not only the mates, but also samples taken from their chick, allowing checks of Mendelian inheritance of MHC alleles. The total dataset consisted of 226 samples, 112 from *H. castro* and 114 from *H. monteiroi* (Tables 4.1 and 4.2, respectively). DNA was isolated as explained in previous chapters, and samples were sexed using PCR with primers by Fridolfsson and Ellegren (1999).

**Table 4.1** Samples from H. castro used in Illumina HTS sequencing, utilising MID-tagged primers. Family code is denoted with the prefix 'FC…'. Sample names are prefixed 'OC…'. Samples highlighted in red indicate chicks. Additional sample details, including sex and sample date, can be found in the Appendix for Chapter 4

***H. castro***

| Family | Sample |
|---|---|
| FC002 | OC042 |
| FC002 | OC080 |
| **FC003** | **OC041** |
| FC003 | OC053 |
| FC003 | OC066 |
| FC004 | OC008 |
| FC004 | OC088 |
| FC005 | OC036 |
| **FC005** | **OC202** |
| FC005 | OC085 |
| FC009 | OC045 |
| FC009 | OC050 |
| FC008 | OC035 |
| FC008 | OC086 |
| FC011 | OC010 |
| FC011 | OC011 |
| FC014 | OC016 |
| FC014 | OC030 |
| FC019 | OC033 |
| FC019 | OC034 |
| FC022 | OC013 |
| FC022 | OC082 |
| FC023 | OC003 |
| FC023 | OC052 |
| **FC023** | **OC004** |
| FC026 | OC017 |
| **FC026** | **OC199** |
| FC026 | OC058 |

| Family | Sample |
|---|---|
| FC027 | OC022 |
| FC027 | OC072 |
| FC028 | OC009 |
| FC028 | OC039 |
| FC032 | OC023 |
| FC032 | OC078 |
| **FC032** | **OC073** |
| FC035 | OC014 |
| FC035 | OC067 |
| FC036 | OC018 |
| FC036 | OC059 |
| FC039 | OC021 |
| FC039 | OC061 |
| FC041 | OC026 |
| FC041 | OC091 |
| FC042 | OC005 |
| FC042 | OC090 |
| FC045 | OC007 |
| FC045 | OC055 |
| FC047 | OC012 |
| FC047 | OC040 |
| FC048 | OC031 |
| FC048 | OC064 |
| FC054 | OC024 |
| FC054 | OC025 |
| FC056 | OC028 |
| FC056 | OC071 |

| Family | Sample |
|---|---|
| FC059 | OC153 |
| FC059 | OC154 |
| **FC059** | **OC170** |
| FC060 | OC132 |
| FC060 | OC193 |
| FC062 | OC099 |
| FC062 | OC168 |
| FC064 | OC101 |
| FC064 | OC151 |
| **FC064** | **OC102** |
| FC065 | OC137 |
| FC065 | OC149 |
| FC066 | OC167 |
| FC066 | OC182 |
| **FC066** | **OC104** |
| FC071 | OC110 |
| FC071 | OC172 |
| FC073 | OC150 |
| FC073 | OC164 |
| **FC073** | **OC112** |
| FC074 | OC113 |
| FC074 | OC189 |
| FC076 | OC116 |
| FC076 | OC173 |
| FC078 | OC118 |
| FC078 | OC190 |
| FC083 | OC097 |
| FC083 | OC176 |

| Family | Sample |
|---|---|
| FC084 | OC098 |
| FC084 | OC178 |
| FC086 | OC124 |
| **FC086** | **OC214** |
| FC086 | OC175 |
| FC090 | OC128 |
| FC090 | OC196 |
| **FC090** | **OC184** |
| FC092 | OC131 |
| FC092 | OC191 |
| FC094 | OC135 |
| FC094 | OC136 |
| FC097 | OC139 |
| FC097 | OC174 |
| FC098 | OC141 |
| FC098 | OC186 |
| FC099 | OC142 |
| FC099 | OC188 |
| FC101 | OC140 |
| FC101 | OC179 |
| FC102 | OC120 |
| FC102 | OC165 |
| FC104 | OC166 |
| FC104 | OC192 |
| **FC104** | **OC119** |
| FC105 | OC145 |
| FC105 | OC147 |
| FC111 | OC130 |
| FC111 | OC185 |

**Table 4.2** Samples from H. monteiroi used in Illumina HTS sequencing, utilising MID-tagged primers. Family code is denoted with 'FM…'. Sample names are prefixed 'OC…'. Samples highlighted in red indicate chicks. Additional sample details, including sex and sample date, can be found in the Appendix for Chapter 4

***H. monteiroi***

| Family | Sample |
|---|---|
| FM001 | OC225 |
| FM001 | OC255 |
| **FM001** | **OC397** |
| FM002 | OC239 |
| FM002 | OC279 |
| FM003 | OC247 |
| FM003 | OC288 |
| **FM003** | **OC383** |
| FM004 | OC246 |
| FM004 | OC297 |
| FM005 | OC248 |
| FM005 | OC299 |
| **FM005** | **OC421** |
| FM005A | OC245 |
| FM005A | OC248 |
| **FM005A** | **OC392** |
| FM006 | OC216 |
| FM006 | OC270 |
| FM007 | OC217 |
| FM007 | OC271 |
| FM008 | OC251 |
| FM008 | OC283 |
| FM009 | OC218 |
| FM009 | OC276 |
| FM010 | OC219 |
| FM010 | OC252 |
| FM011 | OC338 |
| FM011 | OC340 |
| FM011A | OC338 |
| FM011A | OC419 |

| Family | Sample |
|---|---|
| FM012 | OC220 |
| FM012 | OC272 |
| FM014 | OC223 |
| FM014 | OC253 |
| **FM014** | **OC310** |
| FM014A | OC252 |
| FM014A | OC253 |
| **FM014A** | **OC408** |
| FM014B | OC223 |
| FM014B | OMA40 |
| FM017 | OC229 |
| FM017 | OC235 |
| FM017 | OC277 |
| **FM017** | **OC314** |
| FM018 | OC243 |
| FM018 | OC285 |
| FM019 | OC244 |
| FM019 | OC286 |
| FM019A | OC286 |
| FM019A | OC421 |
| FM020 | OC245 |
| FM020 | OC287 |
| FM024 | OC228 |
| FM024 | OC258 |
| FM025 | OC341 |
| FM025 | OC351 |
| FM026 | OC221 |
| FM026 | OC390 |
| FM027 | OC232 |
| FM027 | OC266=OC348 |

| Family | Sample |
|---|---|
| FM027A | OC266=OC348 |
| FM027A | OC369 |
| FM029 | OC233 |
| FM029 | OC265 |
| FM031 | OC231 |
| FM031 | OC284 |
| FM033 | OC237 |
| FM033 | OC278 |
| FM034 | OC238 |
| FM034 | OC266=OC348 |
| FM035 | OC267 |
| FM035 | OC268 |
| FM039 | OC224 |
| FM039 | OC273 |
| **FM039** | **OC316** |
| FM039A | OC273 |
| FM039A | OC373 |
| **FM039A** | **OC409** |
| FM040 | OC269 |
| FM040 | OC298 |
| FM041 | OC275 |
| FM041 | OC301 |
| FM042 | OC337 |
| FM042 | OC343 |
| FM043 | OC261 |
| FM043 | OC295 |
| FM044 | OC242 |
| FM044 | OC300 |
| FM045 | OC250 |
| FM045 | OC290 |

| Family | Sample |
|---|---|
| FM046 | OC263 |
| FM046 | OC293 |
| FM047 | OC262 |
| FM047 | OC296 |
| FM048 | OC403 |
| FM048 | OC410 |
| FM048A | OC410 |
| FM048A | OMA42 |
| FM049 | OC389 |
| FM049 | OC414 |
| FM049A | OC389 |
| FM049A | OMA41 |
| FM054 | OC383 |
| FM054 | OC385 |
| **FM054** | **OC379** |
| FM056 | OC367 |
| FM056 | OC390 |
| **FM056** | **OC413** |
| FM059 | OC249 |
| FM059 | OC289 |
| **FM059** | **OC378** |
| **FM059** | **OC420** |
| FM059A | OC249 |
| FM059A | OC391 |

### 4.3.2: MID-tagging and PCR Amplification

The primers designed to amplify DAB1 and DAB2 (DAB1 F2 + DAB1/2 R2; DAB2 F1 + DAB1/2 R2) were adapted for use in Illumina HTS, by adding Multiplex Identifier tags (MID-tags). MID-tags are 8-10 bp fragments added to the 5' end of the primers, providing a unique identifier for each individual. In downstream analysis of HTS outputs, data can be de-multiplexed according to the unique identifiers, allowing individual identification (Binladen et al. 2007; Brown et al. 2014).

In this study, 8 bp MID-tags were added to both the forward and reverse primers, providing in combination a unique identifier for each individual or the various types negative/amplification controls. Negative controls served as indicators for contamination and tag-jumping in downstream analyses. These negative controls consisted of 7 PCR negatives (PCR reactions containing water in lieu of DNA), and 22 extraction negatives (created as a control during DNA extractions), bringing the total sample size to 255 for each DAB lineage. To account for some PCRs failing, MID-tags for at least 300 unique combinations were initially considered, ensuring there would be enough additional combinations for repeats. Due to a conserved region at the end of exon III, both DAB lineages share the same reverse primer, locus specificity provided by unique forward primers (**see Chapter 3**). The MID tags were taken from Taberlet et al. (2018), and checked to ensure that they did not coincidentally match the region of MHC flanking each primer, ensuring tags uniqueness. Subsequently 50 tags were chosen, consisting of 13 unique MID-tagged forward primers for each DAB lineage, and 24 shared MID-tagged reverse primers (Table 4.2 and 4.3). This resulted in 312 possible combinations for each DAB lineage, adequately covering all samples, negative controls and any repeats. MID-tagged primers were ordered from Eurofins Genomics (Germany).

**Table 4.3** Forward primers and the MID-tags generated for use in Illumina HTS sequencing

| **DAB1 Forward Primer (5'-3'):** | **GACCTCTCCATGTYTGCACGAR** | | **DAB2 Forward Primer (5'-3'):** | **GACCTGCCTCCCTGCACAAACA** |
|---|---|---|---|---|
| **MID-Tag Primer** | **MID-Tag Sequence (5'-3')** | | **MID-Tag Primer** | **MID-Tag Sequence (5'-3')** |
| Locus 1 F1 | ATGCTGAC | | Locus 2 F1 | ACAAGACC |
| Locus 1 F2 | CAACGAGA | | Locus 2 F2 | CAGTAGAC |
| Locus 1 F3 | TCATCCTG | | Locus 2 F3 | CAACGTAC |
| Locus 1 F4 | TCTCCAGA | | Locus 2 F4 | AGAGAACG |
| Locus 1 F5 | GATAGAGG | | Locus 2 F5 | CATCAGCA |
| Locus 1 F6 | GCCAGTTA | | Locus 2 F6 | CATCGGAT |
| Locus 1 F7 | ACGCAGTA | | Locus 2 F7 | GGACTATG |
| Locus 1 F8 | TCGTATGG | | Locus 2 F8 | CTAACTCC |
| Locus 1 F9 | CTGCTATC | | Locus 2 F9 | CAAGACTC |
| Locus 1 F10 | TCATACGC | | Locus 2 F10 | CTAGCAGA |
| Locus 1 F11 | TCCAACAC | | Locus 2 F11 | AATACGGC |
| Locus 1 F12 | ATACCGGA | | Locus 2 F12 | TGTCGTAC |
| Locus 1 F13 | CCTTGGAA | | Locus 2 F13 | TGTGCTTG |

**Table 4.4** Reverse primers and MID-tags generated for Illumina HTS sequencing.

| **Reverse Primer (5'-3'):** | **GCAATGTTCTGCCMAGCACT** |
|---|---|
| **MID-Tag Primer Name** | **MID-Tag Sequence (5'-3')** |
| Both Loci R1 | TTACGCCA |
| Both Loci R2 | CAGCTCAT |
| Both Loci R3 | GTGTAGTC |
| Both Loci R4 | AGAATGCC |
| Both Loci R5 | TTAGGCAC |
| Both Loci R6 | TGCACTAC |
| Both Loci R7 | AGAGCTAC |
| Both Loci R8 | GATGCCTT |
| Both Loci R9 | ATGAGCAC |
| Both Loci R10 | TGGAACCT |
| Both Loci R11 | TTCCTCAC |
| Both Loci R12 | TACTCTCG |
| Both Loci R13 | AAGGACAC |
| Both Loci R14 | GAACTGCA |
| Both Loci R15 | GTGTACCA |
| Both Loci R16 | TGCCATCT |
| Both Loci R17 | GTAGCGAA |
| Both Loci R18 | TGTATCGG |
| Both Loci R19 | TAGAGGAC |
| Both Loci R20 | AGGTGACT |
| Both Loci R21 | TTCGTGCT |
| Both Loci R22 | AGAGGCAA |
| Both Loci R23 | AACAGGAG |
| Both Loci R24 | AAGCCGAA |

A subset of MID-tagged primers were initially tested on a small number of samples, with minor adjustments made to PCR cycling conditions to achieve strong amplification. MID-tagged PCRs were carried out in 15 µL reaction volumes containing 0.5 µL template DNA, 1X GoTaq colourless buffer, 1.5 µM $MgCl_2$, 0.2 mM of each dNTP, 0.2 µM of each MID-tagged primer, 0.03 U/µL Taq polymerase and 9 µL ultrapure water. PCRs were run on an Applied Biosystems SimpliAmp thermal cycler. PCR reaction conditions began with a 2-minute denaturation step at 95°C, 40 cycles of 95°C for 30 seconds, 55°C for 45 seconds and 72°C for 1 minute, finishing with a final extension of 5 minutes at 72°C. PCR cycles were capped at 40 to limit the formation of chimeras (sequences formed from two or more different alleles) at least somewhat (Sommer et al 2013). Chimeras often form in the latter cycles of PCR, when dNTP and primer concentrations are lower, so limiting cycles helps avoid this (Lenz and Becker 2008). In this study where multiple alleles were expected and reducing chimera formation was therefore important to avoid misidentifying chimeras for true alleles. All PCR products were run on 1.5% agarose gel, stained with SYBRSafe (ThermoFisher Scientific, UK) and visualised under UV light. Each PCR reaction included three PCR negatives to monitor for contamination. In cases where storm petrel samples failed to produce a band, the samples were repeated using a different MID-tag combination. One MID-tag, Locus 1 F1, failed to produce any bands, regardless of which reverse primer was used. The corresponding 24 failed samples were repeated with Locus 1 F11 and then amplified well.

### 4.3.3: Pooling and Library Preparation for Illumina HTS Sequencing.

All samples that produced a band on agarose gel were pooled for Illumina sequencing, whereas only a proportion of all negatives (10% of total pool volume) were included in the pools. PCR products were mixed in approximate equimolar amounts into one single 'pool'. As the DAB lineages used the same reverse primer, pooling was performed separately for these. For each gel image and PCR result, the bands presented were put into two categories according to brightness (bright or faint; Figure 4.1). The brightness was taken as a proxy for concentration and used as a guide for deciding amounts of PCR product to pool. For bands that were considered 'bright', 1 µL of PCR product was added to the pool, and for 'faint' bands the volume was 3 µL. Where negatives were added to a pool, 1 µL was used; this lower volume was used to avoid or reduce the

possibility of diluting DNA within the pools. This initial mixing produced a total of 8 pools, representing four pools for each lineage.

Samples were initially pooled by PCR plate (four for each DAB lineage), and later combined into one 'super pool' per DAB lineage. DAB1 super pool contained a total of 210 samples and 15 negatives. DAB2 super pool contained 204 samples and 17 negatives.

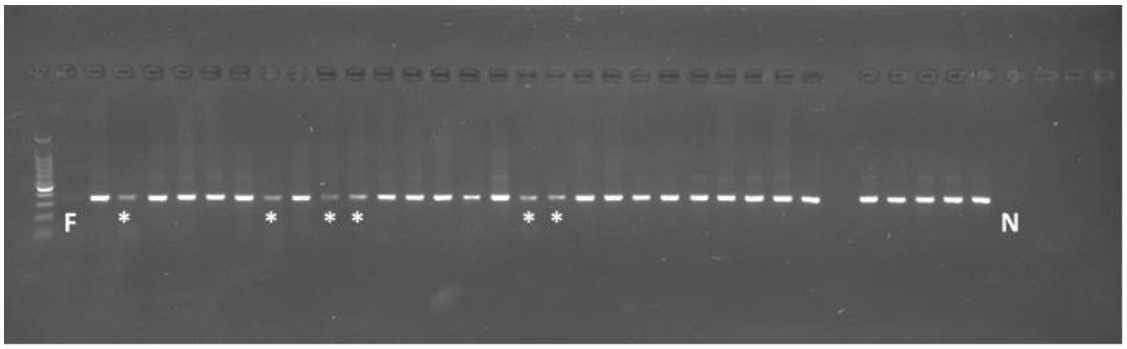



**Figure 4.1** Gel electrophoresis image depicting PCR results from a MID-tagged PCR, and the two different band strengths observed. Failed samples are indicated with an 'F' and were repeated using a different MID-tag combination. Bands labelled with a '*' indicate PCR products which were deemed 'faint' and therefore pooled with a higher volume. This figure displays results for DAB1, however results for DAB2 were equivalent. A 100 bp DNA ladder (Promega) was used, confirming fragment sizes to be ca. 350-400 bp.

The concentration of each PCR pool was quantified on a Qubit Fluorometer before being combined into a super pool. Concentrations, volumes and number of samples in each pool, and their subsequent consolidation into a super pool can be seen in Table 4.5. Tapestation results for the two final pools showed only low levels of primer dimer, so no more purification was needed prior to library preparation.

**Table 4.5** Results from the Qubit for each initial pool, and the subsequent volumes and concentrations added into the 'super pool' for each DAB locus. The final super pool volumes and concentrations are also included.

| Locus | PCR Pool | Number of samples | Initial pool volume (µL) | Post-SPRI volume (µL) | QuBit concentration (ng/µL) | Volume and concen-tration added into the 'super' pool | Final pool total volume and concentration |
|---|---|---|---|---|---|---|---|
| DAB1 | 1 | 72 | 219 | 72 | 52 | 6.58 µL<br>4.75 ng | 20.33 µL<br>45.6 ng/µL |
| DAB1 | 2 | 97 | 297 | 97 | 53 | 8.69 µL<br>4.75 ng | |
| DAB1 | 3 | 27 | 90 | 27 | 59 | 2.17 µL<br>4.75 ng | |
| DAB1 | 4 | 29 | 96 | 29 | 47.5 | 2.9 µL<br>4.75 ng | |
| DAB2 | 5 | 98 | 294 | 98 | 58.8 | 7.71 µL<br>4.63ng | 19.52 µL<br>51.9 ng/µL |
| DAB2 | 6 | 97 | 300 | 97 | 57.6 | 7.79 µL<br>4.63 ng | |
| DAB2 | 7 | 23 | 69 | 23 | 60 | 1.77 µL<br>4.63 ng | |
| DAB2 | 8 | 3 | 13 | 10* | 6.17 | 2.25 µL<br>4.63 ng | |

Prior to library preparation, the sample pools needed to be purified. This was achieved using the SPRIselect bead-based system (Beckman Coulter, USA). The SPRI bead clean up used followed the 'Left Side Size Selection' documented in the user guide, with a few modifications (Appendix 4.1). With a fragment size of approximately 270 bp for both lineages, a 1:1 ratio of SPRIselect volume to pool volume was chosen [according to the Agilent high sensitivity DNA chip electropherogram]. At the end of the SPRIselect protocol, each pool was eluted

with ultrapure water. The volume of ultrapure water used reflected the number of samples that were put into the pool after PCR. For example, in a pool where 27 samples had been added, 27 µL was used in the elution.

Once cleaned, the four pools for each locus could be combined into two, DAB-specific, equimolar 'super pools' for HTS library prep. The DNA concentration of each cleaned pool was measured using a Qubit Fluorometer (ThermoFisher Scientific, UK), and used to calculate (Appendix 4.2) the volume of each pool needed in the 'super pool'. The Qubit Fluorometer was again used on the two super pools, to measure the final concentrations before library preparation could begin. The final concentration of each pool going into the DAB1 super pool was 4.75 ng/µL, leading to a super pool volume and concentration of 20. 33 µL and 45.6 ng/µL, respectively. The super pool for DAB2 contained 6.87 ng/µL of the four pools, with a total volume of 30.85 µL and concentration of 55 ng/µL.

Following creation of a single pool for each DAB locus, both were run on a Tapestation, to ensure the concentrations reflected the correct fragment size, and that the integrity of samples was good. The Tapestation revealed low levels of primer dimer and confirmed that purification had been successful.

Library preparation was carried out using a Nextflex® Rapid DNA Sample Preparation kit (PerkinElmer®, USA), to prepare the pools for Illumina HTS sequencing. Size selection was not necessary, as all the pooled samples were of almost the same length and only the removal of low-molecular weight material was needed. As a result, the protocol for library preparation without size selection was followed throughout. Details of the library preparation protocol are shown in the Appendix for Chapter 4. Each DAB pool was processed separately, with the following adjustments made specific to the two DAB libraries needed:

i) Beginning with End-Repair and Adenylation (Step A1), the pools had to be mixed with water, NEXTflex buffer mix and NEXTflex adenylation enzyme, to a total volume of 50 µL. Both libraries contained 15 µL of End-Repair and Adenylation Buffer Mix, and 3 µL of End-Repair and Adenylation Enzyme Mix. To calculate the volume of DAB pool needed, 100 was divided by the Qubit concentration of the super pools. This volume was added to enzyme and buffer mix,

then topped up to 50 µL with water. For DAB1, the adenylation mix contained 2.19 µL of the super pool, and 29.81 µL of water. DAB2 contained 1.93 µL of the pool, and 30.07 µL water.

ii) In Adapter Ligation (Step B1), step 2, the manual suggested an adapter dilution of 1:8.3, according to an input amount of up to 100 ng DNA. To achieve this, 1 µL of adapter was diluted in 8.3 µL water. One adapter was chosen for each pool, to allow downstream demultiplexing of the two DAB lineages. To the adenylation mix from step A1, 2.5 µL of diluted adapter and 47.5 µL NEXTflex Ligase Enzyme mix was added, before incubation in a thermocycler.

iii) After incubation, the libraries were cleaned using AMPure XP beads (Beckman-Coulter, USA). The protocol suggested 50 µL beads were used, but here 80 µL was used. This 0.8x ratio of beads to library mix would ensure a size selection of approximately 300bp, according to the user manual. After the ethanol washes, 52 µL of resuspension buffer was used, instead of the suggested 22 µL These were the only adjustments made to the AMPure bead clean stage.

The final stage of library preparation, PCR Amplification (Step C1), was carried out according to the manufacturer's instructions, with no adjustments made.

Once the two DNA libraries for each DAB lineage had been generated, both were checked on a tapestation and Qubit once more, supplying the average library size in bp, and the concentration respectively. The final concentration of the DAB1 library was 9.88 ng/µL, and DAB2 was 16.3 ng/µL. Both pools were then diluted to 4 nm (Appendix 4.3) using 5 µL of each library, and 32.75 µL water for DAB1, and 58.8 µL for DAB2. The concentration of these libraries was then once more confirmed using a Qubit, and adjusted closer to 4 nmol if necessary, by adding more water or more undiluted library pool. After library preparation and dilution to 4 nmol, the final stages of the libraries were prepared for Illumina sequencing by the Genome Hub at Cardiff University. Once both libraries were diluted to 4 nmol, the two libraries were sequenced jointly, generating 2 x 250 bp paired end reads on an Illumina MiSeq sequencer (Illumina, USA).

### 4.3.4: Demultiplexing Illumina Data Using FastP and jMHC

Scripts used to demultiplex data can be found in Appendix 4.4, following bioinformatics scripts modified from Drake et al. (unpubl.). Files from Illumina were first unzipped into. fastq files. In turn each of these files were checked for truncation in both sequencing directions. The 'grep' command was used to extract reads into a .txt file, with 8 bases preceding the primer sequence. This would account for reads that included the MID-tags. To confirm this had worked, the 'less' command was used to list a selection of reads. These reads were aligned in Geneious (www.geneious.com) using a sequence (GenBank accession number KJ162508) from Burri et al. (2014) for DAB1, and a sequence from Dearborn et al. (2014) (taken from their supplementary information) for DAB2. In addition to the reads that included the MID-tags, the number of total reads, and reads without an 8-bp addition were also calculated. These values were then used to calculate the percentage of reads that had been truncated and would therefore be lost from the dataset due to missing MID-tags or primer sequence.

Next, sequences were trimmed, aligned and the quality was checked using 'fastp' (Chen et al. 2018). FastP is an all-in-one tool that pre-processes fastq files for further analysis. Using a shell script, the two Illumina files for each DAB lineage were filtered by a length limit of 250 bp. The forward and reverse reads for each DAB lineage that passed the length filter were then merged and compiled into one fastq file. The fastq file of reads for each DAB lineage was then quality-filtered with a Q-score limit of 30. The reads that maintained a Q-score of 30 and above were then retained and converted into a fasta file for each DAB lineage. The final reads were then calculated for each DAB lineage.

Next, the software jMHC (Stuglik et al. 2011) was used to genotype the reads into distinct alleles. Designed specifically for analysing high-throughput sequencing data from MHC genes, the program uses fasta files to extract variants and produces a table listing their frequencies in the dataset.

First, a 'tags' file was created for use in jMHC, including a list of all sample names present in the dataset, along with their corresponding MID-tag combination. In the first column, the concatenated MID-tag combinations were placed, in the next column the .fasta file of sequences was included, whilst the final column had the sample names for each MID-tag combination.

As well as true samples, negatives were given sample names of 'ENEG' for extraction negatives, and 'PNEG' for PCR negatives. Negatives were included to detect any contamination. Whilst they did not have DNA added, the presence of reads for their MID-tag combinations was used to monitor for contamination. In addition to these negative controls, 'NA' samples were included in the tags file, corresponding to tag combinations that were not chosen for any PCR reactions, but which could hypothetically arise due to tag-jumping. For example, L1F13 was used with R1-12 in PCR. Primer R13-24 was not combined with L1F13 but had been used in 12 reactions together with primers L1F1 to L1F12. In the tags file, entries labelled 'NA' were therefore created for the MID-tag combination of L1F13 and R13-24. As this combination was never used in PCR, sequences containing both these tags could be regarded as chimeric. Such tag jumping is well known (Schnell et al. 2015), but should be accounted for in data analysis, so that false reads do not inflate allele numbers.

In jMHC a separate database was created for each DAB lineage. The relevant tags file along with sequences from FastP as well as the primer sequences were inputted, specifying tag length as 8. Following Stuglik et al. (2011), '2-sided tags', 'Forward' and 'Reverse' were selected, ensuring that jMHC would only extract reads from the .fasta file that had the correct primer sequence, MID-tag length, looking for these in both the forward and reverse direction. When extracting reads to a tabular format, jMHC clusters reads into variants or alleles. During the extraction, samples with 'NO_TAG' were discarded, and only variants with ≥3 reads total were included, considering variants with <3 reads total were automatically considered artefacts (Rekdal et al. 2018). The resulting jMHC output provided a list of potential alleles and their associated individual reads, with tags and primers trimmed away.

### 4.3.6: Data Filtering and Allele Calling Pipeline

The jMHC output included all potential alleles, but presumably also any PCR artifacts (primer dimers, contamination, chimeras and hairpins), inflating the number of alleles and requiring additional filtering. To this end, the following steps were taken to reduce the dataset to only 'true' alleles (summarised in Figure 4.1):

First, the unique sequences were sorted by length. From previous Sanger sequencing data, the target size for DAB1 was expected to be 273 or 276 bp, depending on presence or absence of a 3-bp indel. For DAB2, the target size was 274 or 271 bp, also depending on a 3-bp indel. Sequences that did not adhere to these lengths were removed. Initially it was considered that sequences with indels in multiples of 3 would be considered (i.e., whole codons), as per Sommer et al (2013). However, due to the huge number of sequences obtained, a stricter length filter, described here, was used. Sequences not falling within this strict filter were briefly screened for high read counts, however it is possible that true sequences with indels may have been missed out. Next, as per Rekdal et al. (2018) total reads for each sequence were counted, and any sequences with <500 reads across all samples were removed, as this was considered too low to be considered an allele rather than an artifact (Lighten et al. 2014). To account for reads assigned to contamination, the maximum read count among the negative samples for each allele was determined, and this value was then subtracted from all other read counts for the same allele, serving to remove reads from each sample that could be attributed to contamination (an approach commonly used in dietary metabarcoding studies, e.g., Drake et al. unpubl).

After this removal of contamination, low-read artifacts and sequences with non-target lengths, the total number of reads across all unique sequences (putative alleles) was determined, calculated as the percentage of the total read count for this individual. Burri et al. (2014) estimated that *H. castro* and *H. monteiroi* possess a minimum of 3 loci and 5 alleles each at DAB1 and DAB2. Hence, to account for the presence of more than 3 locus copies of each DAB lineage, we applied cut-offs to remove alleles with read counts too below expected thresholds: e.g., for four DAB1 loci, a maximum of eight distinct alleles with on average 12.5% read proportion would be expected. Conservatively, a read percentage cut-off of 10% was therefore applied, representing the lowest read count percentage expected for 10 alleles (5 loci). This cut-off represented a compromise between allowing for potentially larger numbers of loci, against inclusion of sequences as alleles that from Mendelian inheritance checks in family trios (mother, father, chick) predominantly appeared to be artefacts. Any sequences that attributed to <10% of the total reads in a sample were thereby removed for that sample.

Following this filtering, the total reads per sequence were re-calculated, and sequences with <500 total reads were discarded.

### 4.3.7: Checking Allele Inheritance, Copy Number and Negative Reads

#### 4.3.7.1: Assessing Copy Number of MHC Alleles

Ideally, read share percentages from the previous calculations could be used to inform about copy numbers of MHC loci. For example, if 4 copies of 2 loci were observed, alleles could be present in a 25%, 50%, 75% or 100% share. Should an individual have 3 copies of one allele, and 1 copy of another, the read share percentages for the two alleles should be 75% and 25% respectively. To visualise the read share percentages of alleles across the dataset, histograms were created using R and Rstudio using the 'hist' command, setting 'breaks=50' to allow more precise visualisation of read count proportions. These histograms were inspected to see whether the obtained read count proportions fit the expected distributions for given locus/allele numbers. The full R script used can be found in the Appendix for Chapter 4

#### 4.3.7.2: Verifying Mendelian Inheritance

For each storm petrel family where chick samples were taken (n=24 for DAB1, n=23 for DAB2), Mendelian inheritance of alleles was checked. If observed, Mendelian inheritance should show all chick alleles as inherited from their parents, and an equal share of parent alleles represented in the offspring. The read counts and alleles for individuals of each family were extracted from the final, filtered dataset. Read percentages of alleles in each family were calculated, and the alleles present in each chick were checked and compared to the alleles present in the parents. The read counts, percentage shares and subsequent inheritance checks were carried out manually in Microsoft Excel.

Each family was assessed on criteria compatible with Mendelian inheritance. Families were categorised based on these criteria, and their overall conformation to Mendelian inheritance was determined. The criteria used were:

i) All alleles found in the chick must be represented in the parents.
ii) Chicks inherit at least one allele from each parent (larger numbers for higher numbers of loci).

Mendelian inheritance would imply equal inheritance of alleles from both parents. Copy number could however not be determined in this, therefore precisely equal inheritance was not considered as part of the above criteria. Hence, inference of Mendelian inheritance could only be made tentatively, precluding definitive assessments.

If a chick appeared to have inherited alleles from both parents, this was categorised as 'conforming' to Mendelian inheritance. In some cases, chicks shared alleles with both parents, but also had extra alleles that did not present in the parents - here, families were categorised as 'inconclusive'. In cases where only one parent shared alleles with the chick, this was categorised as a 'undetermined'. Table 4.6 displays each category name and the criteria for sorting. On the rare occasions where chicks had no alleles shared with either parent, these were categorised as 'no inheritance'.

**Table 4.6** Categories used to assign families and check their conformation to Mendelian inheritance. Criteria for each category is explained.

| | |
|---|---|
| **Category 1: Conforming to Mendelian inheritance** | • All alleles possessed by the offspring are found in the parents.<br>• Both parents have contributed at least one allele to the offspring. |
| **Category 2: Unexpected alleles in offspring** | • The offspring has alleles from both parents, but also has additional alleles **not** seen in the parents. |
| **Category 3: Single-parent match** | • The offspring has inherited alleles from one parent but not the other. |
| **Category 4: No shared alleles.** | • No alleles present in the offspring are seen in either parent. |

All families were assessed by these criteria, and conformation to Mendelian inheritance was evaluated.

#### 4.3.7.3: Confirming Presence / Absence of Indels

Dearborn et al (2015) recorded a 3-bp indel (one codon position) present in both DAB lineages in Leach's storm petrel. To check whether the 3-bp indel in *H. castro* and *H. monteiroi* included the same codon position, allele sequences were aligned with those from Dearborn et al. (2015) (KP090142 for DAB1 and KP090151 for DAB2), using Geneious. The alignment confirmed that for some alleles (n=13 for DAB1; n=6 for DAB2), the same 3-bp region contains a polymorphic indel all three petrel species. This is consistent with a single origin indel that may predate the radiation of all 3 species. As such, it is possible that this 3 bp indel could have evolved in response to an environmental pathogen or undergone deletion if not functional. However, Dearborn et al (2015) found that the indel in *H. leucorhous* did not occur at a peptide-binding site, which suggests it may not be related to pathogen response.

#### 4.3.7.4: Assessing PCR Amplification Outcomes Using BLAST

To investigate the phylogenetic affiliation of the final set of alleles obtained for *H. castro* and *H. monteiroi*, all alleles were BLASTed using a shell script against (a) the nucleotide database available on Genbank (Benson et al. 2013), and for putative DAB2 alleles also (b) a bespoke database that contained both Sanger sequencing data from DAB2 (**Chapter 3**), plus previously published MHC data for related species (Burri et al. 2014b; Dearborn et al. 2014). The latter was done because no *H. castro* or *H. monteiroi* DAB2 sequences were represented in GenBank's nucleotide database.

In addition, reads contained in negative controls were also blasted against the NCBI database. This confirmed that reads in the negatives were the result of low levels of contamination with *H. castro* and *H. monteiroi* DAB sequences, most likely at the PCR stage, with no evidence of primers cross-reacting with template DNA from other species.

#### 4.3.7.5: Read Counts in Negative Controls and Storm Petrel Samples

To compare the number of reads in samples versus those found in negatives or 'NAs', boxplots were created in R, using unfiltered read count data taken from the initial output from jMHC. As explained above, non-used MID-tag combinations were assigned as 'NA' groups, allowing investigation of tag-jumping (Schnell et al. 2015). Reads for DNA samples were simply indicated as

'Samples'. PCR and Extraction negatives that were pooled were each given their own category. 'Failed samples' represented MID-tag combinations that were used in PCR but did not produce a band when viewed on agarose gel. 'Unpooled MID-tags' were combinations that were used on negatives, but the PCR product was not included in the pool. 'Unused MID-tags' represents combinations that were not used in PCR but could exist if tag-jumping were present. These were included so that tag-jumping could be detected. The boxplots for each DAB lineage were produced using the 'boxplot' function in R. Scripts are shown in the appendix.

### 4.3.8: Phylogenetic analysis of alleles

Alignments of the final alleles were generated in Geneious, along with *H. castro* and *H. monteiroi* sequences from Burri et al. (2014), Leach's storm petrel sequences from Dearborn et al (2015) (representing data from both DAB1 and DAB2 in a closely related species from the same genus). All sequences (n=118) were trimmed to target Exon 2 (ranging from 270 to 291 bp). Using the phylogenetic tree created above as a reference point, two egret sequences (Wang et al. 2013) were included to act as an outgroup to root the phylogeny.

The alignment was analysed by ModelFinder in IQTree (Nguyen et al. 2015) to find the most suitable phylogenetic model for the data. After assessing 280 models, a TVMe+R3 model was chosen according to BIC criterion (a transversion model with equal base frequencies, and a FreeRate category of 3). The model was run with 100 bootstraps. The resulting tree was edited using FigTree. Genbank sequences used in the alignment are listed in Table 4.7.

**Table 4.7** Genbank sequences used in the alignment to check the phylogeny of DAB1 and DAB2 alleles. In addition to these sequences, all filtered, true alleles generated by jMHC were used. These sequences provided a reference for checking the divergence of alleles

| Species | DAB lineage | GenBank Accession Code | Reference |
|---|---|---|---|
| **Leach's storm petrel (*H. leucorhous)*** | DAB1 | KP090143 | (Dearborn et al. 2014) |
| | | KP090144 | |
| | | KP090148 | |
| | | KP090146 | |
| | | KP090150 | |
| | DAB2 | KP090158 | |
| | | KP090154 | |
| | | KP090160 | |
| | | KP090156 | |
| | | KP090162 | |
| **Band-rumped storm petrel (*H. castro*)** | Unspecified | KJ162507 | (Burri et al. 2014a) |
| | | KJ162508 | |
| | | KJ162509 | |
| | | KJ162510 | |
| | | KJ162511 | |
| **Monteiro's storm petrel (*H. monteiroi*)** | | KJ162513 | |
| | | KJ162514 | |
| | | KJ162515 | |
| | | KJ162516 | |
| | | KJ162517 | |
| **Chinese egret (*E. eulophotes*)** | DAB1 | KC282848 | (Wang et al. 2013) |
| | DAB2 | KC282849 | |

Separate haplotype networks of DAB1 and DAB2 alleles were created using PopArt using the median joining algorithm. Sequence diversity statistics were calculated using ARLEQUIN 3.5.2.2 (Excoffier and Lischer 2010), including haplotype diversity, nucleotide diversity, Fu's $F_S$ (Fu 1997) and Tajima's D (Tajima 1989). Haplotype data files were created for each group using DAB allele sequences, including each allele as many times as the number of individuals carrying it. This was done in lieu of having actual allele counts, since the number

of copies of each allele within individuals could not be determined based on the present data.

### 4.3.9: Assessment of genetic differentiation among species for DAB allele profiles

To investigate differentiation among individuals for DAB allelic profiles, a presence (1) / absence (0) matrix of all DAB1 and DAB2 alleles in all individuals was generated in R. Jaccard distances (Jaccard 1912), suitable for presence-absence data, were calculated for all pairwise comparisons among individuals, using the vegdist function from the vegan package (Oksanen et al. 2015). The resulting distance matrix was used to reconstruct a neighbour-joining tree in MEGA-X (Kumar et al. 2018) and to conduct non-metric dimensional scaling (NMDS) analysis in R, using the metaMDS function from the vegan package (script is included in the appendices).

## 4.4: Results

### 4.4.1: MID-tag PCR Results and Pooling

For both DAB lineages, a total of 210 samples, seven PCR negatives and 23 extraction negatives were PCR amplified using MID-tag primers, and subsequently Illumina sequenced. For DAB1, 27 samples initially failed to amplify, and all were successfully repeated with a different MID-tag combination. For DAB2, eight samples failed to amplify, and only two samples were successful on repeating with a different MID-tag combination. These six samples equated to four adults and two chicks. A loss of these six samples for DAB2 resulted in five pairings being removed from the dataset.

### 4.4.2: Bioinformatics

#### 4.4.2.1: Truncation and FastP

Only a small proportion of reads was truncated, ranging from ca. 4% for DAB1-read 1 to almost 7% for DAB2 read 2 (Table 4.8), implying that most reads contained the whole target fragment, as well as the primers and MID-tags. After FastP processing, DAB1 had 6,354,407 merged reads, and DAB2 7,085,167 merged reads.

**Table 4.8** Read counts and truncation percentage calculations for initial Illumina data. The truncation percentage was calculated by dividing reads without an 8-bp MID tag by the total reads present. The number of reads with an 8-bp MID tag would be considered the true number of successful reads.

| DAB, Illumina File and Direction | Total Reads | Reads with 8-bp MID Tag | Reads without 8-bp MID Tag | Truncation (%) |
|---|---|---|---|---|
| DAB1; Illumina Read 1; Forward | 3,100,318 | 2,974,024 | 126,294 | 4.07 % |
| DAB1; Illumina Read 1; Reverse | 3,046,057 | 2,900,300 | 145,757 | 4.78 % |
| DAB1; Illumina Read 2; Forward | 2,917,607 | 2,737,601 | 180,006 | 6.16 % |
| DAB1; Illumina Read 2; Reverse | 3,112,788 | 2,935,088 | 177,700 | 5.70 % |
| DAB2; Illumina Read 1; Forward | 3,512,198 | 3,304,890 | 209,691 | 5.97 % |
| DAB2; Illumina Read 1; Reverse | 3,364,408 | 3,186,185 | 178,223 | 5.29 % |
| DAB2; Illumina Read 2; Forward | 3,245,852 | 3,046,639 | 199,213 | 6.13 % |
| DAB2; Illumina Read 2; Reverse | 3,524,731 | 3,283,053 | 241,678 | 6.85 % |

#### 4.4.2.2: jMHC Output and Filtering.

An overview of the jMHC and manual filtering process, and the remaining number of reads and sequences after each stage, can be seen in Figure 4.2. For DAB1 and DAB2, 6,354,407 and 7,085,167 reads were imported into jMHC, respectively. After assessing the reads for presence of primers and tags, 5,132,061 DAB1 reads, and 5,586,060 DAB2 reads remained. These reads were then clustered into alleles, discarding any alleles with <3 total reads. After assigning reads to alleles, DAB1 presented with 51,499 unique sequences, whilst DAB2 presented with 38,381 unique sequences.

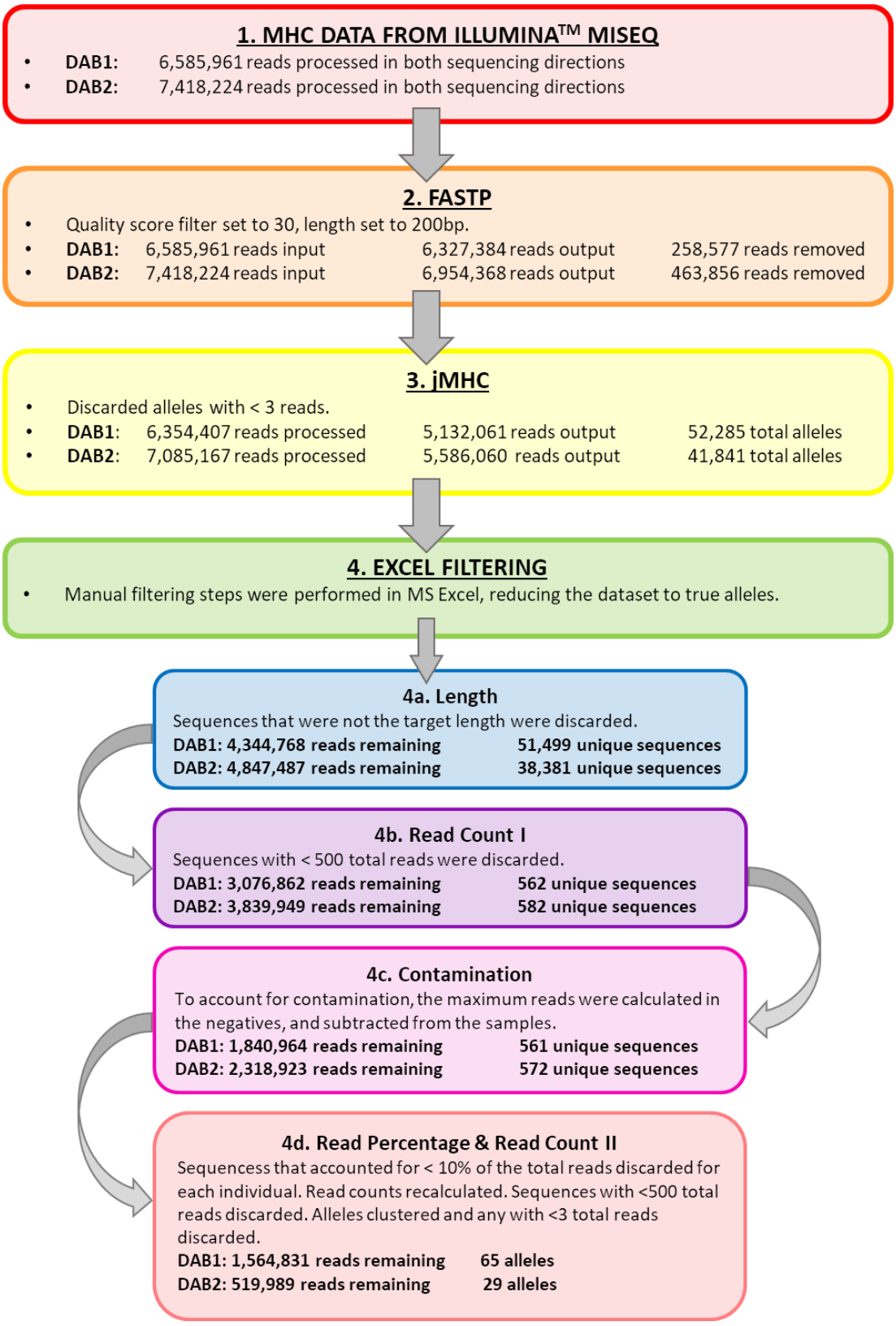



**Figure 4.2** Flow chart representing the filtering steps taken to reduce the dataset to true alleles. Flow chart representing the filtering steps taken to reduce the dataset to true alleles. The total number of Individual successfully sequenced and genotyped for DAB1/DAB2 are: H. castro n=111/110; H. monteiroi n= 99/91, respectively

Manual filtering in Excel only led to a small number of samples being removed. Where samples initially had a low number of reads, contamination filtering reduced those reads to 0, removing 3 samples from the dataset. Following these steps, 65 alleles were found for DAB1, totalling 1,564,831 reads across the 204 samples, with 6 samples rejected due to low overall read count. Of these DAB1 alleles, 23 alleles were unique to *H. castro,* 35 were unique to *H. monteiroi*, and 7 alleles were shared between the two species. A total of 27 alleles were found for DAB2, containing 519,989 reads, across 201 samples, with a total removal of 3 samples due to low read count. Of these 27 alleles, 14 were unique to *H. castro*, 11 were unique to *H. monteiroi,* and 2 alleles were shared between the two species. After filtering, each retained allele in the dataset was supported by a minimum of 500 remaining reads. All final retained alleles for each DAB lineage can be seen in Table 4.9 and 4.10.

**Table 4.9** DAB1 alleles, including length in bp and the total reads across the dataset of 204 samples

| Allele | Length | Total Reads | | Allele | Length | Total Reads |
|---|---|---|---|---|---|---|
| DAB1*01 | 276 | 183755 | | DAB1*35 | 273 | 11207 |
| DAB1*02 | 276 | 183445 | | DAB1*36 | 273 | 11065 |
| DAB1*03 | 276 | 175896 | | DAB1*37 | 273 | 9412 |
| DAB1*04 | 276 | 81115 | | DAB1*38 | 276 | 9027 |
| DAB1*05 | 273 | 65474 | | DAB1*39 | 276 | 8459 |
| DAB1*06 | 276 | 59233 | | DAB1*40 | 276 | 7834 |
| DAB1*07 | 276 | 57951 | | DAB1*41 | 276 | 7475 |
| DAB1*08 | 276 | 55637 | | DAB1*42 | 276 | 6639 |
| DAB1*09 | 276 | 53416 | | DAB1*43 | 276 | 5775 |
| DAB1*10 | 276 | 51150 | | DAB1*44 | 276 | 5645 |
| DAB1*11 | 276 | 34482 | | DAB1*45 | 276 | 5096 |
| DAB1*12 | 276 | 33824 | | DAB1*46 | 276 | 4694 |
| DAB1*13 | 276 | 30343 | | DAB1*44 | 276 | 5645 |
| DAB1*14 | 276 | 30150 | | DAB1*45 | 276 | 5096 |
| DAB1*15 | 276 | 26205 | | DAB1*46 | 276 | 4694 |
| DAB1*16 | 276 | 25082 | | DAB1*47 | 276 | 4298 |
| DAB1*17 | 276 | 23374 | | DAB1*48 | 273 | 4133 |
| DAB1*18 | 276 | 22216 | | DAB1*49 | 276 | 3814 |
| DAB1*19 | 276 | 18975 | | DAB1*50 | 276 | 3761 |
| DAB1*20 | 276 | 16895 | | DAB1*51 | 276 | 3728 |
| DAB1*21 | 276 | 16518 | | DAB1*52 | 276 | 3701 |
| DAB1*22 | 276 | 15746 | | DAB1*53 | 273 | 3698 |
| DAB1*23 | 273 | 15475 | | DAB1*54 | 273 | 3672 |
| DAB1*24 | 276 | 15370 | | DAB1*55 | 276 | 3403 |
| DAB1*25 | 276 | 14375 | | DAB1*56 | 276 | 3171 |
| DAB1*26 | 276 | 14271 | | DAB1*57 | 273 | 2797 |
| DAB1*27 | 276 | 13876 | | DAB1*58 | 276 | 2233 |
| DAB1*28 | 276 | 13703 | | DAB1*59 | 273 | 2221 |
| DAB1*29 | 276 | 13013 | | DAB1*60 | 273 | 2203 |
| DAB1*30 | 276 | 12221 | | DAB1*61 | 276 | 1764 |
| DAB1*31 | 276 | 12218 | | DAB1*62 | 273 | 1347 |
| DAB1*32 | 276 | 11987 | | DAB1*63 | 276 | 1093 |
| DAB1*33 | 276 | 11373 | | DAB1*64 | 276 | 879 |
| DAB1*34 | 273 | 11220 | | DAB1*65 | 276 | 602 |

**Table 4.10** DAB2 alleles, including length in bp and the total reads detected across 201 samples

| DAB Name | Length | Total Reads |
|---|---|---|
| DAB2*01 | 274 | 485118 |
| DAB2*02 | 274 | 211086 |
| DAB2*03 | 274 | 174230 |
| DAB2*04 | 274 | 110117 |
| DAB2*05 | 271 | 108364 |
| DAB2*06 | 271 | 96732 |
| DAB2*07 | 274 | 68821 |
| DAB2*08 | 274 | 75967 |
| DAB2*09 | 274 | 64997 |
| DAB2*10 | 274 | 62553 |
| DAB2*11 | 274 | 53967 |
| DAB2*12 | 274 | 60585 |
| DAB2*13 | 274 | 50627 |
| DAB2*14 | 274 | 49718 |
| DAB2*15 | 274 | 41918 |
| DAB2*16 | 274 | 39094 |
| DAB2*17 | 271 | 30096 |
| DAB2*18 | 274 | 35613 |
| DAB2*19 | 271 | 24710 |
| DAB2*20 | 271 | 23793 |
| DAB2*21 | 274 | 17798 |
| DAB2*22 | 274 | 12952 |
| DAB2*23 | 274 | 7097 |
| DAB2*24 | 271 | 1128 |
| DAB2*25 | 274 | 3668 |
| DAB2*26 | 274 | 1554 |
| DAB2*27 | 274 | 2596 |

For DAB1, the minimum number of reads per sample was 1,326, the maximum 26,375 and the median number of reads across samples was 6,252. For DAB2, the minimum number of reads was 1,026, the maximum 18,449, with median reads per sample of 9,557. The number of reads in negative controls or 'NA' MID-tag combinations was low in comparison to those obtained for storm petrel samples. In both DAB lineages, the average number of reads in negative controls contributed to 3% of the total read count. Pre-filtering, the total number of DAB1 reads attributed to contamination was 63,761, with a maximum of 11,843 and a minimum of 390. The average number of reads per sample was 4,024. For DAB2, reads attributed to contamination was 61,788, with a maximum of 13,019 and a minimum of 412 reads. The average number of reads contained in negative

samples was 3,861. Tag-jumping was detected, as evidenced by reads being attributed to 'NA' samples. In DAB1, the total number of 'NA' reads was 62,681, with a maximum of 7,502 and a minimum of 504 reads. On average, NA samples contained 7,502 reads. For DAB2, the total reads in NAs was 79,411, the maximum 12,178, the minimum 149 and average 1,804.

The boxplots for read counts in both DAB1 (Figure 4.3) and DAB2 (Figure 4.4) data showed that read depths present in sequenced samples were considerably higher than read depths in negative controls or 'NA' samples. For both DAB lineages, a small number of reads was found in PCR and extraction negatives - indicating presence of low levels of contamination, either at the DNA extraction stage (in the case of extraction negatives), or the PCR stage (for extraction and PCR negatives). Reads attributed to failed samples or unpooled MID-tags were consistent with some tag-jumping, or small amounts of failed-sample PCR mix contaminating other wells in a PCR plate. Reads from the unused MID-tags indicate that a small amount of tag-jumping has occurred. These MID-tag combinations were never combined in a PCR reaction and are therefore unlikely to be the result of contamination.

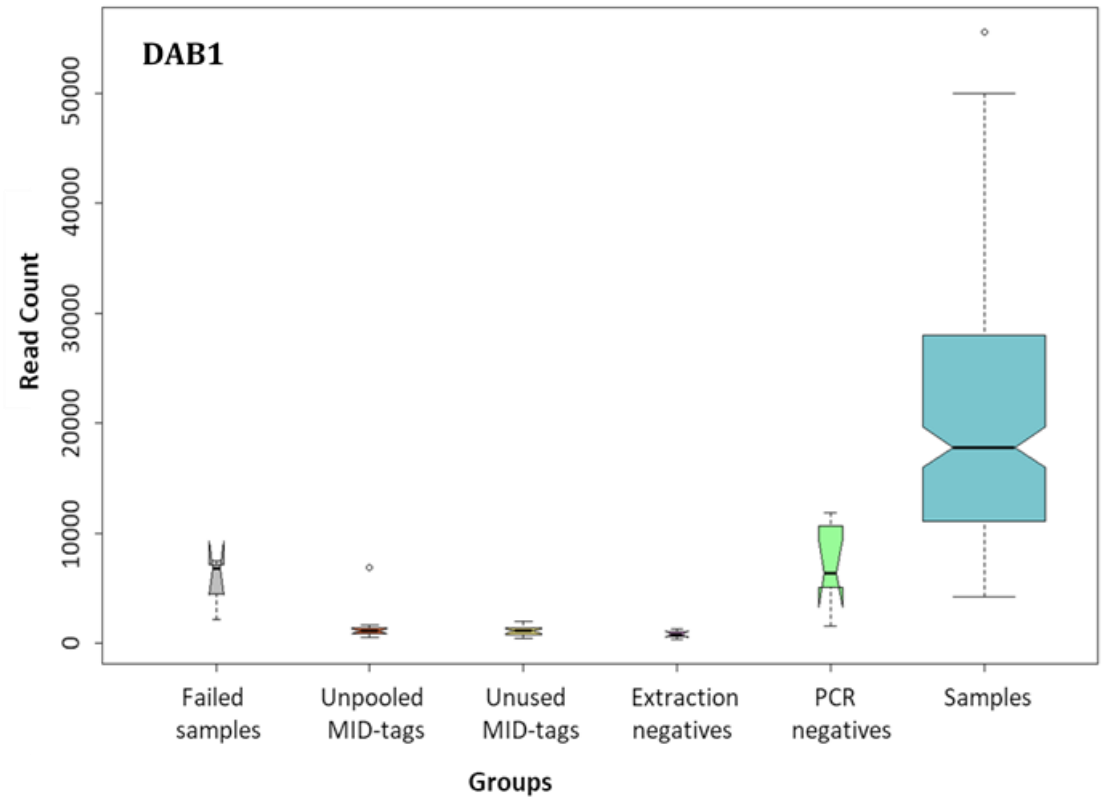



**Figure 4.3** Boxplot representing the total reads per each MID-tag in the DAB1 dataset, categorised into types. The boxplot indicates that overall, total reads in true samples are much higher than those found in other categories. The reads in the other categories indicates low levels of tag-jumping and contamination.

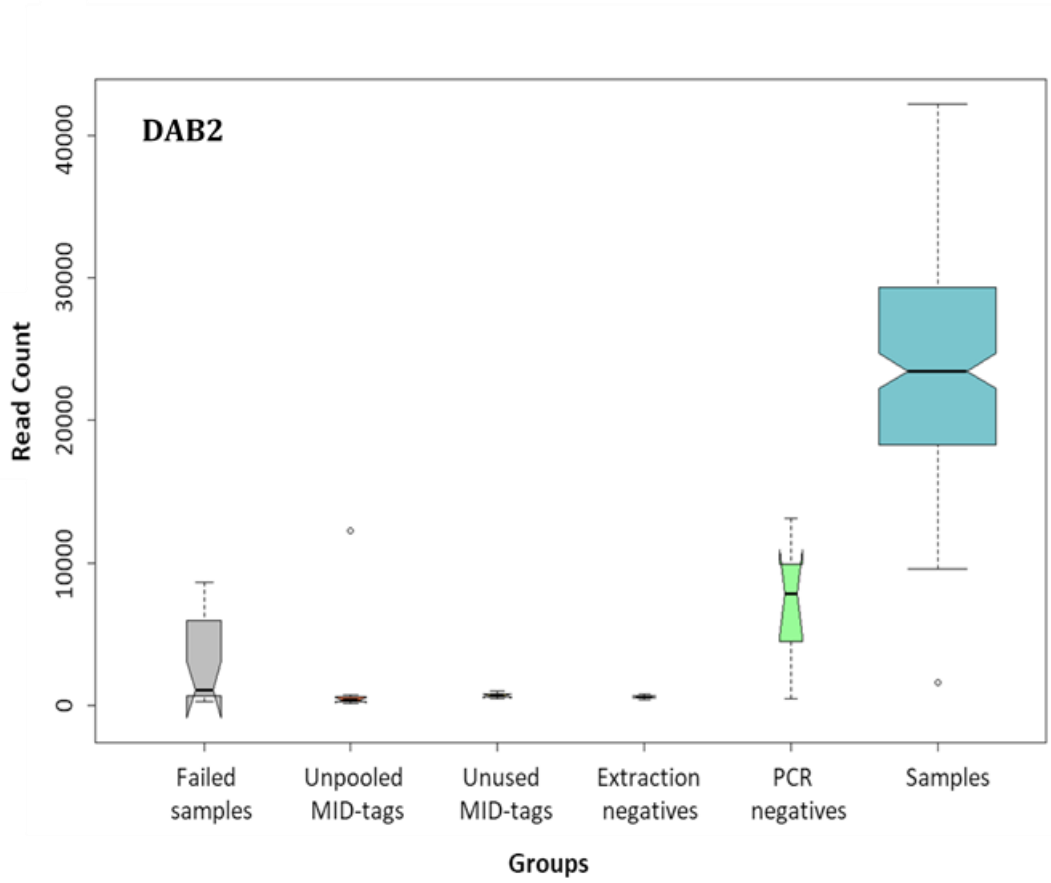



**Figure 4.4** Boxplot representing the total reads per each MID-tag in the DAB2 dataset, categorised into types. The meanings for each category are specified in the main text. The boxplot indicates that overall, total reads in true samples are much higher than those found in other categories. The reads in the other categories indicates low levels of tag-jumping and contamination

## 4.4.3: Assessing Copy Number and Mendelian Inheritance

### 4.4.3.1: Evaluation of the Potential to Use Read Count Percentage as a Proxy for Copy Number

For each sample, calculating the read count share of each allele could ideally reveal copy numbers of MHC alleles and loci – assuming allelic frequency in the individual's genome would be tightly correlated with Illumina sequencing read counts. Anticipating at least 3 loci (Burri et al. 2014) could produce up to 6 allelic (autosomal) copies in diploid organisms such as storm petrels. Three copies of a diploid locus could encompass 1-6 different alleles within one individual bird. Examples of the expected read counts for 3 copies of loci can be seen in Table 4.11.

**Table 4.11** Representative read shares and copy numbers of alleles, if 6 copies were present. This table demonstrates how read share percentages could hypothetically be used to determine copy number in the two storm petrel species. Alleles have been given arbitrary labels of A-E. If an individual has 5 copies of allele A, and one of allele B, the read share percentages should equal 83% and 17%, respectively. If read count was indeed consistent with copy number, the histograms shown in Figure 4.5 would be expected to fit these specified, discrete percentage categories, rather the random, non-discrete pattern that was observed

| **Alleles Present** | **Percentage Shares** |
|---|---|
| A A A A A A | 100% - A |
| A A A A A B | 83% - A 17% - B |
| A A A A B B | 67% - A 33% - B |
| A A A A B C | 67% - A 17% - B 17% - C |
| A A A B B B | 50% - A 50% - B |
| A A A B B C | 50% - A 33% - B 17% - C |
| A A A B C D | 50% - A 17% - B 17% - C 17% - D |
| A A B B C C | 33% - A 33% - B 33% - C |
| A A B B C D | 33% - A 33% - B 17% - C 17% - D |
| A A B C D E | 33% - A 17% - B 17% - C 17% - D 17% - E |
| A B C D E F | 17% - A 17% - B 17% - C 17% - D 17% - E 17% - F |

If read count proportions indeed followed these percentages, this evaluation approach could reveal the MHC copy number in these populations. However, as shown in Figure 4.5, the read count percentage share of alleles for both DAB lineages varied continuously - each bar on the histogram represents a 2%

change in read share proportions, showing that these values do not fit the distinct categories representative of 3 loci in Table 4.8. Hence, read count proportions did not seem to fall into discrete categories, precluding assessment of locus copy number with this approach. From the histograms it is also clear that DAB2 contained many homozygotes, represented by a large proportion of single alleles accounting for 100% of all reads of a given sample.

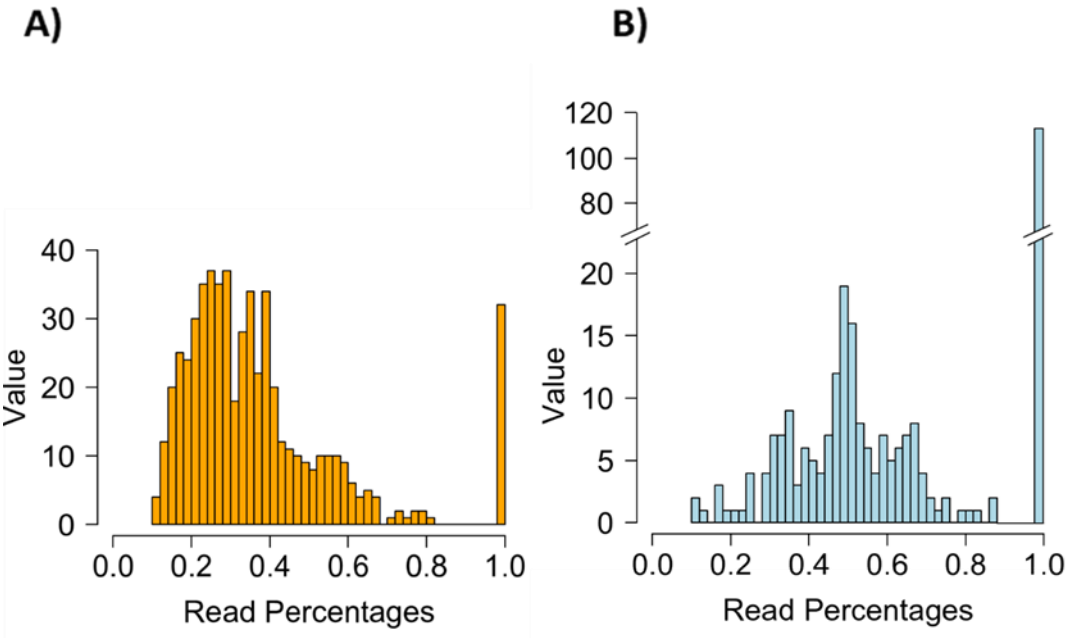



**Figure 4.5** Histograms for (A) DAB1 and (B) DAB2 of read share percentages of alleles within individuals, across the whole dataset. Read share percentage was calculated by dividing the read counts of each allele by the total reads for each sample. The histogram for DAB2 has a truncated y axis to better display the lower frequency read percentages.

#### 4.4.3.2: Mendelian Inheritance

Families were assigned to four different categories, based on how well they conformed to Mendelian inheritance. Full details for each family can be seen in the appendix, whilst Table 4.12 displays the overall number of families that were assigned to each category.

For DAB1, 12 families from each species were assessed, totalling 24 families overall. One family (FC003) was removed after both adults were revealed to be female. For DAB2, 12 families for *H. castro* and 11 families for *H. monteiroi* were initially checked. Two families (FC003 and FM005a) were not considered for Mendelian inheritance checks – as mentioned for DAB1, both presumed parental

samples in FC003 were found to correspond to females. The PCR for the father in FM005a failed, precluding further assessments of inheritance.

**Table 4.12** Numbers of families assigned to each category, and their conformation to Mendelian inheritance. For DAB1, one family was removed from the dataset; in DAB2, two families were removed from the dataset, after filtering led to sample loss, or sexing revealed both previously presumed biological parents to be female. For Category 3, which parent mismatched is noted for the families within this category.

| | |
|---|---|
| **Category 1: Consistent with Mendelian inheritance**<br>- *All alleles possessed by the chick are present in the parents.*<br>- *The chick has inherited alleles from **both** parents.* | **DAB1 –** 18 families<br>**DAB2 –** 13 families |
| **Category 2: Unexpected alleles in offspring**<br>- *The chick has alleles from both parents, but also has additional alleles not seen in the parents.* | **DAB1 –** 1 family<br>**DAB2 –** 0 families |
| **Category 3: One parent match**<br>- *The chick has inherited alleles from one parent but not the other.* | **DAB1 –** 4 families* (3 fathers and 1 mother mismatched)<br>**DAB2 –** 7 families* (4 fathers and 1 mother mismatched) |
| **Category 4: No shared alleles**<br>- *The alleles present in the chick are not seen in either parent.* | **No Families for either DAB lineage** |

* Only one family presented a one parent match for both DAB1 and DAB2, FM003.

For a locus showing Mendelian inheritance, one would predict that the alleles present in offspring have been inherited from both parents in equal shares. If the copy number of MHC alleles for these two species were known, read share percentages could be used to estimate whether alleles are inherited equally. As shown in the histograms, however, the read share percentages for alleles did not fall into discrete categories and did not allow for an estimation of copy number. Without copy number, Mendelian inheritance could not be assessed with complete rigour (see an example in Table 4.13). Without considering equal inheritance, for DAB1 and DAB2, allele sharing within 18 and 13 families

respectively was at least *consistent* with Mendelian inheritance. In these cases, all chick alleles were represented in the parents, and chicks had alleles from both parents.

**Table 4.13** OC128 is the mother (homozygote), and OC196 is the father. OC184 – the chick – has received copies from each parent, with no unexpected, extra alleles. Read shares are close to equal, but without copy number cannot be used as a proxy for estimating equal inheritance

| FC090 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| OC128 | | | OC196 | | | OC184 | | |
| seq45537 | 3352 | 100% | seq45537 | 1356 | 68% | seq45537 | 933 | 59% |
| | | | seq27101 | 647 | 32% | seq27101 | 638 | 41% |
| | | | | | | | | |

In families where Mendelian inheritance appeared doubtful, chicks had apparently not received alleles from both parents, or, the chicks carried extra alleles not seen in either parent.

For DAB1, four families were found where only one parent matched the alleles present in the chick. For DAB2, seven families displayed this single-parent match. In one DAB1 family a chick displayed an additional allele that was absent from both parents. This allele showed 13 apparent substitutions when compared to the closest parent allele and was common in the whole data set (found in 51 individuals), suggesting sequencing error or mutation is unlikely. It appears mismatches are most common between presumed fathers and the chicks, with 3 out of 4 one-parent mismatches occurring between chick and the male partner in DAB1, and 4 out of 7 chick-male mismatches occurring for DAB2. Without more in-depth parentage data such as microsatellites, and the ability to conclusively confirm actual parentage, Mendelian inheritance in these families could not be reliably assessed.

Knowing the copy number of MHC alleles in these two species would allow for a more robust confirmation of Mendelian Inheritance. As copy number could not be confirmed, estimations of Mendelian inheritance are only tentative and are not truly reliable.

### 4.4.4: Phylogenetic Analysis of Alleles and Diversity Statistics

The tree produced in IQTree (Figure 4.6) demonstrated that divergence of DAB lineages occurred prior to the radiation of the studied Hydrobatidae species, with alleles clustering by DAB lineage, not by species. In contrast to this, *H. castro* and *H. monteiroi* DAB1 alleles were intermingled, not separated by species – and the same was observed for their DAB2 alleles. Further, *H. castro* and *H. monteiroi* DAB1 sequences clustered with Leach's storm petrel DAB1 sequences, and the same pattern was observed for DAB2. Support values for the branches were generally low, although high support was obtained for the Hydrobatinae (100), and a clade of DAB1 sequences in Hydrobatinae (82).

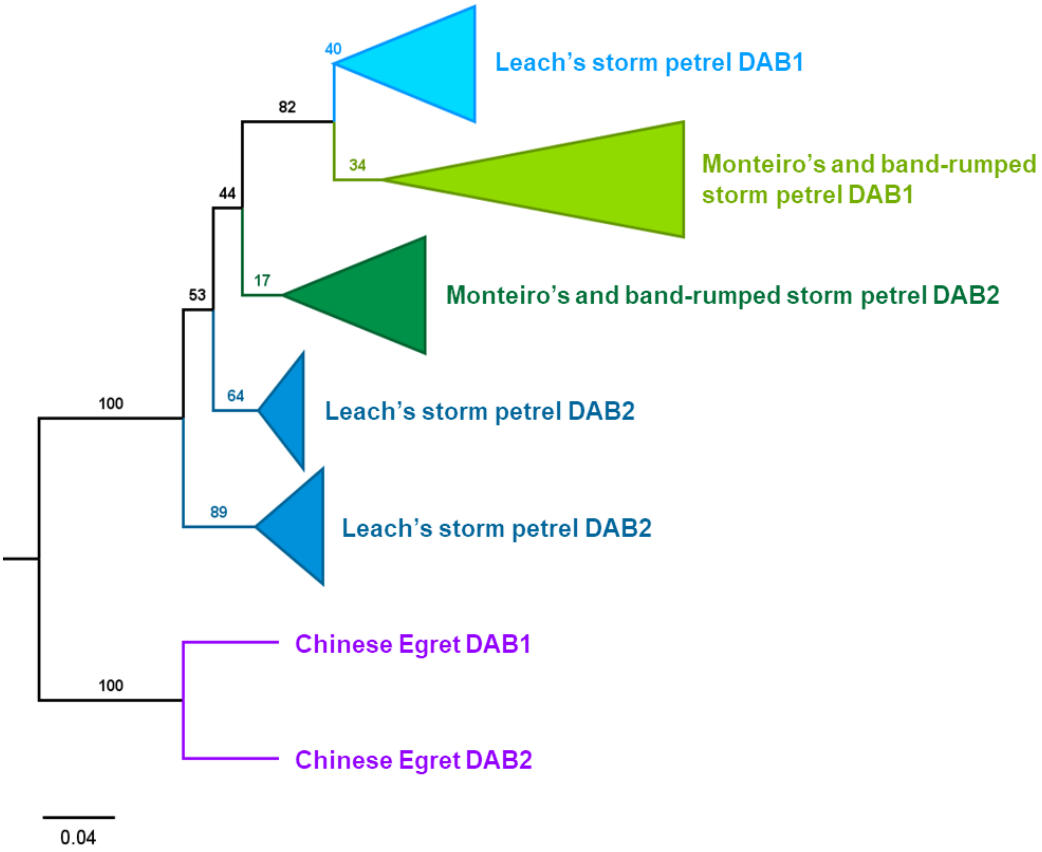



**Figure 4.6** Maximum likelihood tree based on TVMe+I nucleotide distances among MHC Class IIB exon 2 alleles within Hydrobatidae storm petrels, using egrets as an outgroup. Numbers on branches denote bootstrap support

The median-joining haplotype network for DAB1 showed 23 alleles unique to *H. castro*, 35 alleles unique to *H. monteiroi* and 7 alleles shared between the two species (Figure 4.7). For DAB2, 10 alleles were unique to *H. monteiroi*, and 14 were unique to *H. castro*, with only 2 shared alleles. In both haplotype networks, alleles did not group into distinct, species-specific clusters, but rather were extensively intermingled – albeit less so for DAB2 (Figure 4.8). To confirm that

shared alleles were not artefacts, the read count and number of individuals with those alleles were checked for both species. A high read count and/or presence in more than one individual suggested that these alleles were 'true alleles' as opposed to any remaining artefacts

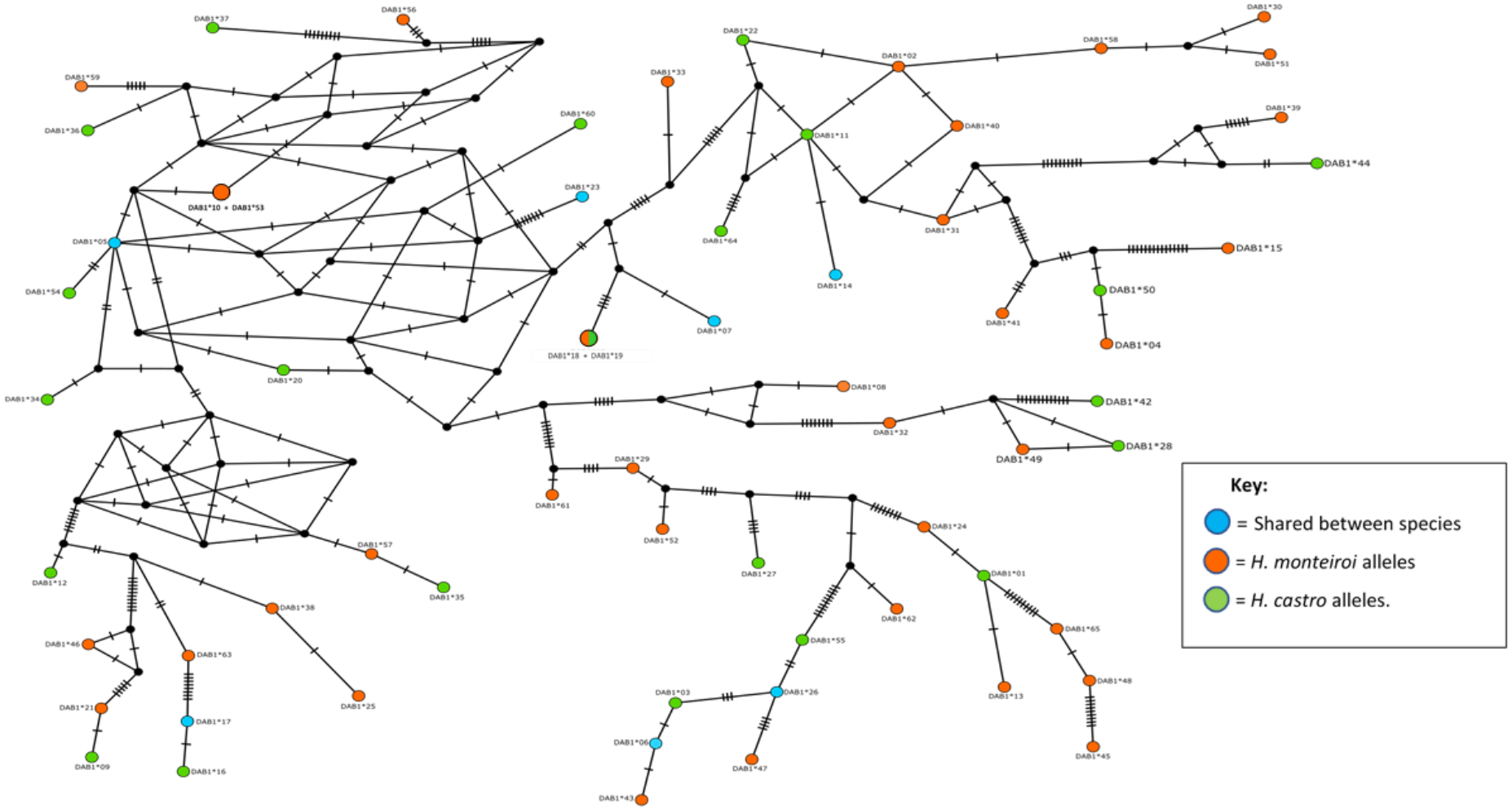



**Figure 4.7** Median-joining haplotype network of 65 DAB1 exon 2 alleles, created using PopArt. Hash marks on branches denote substitutions, nodes represent alleles which are colour coded (see legend). The two larger circles each represent two alleles that PopArt deemed identical but that in fact differed by a 3-bp indel. One of these larger nodes contains an allele unique to *H. monteiroi* and another allele unique to *H. castro*; this node is therefore split into both species' colours (orange and green).

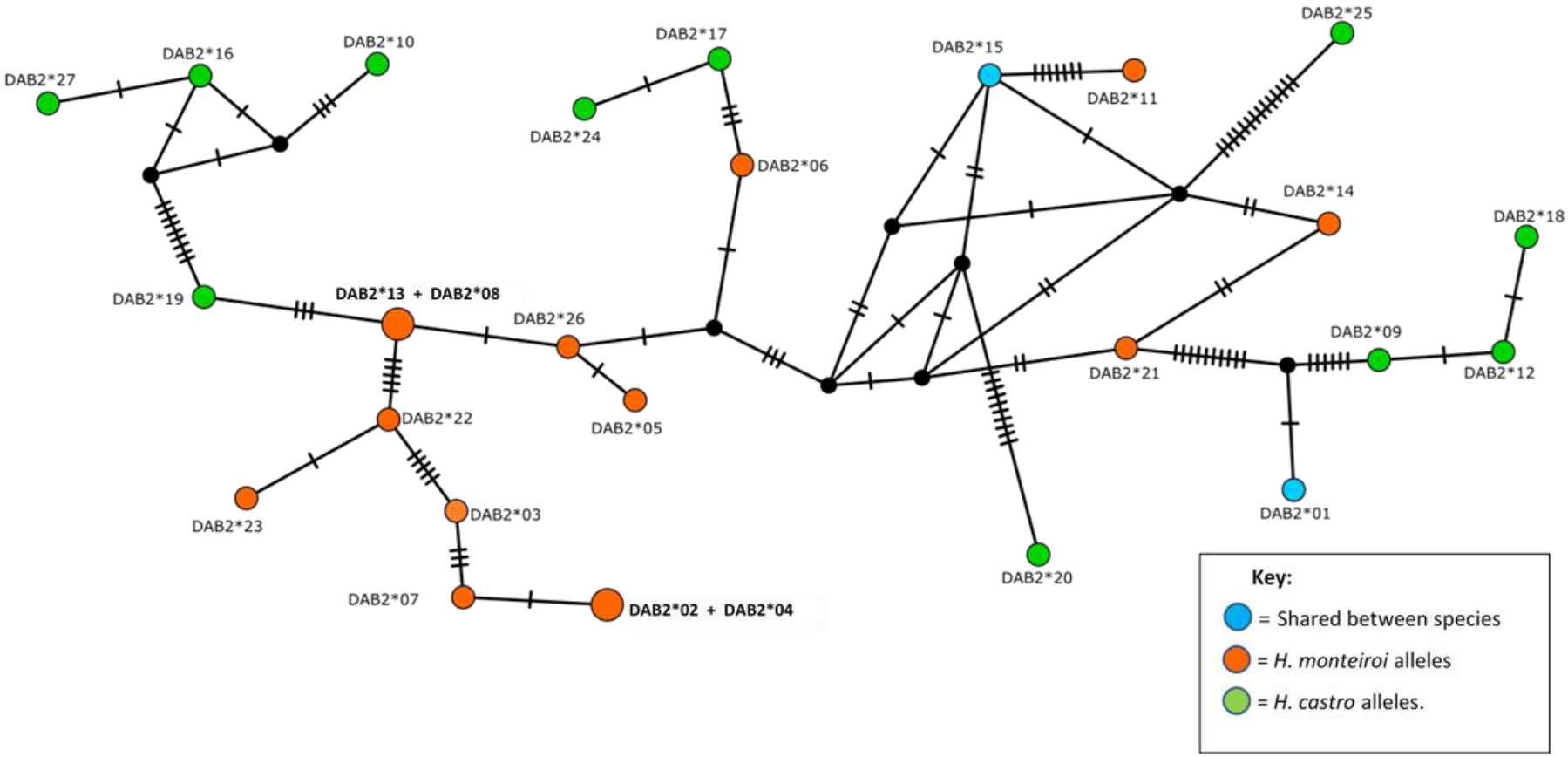



**Figure 4.8** Median-joining haplotype network of 27 DAB2 exon 2 alleles, created using PopArt. Hash marks on branches denote substitutions, nodes represent alleles which are colour coded (see legend). The two larger circles denote two cases where PopArt clustered two alleles together that in fact differ by 1 SNP in region with an indel for other alleles. PopArt's algorithm excludes indels, and so has not considered or recognised a difference in these regions

Diversity statistics calculated in ARLEQUIN (Table 4.14) revealed that DAB1 showed higher haplotype and nucleotide diversity than DAB2, within each of the two species separately, and when considered jointly. When comparing the two species, DAB1 tended to show similar or slightly higher genetic variability in the island-endemic *H. monteiroi* than in the more widely distributed *H. castro*. For DAB2, *H. monteiroi* again demonstrated higher nucleotide diversity than *H. castro*, however the reverse was seen for haplotype diversity. In all tests for Tajima's D and Fu's $F_s$, values were positive (rather than negative), but not deviating significantly from zero (all $p>0.05$), suggesting that neither signals of selection nor population size changes were detectable on average along the entirety of each DAB exon 2 alignment

.

**Table 4.14** Diversity statistics for DAB sequences in *H. castro* and *H. monteiroi*, calculated using ARLEQUIN (Excoffier and Lischer 2010). Haplotype and nucleotide diversity are shown with their standard deviations, Tajima's D and Fu's FS along with their associated p-values. Sample sizes were -*H. castro* n=111 DAB1 / 110 DAB2; *H. monteiroi* n= 99 DAB1/ 91 DAB2

| DAB | Species | Number of alleles | Number of sequences | Haplotype Diversity (+/- S.D.) | Nucleotide Diversity (+/- S.D.) | Tajima's D | Tajima's D p-vale | Fu's $F_S$ | $F_S$ p-value |
|---|---|---|---|---|---|---|---|---|---|
| **DAB1** | **Combined** | 65 | 568 | 0.959 (+/- 0.002) | 0.069 (+/- 0.034) | 2.806 | 0.994 | 0.337 | 0.598 |
| **DAB1** | ***H. castro*** | 30 | 271 | 0.905 (+/- 0.009) | 0.069 (+/- 0.034) | 2.467 | 0.985 | 11.104 | 0.943 |
| **DAB1** | ***H. monteiroi*** | 42 | 297 | 0.936 (+/- 0.007) | 0.066 (+/- 0.033) | 2.493 | 0.991 | 3.583 | 0.808 |
| **DAB2** | **Combined** | 27 | 287 | 0.914 (+/- 0.008) | 0.054 (+/- 0.027) | 3.468 | 1.000 | 8.796 | 0.933 |
| **DAB2** | ***H. castro*** | 16 | 178 | 0.897 (+/- 0.008) | 0.0348 (+/- 0.017) | 1.620 | 0.953 | 7.543 | 0.957 |
| **DAB2** | ***H. monteiroi*** | 13 | 109 | 0.689 (+/- 0.046) | 0.0401 (+/- 0.020) | 1.371 | 0.926 | 9.645 | 0.966 |

## 4.4.5: Genetic differentiation between *H. monteiroi* and *H. castro*, based DAB allele sharing

A neighbour-joining tree based on Jaccard distances of DAB1 and DAB2 allele sharing among individuals (Figure 4.9) recovered *H. monteiroi* and *H. castro* in two separate clades, with no mismatches. Similarly, an NMDS plot (Figure 4.10) showed the two species in separate positions, with no obvious overlap, but presence of a few individuals in intermediate placement between the groups.


**Figure 4.9** Neighbour-joining tree of all sampled individuals, based on Jaccard distances of DAB1 and DAB2 allele sharing (presence/absence). Orange: *H. monteiroi*, black: *H. castro* individuals.

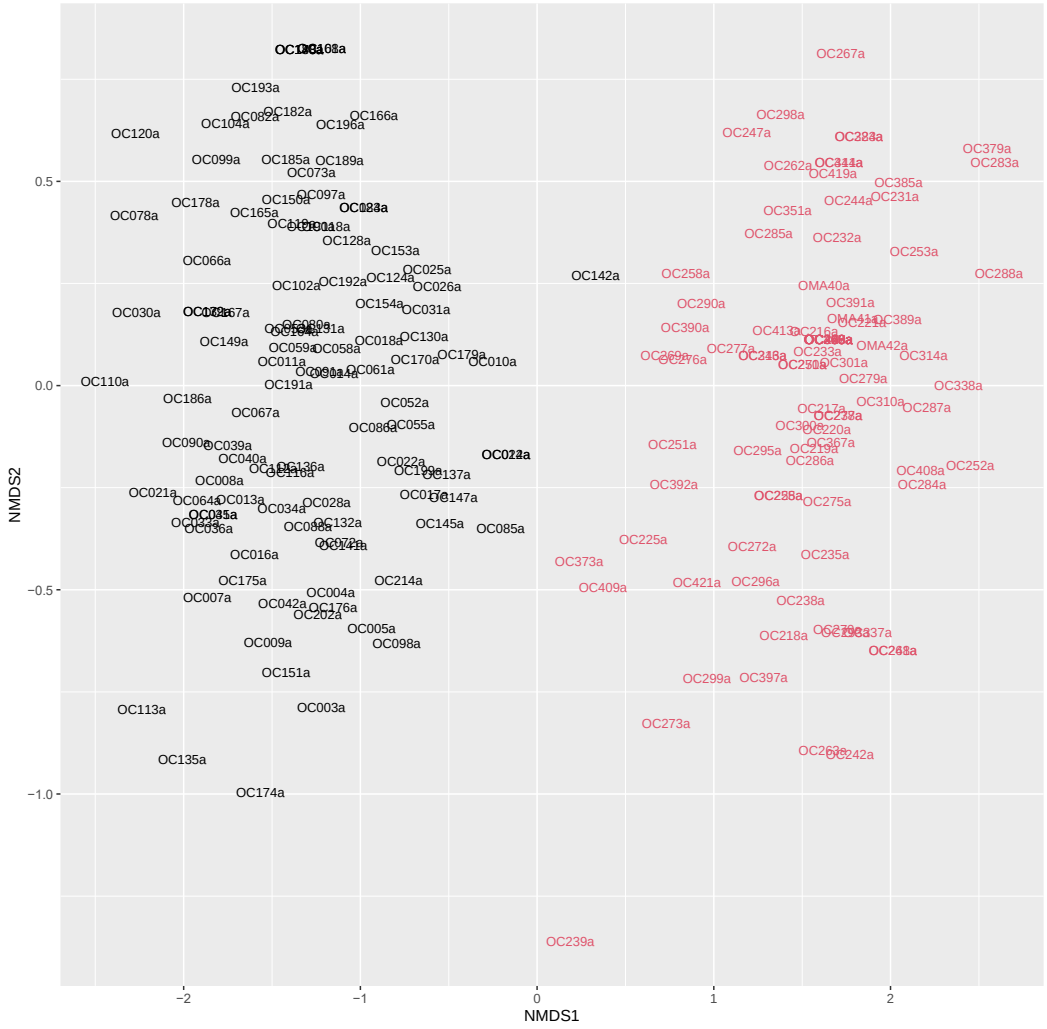



**Figure 4.10** Non-metric dimensional scaling (NMDS) plot of DAB1 and DAB1 allele sharing among all sampled individuals, based on pairwise Jaccard distances. Individuals are denoted by their ID codes (black: *H. castro*, red: *H. monteiroi*), with the middle of each ID label corresponding to the inferred placement in the NMDS plot.

## 4.5: Discussion

### 4.5.1: Differing Patterns of Divergence at Species and MHC Level

Phylogenetic analysis showed a distinct divergence between DAB1 and DAB2 lineages in storm petrels, forming distinct DAB lineage-specific clades rather than grouping by species. The clustering of DAB sequences of *H. leucorhous* sequences with *H. castro* and *H. monteiroi* suggests that lineages from an ancestral duplication of MHC Class IIB loci have been retained in all three species (Dearborn et al. 2015). Similarly, haplotype networks for the alleles of each DAB lineage showed no species-specific clusters, with haplotypes from both species dispersed and intermingled across the networks. These results document incomplete lineage sorting of *H. castro and H. monteiroi* alleles, suggesting that not enough time has passed since their divergence to allow for species-specific clusters to form (Pereira and Schrago 2018).

These results do not imply free intermixing of MHC sequences in the two focal species. Firstly, only a small proportion of encountered alleles was shared between the species. Furthermore, and in contrast to phylogenetic signals for individual alleles, joint multi-allele (genotypic) analyses of DAB1 and DAB2 variation revealed a clear overall separation of *H. castro* and *H. monteiroi,* as seen in neighbour-joining and NMDS analyses of Jaccard distances among individuals. This distance-based analysis is likely not suitable for detailed interpretation of inter-species distances and of the few possible outliers in the NMDS plot (i.e., individuals that clustered in apparently intermediate position between the main clouds of points). Regardless of these details, the contrasts seen here between analyses of single alleles versus composite MHC genotypes demonstrate the importance of characterising the full MHC profile of individuals as part of detailed evolutionary investigations.

In the Azores, the winter-breeding *H. castro* and the summer-breeding *H. monteiroi* share the same wider environment and breeding sites (Bolton et al. 2008). The divergence of species based on combined DAB genotypes, could result from two main mechanisms. MHC distinctiveness could result from diverging selection (sexual or environmental). For example, selection imposed by different pathogens experienced by the two species may be driving their MHC divergence (Spurgin and Richardson 2010). Alternatively, MHC divergence

between *H. castro* and *H. monteiroi* could be a consequence of reproductive isolation between the species, reflecting genome-wide differentiation patterns as seen for ddRAD data (Taylor et al. 2019). Lack of allele sharing between species could thus be a direct result of mate choice, where individuals of one species detect and avoid heterospecifics, ultimately resulting in population-level distinction (Santos et al. 2018). In this scenario, physical reproductive isolation (presumed to be mediated/reinforced by vocalisations; Bolton et al. 2008) may be driving species divergence. Phylogenetic clustering of individual MHC alleles may be a lagging behind this process, still showing extensive signals of incomplete lineage sorting (Zink and Barrowclough 2008). Incomplete lineage sorting between recently diverged species is common, complicating phylogenetic inferences (Maddison and Knowles 2006), especially if selection favours alleles common to both species (Mailund et al. 2014).

Here it is demonstrated that MHC divergence between species can be detectable, and that evaluating MHC genotypes separately may result in inaccurate depictions of species divergence. To further quantify MHC divergence both at the DAB and species level, it would be interesting to compare amino acid divergence, 'MHC supertype', and expression of DAB genes (Dearborn et al. 2015c; Dearborn et al. 2016b; Marmesat et al. 2017), and to investigate how patterns of divergence are distributed across the entire genome.

### 4.5.2: Species-level variability for DAB1 and DAB2 in *H. castro* and *H. monteiroi*

DAB sequence variability was slightly higher in *H. monteiroi* than *H. castro.* This is perhaps surprising, considering the low census size of *H. monteiroi* (328-378 breeding pairs; Oliveira et al. 2016) compared to *H. castro* (6,600-6,900 pairs; BirdLife International). Historic accounts of *H. monteiroi* populations suggest that the species was once far more numerous in the Azores (Bolton et al. 2008), and perhaps MHC variability has been retained. With the long generation time exhibited by storm petrels (16.5 years, Nava et al. 2017) , variability loss is predicted to be a slow process (Hailer et al. 2006). Additionally, the fitness advantages associated with MHC loci could be contributing to the retention of variability. Sexual selection (disassortative mating) or environmentally driven natural selection may retain MHC heterozygosity. Forms of natural selection that

maintain or augment diversity include balancing selection, or spatio-temporally varying selection. *H. castro* and *H monteiroi* inhabit the islet in different seasons and exhibit different migratory patterns (Bolton et al 2008), conceivably exposing them to different pathogens. The summer-breeding *H. monteiroi* experiences higher levels of nest-site competition, inhabiting the islets with other breeding seabirds, and perhaps exposure to more pathogens results in the species retaining higher levels of MHC diversity (Bolton et al 2008). Conversely, *H. castro* demonstrates migratory behaviour when not breeding on the islet, and it is perhaps surprising that higher MHC diversity was not observed, since this behaviour may expose them to a wider array of pathogens. In contrast, for the year-round Azores resident *H. monteiroi*, it is conceivable that temporal fluctuations of environmental pathogens could select for maintenance of diverse DAB allelic repertoires (Penn et al. 2002; Spurgin and Richardson 2010; Weber et al. 2013). In any case, the sharing of some MHC haplotypes between the species suggests that inhabiting the same islets and nest sites may expose the species to some of the same selective pressures (possibly shared pathogens), leading to selection for these haplotypes, despite the temporal separation of their breeding seasons.

High variability at MHC loci despite low population numbers and neutral diversity, has been documented in other species, supporting the theory that selection favours high MHC variability (Aguilar et al. 2004), however this is not a universal pattern across species studied to date. For example, some species show reduced MHC variability due to bottlenecks (Seddon and Ellegren 2004; Manlik et al. 2019). Assessment of other genetic markers alongside MHC can elucidate whether selection pressures influence specifically the MHC or reflect wider genomic patterns that suggest action of whole-genome and/or demographic drivers such as admixture or population growth.

Using non-MHC genetic markers, *H. castro* and *H. monteiroi* have previously been found to exhibit a similar pattern of variability as for DAB data: mitochondrial DNA and ddRAD sequencing data show *H. monteiroi* as similarly or more variable (nucleotide diversity and haplotype diversity) than *H. castro* (Taylor et al. 2019). High MHC variability therefore appears to coincide with relatively high genetic diversity overall in *H. monteiroi*.

### 4.5.3: Sequence Variability of DAB1 and DAB2 loci in *H. castro* and *H. monteiroi*

Prior to this study, no comprehensive assessment of DAB variability had had been done for the two study species. For the DAB1 lineage, previous work by Burri et al. (2014) used cloning and Sanger sequencing of a single individual each from *H. castro* and *H. monteiroi*, finding 5 distinct alleles for each species (plus two additional, tentative alleles for *H. castro* that were not confirmed by independent PCR replicates). In contrast, this study used HTS of 111 *H.* castro and 198 *H. monteiroi* individuals, identifying a total of 65 alleles – 5 of these alleles are identical to those previously found by Burri et al (2014), whilst 60 have not been characterised before. Whilst cloning provides a way of sequencing multiple-allelic PCR templates, obtained sequences can contain incorrect base calls from PCR and bacterial replication errors. To mitigate this, Burri et al. (2014) used sequencing replicates, conducting at least two independent PCRs with subsequent cloning and sequencing of the products. Due to the ability of HTS to sequence multiple alleles rapidly and at lower costs even for large numbers of individuals, it is not surprising that more alleles were detected in the present study, demonstrating the value of HTS in evolutionary studies (Behjati and Tarpey 2013; Sikkema-Raddatz et al. 2013). On the other hand, also HTS can also recover incorrect alleles, due to chimera formation, PCR and sequencing errors and difficulty of effective removal of contamination. Hence, the fact that the five alleles found by Burri et al. (2014) were all encountered in the present dataset is reassuring regarding data quality.

As the DAB2 lineage had not been recovered previously for *H. castro* and *H. monteiroi*, this study demonstrates for the first time that both species have retained both DAB lineages since the ancient duplication of the MHC Class IIB gene. Hence, the aim of providing a comprehensive and robust assessment of DAB variability in the two study species has been achieved.

DAB1 showed higher levels of variability than DAB2, as measured by the number of unique alleles, as well as measures of haplotype and nucleotide diversity. This trend was already apparent at earlier stages of method development, with DAB1 showing a higher number of variable sites than DAB2 in Sanger sequencing results of a smaller number of individuals (**Chapter 3**). Different levels of genetic

variability between the two DAB lineages has previously been observed in Leach's storm petrels, however this species displays the opposite pattern with higher variability observed in DAB2 than DAB1 (Dearborn et al. 2014).

Haplotype and nucleotide diversity was higher for DAB1 than DAB2, and degree of clustering by species appeared more pronounced in DAB2 than DAB1 networks, possibly suggestive of directional selection leading to divergent allele frequencies in *H. monteiroi* than *H. castro*. However, since none of the conducted tests of selection or demographic change revealed significant deviation from neutrality/stasis, factors driving genetic variability at DAB1 and DAB2 in the two study species will require additional research.

Divergence between avian DAB1 and DAB2 genes is not unexpected and has been observed in other species (Burri et al 2008, Dearborn et al 2016). Divergent genes may be retained under selection favouring heterozygosity, where inheritance of divergent genes provides high levels of overall MHC diversity (Gaigher et al. 2018). Divergence between alleles could also be a result of recombination, especially in response to pathogens (Spurgin et al 2011). In the case of *H. castro* and *H. monteiroi*, different selection pressures could be driving the observed differentiation between DAB1 and DAB2 genes. In contrast, many orders in fact only display one MHC lineage, which could be suggestive of a birth-death model, or, gene conversion and concerted evolution between the two DAB lineages could be masking the presence of both (Wittzell et al. 1999; Hess and Edwards 2002b; Strandh et al. 2012; Goebel et al. 2017). The results in this chapter suggest that concerted evolution is not masking the presence of two exon 2 DAB lineages in *H. castro* and *H. monteiroi*, with the two lineages demonstrating some level of sequence divergence. The phylogeny in Figure 3.6 demonstrates that *H. castro* and *H. monteiroi* cluster by DAB lineage, but this is not true for other species in that phylogeny, which may indicate such concerted evolution and/or gene loss occurring in other species. Acquisition of exon 3 sequences for *H. castro* and *H. monteiroi*, which typically demonstrate stronger signals of sequence divergence, may provide better conclusions on phylogenetic differences between DAB1 and DAB2 in the two species.

### 4.5.4: Estimation of DAB Copy Numbers

Determining copy numbers for DAB lineages has proven challenging, due to frequent (i.e. recurrent) locus duplication events and the homogenising force of gene conversion that can render different copies indistinguishable (Miller and Lambert 2004). High levels of variation both between and within species exist (Bentkowski and Radwan 2019), and even inter-individual variation in MHC profile Class I and Class II copy number has been found (Minias et al. 2019). Among birds, passerines can have thousands of copies, while non-passerines generally appear to contain lower copy numbers (Biedrzycka et al. 2017a; O'Connor et al. 2019). The use of HTS short-read technology (e.g. Illumina MiSeq) for MHC genotyping of individuals can unfortunately only provide tentative estimates of copy number (O'Connor et al. 2019; Lighten et al. 2014), resulting from various biases arising throughout the PCR amplification, library preparation and sequencing stages.

For DAB1, individuals in the present dataset had between one and six different alleles, indicative of at least three loci, consistent with findings by Burri et al. (2014). DAB2 was less variable than DAB1, with individuals possessing 1-3 alleles. Notably, only one sample contained three alleles and all others only 1-2. This demonstrates that DAB2 consists of a minimum of 2 loci, although presence of additional locus copies and/or among individual variation in copy numbers cannot be excluded (Bentkowski and Radwan 2019).

It was hoped that the use of HTS would also allow for the copy number of MHC loci to be compared between each DAB lineage, similar to other studies (Dearborn et al. 2016; Hoover et al. 2018). The genotyping pipeline used here did not assume a presence nor absence of CNV, however similar to Dearborn et al. (2016), read depths were used to define alleles, and amplicons with low read counts were discarded as putative artefacts.

Parent-offspring comparisons were used to assess Mendelian inheritance and copy number, since copy number should follow Mendelian inheritance with offspring alleles reflecting those found in the parents (Locke et al. 2006; Wang et al. 2008). If copy number can be determined for individuals, this can be further supported by evidence of family inheritance (Wang et al. 2008). For example, if read depths of alleles in an individual conformed to strict percentage

boundaries, then copy number could be estimated - i.e., if an individual had four alleles, each accounting for 25% of reads, then allele copy number could be estimated as four (or multiples thereof). Histograms of read percentages for both DAB lineages however did not fall into any distinct percentage categories, precluding copy number estimation. In terms of inheritance, for DAB1, 11 *H. castro* and 12 *H. monteiroi* families were assessed. Of these, only one *H. castro* and four *H. monteiroi* families did not follow Mendelian inheritance patterns, with offspring either missing parent alleles entirely, or they had alleles that were not represented in either parent. In DAB2, 11 families for each species were assessed, with three *H. castro* and four *H. monteiroi* families not conforming to Mendelian inheritance. These cases of missing or extra alleles could indicate that the wrong parent was sampled due to field misidentification, extra-pair paternity, gene conversion, or contamination during PCR. Field misidentification is also discussed in section 2.4.2.3; in the case of identifying mate pairings, mistakes are possible if a non-mated individual visits the nest chamber at the time of sampling. This can be avoided by continued sampling of the same nest site to confirm that the same two individuals, and therefore a mated pair, are using the nest. Extra-pair paternity (EPP) has not been widely studied in *H. castro* or *H. monteiroi* and is considered low or non-existent, especially considering their long life spans and associated mate preferences for monogamous bonds and joint parental care (Griffith et al. 2002). Low EPP rates are typical in Procellariiformes (Quillfeldt et al. 2001; Jouventin et al. 2007; Dearborn et al. 2016), and whilst unlikely cannot be ruled out completely - one study of *H. monteiroi* pairings identified just 2 cases of EPP out of 71 sampled offspring, along with one case of brood parasitism (Nava et al. 2017). Indeed, the apparent one-parent mismatches between parent and chicks in the inheritance checks, could indicate EPP where males mismatch to the chicks, or brood parasitism where it is the females that mismatch; further genetic evidence would be needed to conclude this, however. It is also possible that filtering had erroneously excluded true alleles from the parents, or filtering had not been strict enough, and additional chick alleles could in fact represent artefacts. In any case, additional multi-locus data from microsatellites or ddRAD sequencing would provide better insights into parentage (as in Dearborn et al. 2016). Nevertheless, despite some unclear cases, most offspring investigated in this

study contained alleles inherited from both parents in a pattern consistent with Mendelian inheritance. Conclusive determination of allele and locus copy numbers could however not be made.

Genotyped alleles can be used to inform on minimum locus copy number (O'Connor et al. 2019), and *H. castro* and *H. monteiroi* demonstrated between 1-6 and 1-3 alleles per individual, respectively. The proportion of individuals with >2 alleles (indicative of >1 locus copy) for DAB1 was 84%, whilst for DAB2 this was only 1%. This implies that DAB1 comprises with multiple copies in the two study species, contrasting findings in *H. leucorhous*, whilst DAB2 may comprise only 1 locus copy, as in *H. leucorhous* (Dearborn et al. 2016). Some disparity in DAB locus copy numbers between species is not unexpected, considering the wide variation reported across animal taxa (Bentkowski and Radwan 2019; O'Connor et al. 2019). However *why* these species differ in DAB copy number remains less well understood. It is possible that copy number in *H. leucorhous* has been underestimated, or that changes in DAB copy number have occurred since the split of *H. leucorhous* from the ancestor of *H. castro* and *H. monteiroi*. Estimates of copy numbers remain tentative based on the type of data generated as part of this study. Short-read sequencing data from PCR amplicons only allows minimum allele copy number within individuals to be identified, based on the number of distinct sequences encountered. However, this can underestimate true allele counts due to gene conversion, or presence of multiple (identical) copies of the same allele (O'Connor et al. 2019).

DAB2 presented with a surprisingly high number of homozygotes, which could be explained by PCR amplification and/or sequencing bias of certain alleles, or allele drop-out that results in a false positive rate of homozygotes (Guo et al. 2012; Weimer and Sherwood 2019). Alternatively, copy number variation could explain the perceived high number of homozygotes, especially if there are unidentified DAB2 alleles that this study has not unearthed. It is not known how many copies are present in the genome of *H. castro* and *H. monteiroi*, nor if all alleles have been captured, so apparent homozygotes from this chapter might instead be heterozygotes with undiscovered alleles. More work will be required to investigate the factors leading to the high homozygosity for DAB2 observed in this study.

Various types of selective pressures can shape MHC copy number, leading to high levels of CNV between species and individuals. Pathogens in host populations have been shown to drive increases in copy number (Bentkowski and Radwan 2019). Even different MHC classes within an individual can display different selection pressures – and therefore copy number – with life history, exposure to pathogens and selection type (i.e. fluctuating versus balancing, in MHC Class IIB versus Class I) (Minias et al. 2016; Bentkowski and Radwan 2019; O'Connor et al. 2019). Finally, locus copies can be lost through genomic deletions, part of the birth-death model commonly used to characterise evolution of MHC loci (Minias et al. 2018; Minias et al. 2019). It is apparent that commonly-used, current methods may benefit from additional, more detailed sequencing to provide better estimates of copy number, as evidenced in He et al. (2021). Until such estimates are more accurate for *H. castro* and *H. monteiroi*, full characterisation of MHC differentiation is limited to allele comparisons and estimates of copy number. Divergence of alleles between DAB lineages can still be investigated without copy number, however.

### 4.5.5: Technical considerations: Allele Calling Pipeline, and Accounting for Tag Jumping and Contamination

The pipeline developed and applied in this chapter effectively reduced reads into potential alleles. The pipeline filters out sequences that based on a range of criteria do not correspond to true alleles, based on thresholds regarding read count, contamination, and sequence length. Previous genotyping studies (e.g. Rekdal et al. 2018) removed alleles found in only one individual, however that approach was not taken here. For each allele that was present in only one individual, the read depths were used decide whether or not to keep the allele in the data set. An allele was kept if the read count was higher than those found in errors or negatives, if the read percentage of the allele accounted for >10% of the total read count, and in homozygotes, if read depth was considered high (i.e., at least 1,000 reads) for one allele, especially when compared to the relative read count of other more common alleles. It is possible that sampling >200 individuals has resulted in the detection of rare alleles, and so these were kept in for this analysis. In addition, filtering steps to remove reads in negatives or unused MID-tag combinations should have removed artefacts associated with contamination or tag-jumping, that could be mistaken for alleles. In previous

studies (e.g. Stuglik et al. 2011; Sommer et al. 2013; Rekdal et al. 2019), replicates were used to distinguish true alleles from high-read count artefacts that had incidentally high read counts, however replicates were unfortunately not included in this study, precluding such comparisons.

The inclusion of PCR negatives in the pooling process demonstrated good practice for detecting contamination. PCR negatives did not display a band on agarose gel and were thus considered free from contamination. However, HTS returned reads for the MID-tags linked to the negative controls, revealing low levels of contamination, but in too low a level to be visible on agarose gels. Alternatively, low levels of tag jumping may explain the moderate read counts recovered from negative controls. Both tag-jumping and contamination are commonly found for HTS MID-tagged datasets of PCR amplicons, necessitating adequate bioinformatic filtering procedures to obtain accurate read counts for individuals in the dataset (Costello et al. 2018; van der Valk et al. 2020).

Dearborn et al. (2016) investigated the possible presence of CNV for *H. leucorhous*. Incidence of CNV was determined to be rare (at most found in 1% and 14% of individuals at DAB1 and DAB2, respectively).

### 4.5.6: Conclusions

This chapter provides a pipeline for characterisation of the MHC DAB1 and DAB2 in *H. castro* and *H. monteiroi* storm petrels breeding on the Azores. Further, a comprehensive assessment of allelic variability at these loci is provided for population samples of both species, based on Illumina high-throughput sequencing (HTS). Whilst precise copy number of MHC DAB loci could not be determined, the data are consistent with three locus copies for DAB1 and one copy for DAB2 (likely minimum estimates). The obtained data did not yield clear signals of selection or demography at the loci, but the DAB1 lineage was found to be more diverse than DAB2, and most investigated metrics showed the Azores endemic *H. monteiroi* as similarly or even more genetically variable than the geographically widespread *H. castro.*

DAB1 and DAB2 alleles in the two study species clustered with homologous alleles in another species from the same family (Hydrobatidae), Leach's Storm petrel, suggesting conservation of lineage distinctiveness in this family despite gene conversion and lineage loss being previously documented for DAB lineages

in other avian species (Goebel et al. 2017). Reciprocal monophyly of individual alleles was not observed, with haplotype networks showing no distinct species clusters, and alleles intermingled between *H. castro* and *H. monteiroi*, likely resulting from incomplete lineage sorting. However, the two species shared only few alleles and were clearly differentiated at DAB loci in multi-locus analyses. These findings are consistent with previous evidence from various phenotypic characteristics and genetic markers that the two Azores populations warrant species-level distinction, and tentatively suggest that MHC loci are experiencing different selection pressures in the two study species.

## 4.6: Acknowledgements:

I'd like to thank Angela and Trudy in the genomics hub for their assistance in library preparation and Illumina sequencing. I'd like to thank Frank for his input and knowledge on this chapter, particularly some of the analyses. I'd like to thank the BOU for their financial support in awarding me a grant, which funded the sequencing.

# Chapter Six: General Discussion

*Hydrobates monteiroi chick, taken on Praia Islet*

**"Finishing a PhD is like finishing a group project where your partner made a ton of mistakes at the beginning of the assignment. Except your partner is just you 4 years ago."**

**- @_JohnMola on Twitter**

# Chapter 5 General Discussion

## 5.1. Project aims

The main aim of this PhD research was to assess and quantify levels of genetic diversity and divergence between two sympatric storm petrel species, *H. castro* and *H. monteiroi*, especially at the level of the MHC. The MHC plays a vital role in immune response, and therefore characterisation of MHC diversity may inform questions on fitness, mate choice and evolution of the two species. The project aimed to (1) explore current genetic divergence and molecular techniques to identify the visually and morphologically similar *H. castro* and *H. monteiroi,* and (2) characterise the MHC profile of both species to assess how divergence and diversity may differ between the two species.

Specific objectives were to (1) develop a new screening method to aid identification of mismatches between in-field identification and genetic assignment, and apply this to samples collected from storm petrels breeding in the Azores; (2) identify the presence of a retained, ancient duplication of the MHC in both species, representing two distinct lineages of genes; (3) develop lineage-specific primers that allow characterisation of MHC, with a fragment suitable for short-amplicon sequencing using Illumina MiSeq; (4) develop a data analysis pipeline to assess Illumina sequencing data and quantify the number of alleles for each species and MHC lineage, expanding on previous research; (5) verify lineage-specific primers recover allelic variability in storm petrels breeding on the Azores, consistent with Mendelian inheritance; (6) assess the diversity and divergence of both MHC lineages, to compare patterns of variability between the endemic *H. monteiroi* and widely-distributed *H. castro,* and (7) use the above research to compare MHC versus more general genetic divergence, exploring the how the interplay of divergence, selection and diversity may contribute to species divergence. These findings are especially important for the endemic *H. monteiroi*, especially quantifying genetic diversity of the MHC and how this may translate into individual fitness, informing future conservation plans.

## 5.2: Completion of Research Objectives

### 5.2.1. Main findings

A new method to detect mismatches between sampling-location prescribed species, and genetic clade assignment was developed, utilising clade-specific primers. Primers targeted a fixed nucleotide polymorphism between samples in different clades, producing clade-specific bands of different size on agarose gel. This provided a sequencing-free method of identifying mismatches between in-field assumption and genetic assignment of clade. The method proved 100% successful in correct clade assignment, when tested on *H. castro* and *H. monteiroi* samples taken in the Azores. This primer pair can thus be recommended for future studies between *H. castro* and *H. monteiroi* in the Azores and (with appropriate assessment) will have applications across the wider geographic range of the *H. castro* species complex.

The novel description of two DAB lineages in *H. castro* and *H. monteiroi* contributes significantly to current research on MHC for these species (Burri et al. 2014) and is the first such description for the retention of the DAB2 lineage. Phylogenetic analysis confirmed the two lineages are diverged for both species, consistent with the ancient duplication preceding species radiation (Burri et al. 2010). Allele numbers for each DAB lineage were described and compared, again building upon current estimates and demonstrating a discordance with other related species (Burri et al. 2014; Dearborn et al. 2015). Individual genotyping using HTS allowed for the sharing of alleles between species to be investigated, and genetic diversity estimates revealed that *H. monteiroi* possesses slightly higher variability than *H. castro*. This is perhaps surprising for an island endemic with low population size and raises interesting questions about selection at MHC loci. When MHC alleles for each DAB lineage are combined in an individual (i.e., an individual's full MHC genotype is characterised), divergence estimates show a clear separation of species at MHC loci, despite the sharing of some MHC alleles. This contributes even more genetic data regarding speciation of *H. castro* and *H. monteiroi*, and demonstrates the importance of selection in MHC diversity, compared to other neutral genetic markers that do not show patterns of species divergence (Silva et al. 2016). Together, the findings presented here have

implications for understanding the effects of selection and MHC on diversity and speciation.

### 5.2.2.: PhD Chapter Summaries

In chapter 2, a new method using clade-specific primers was developed and described, providing the ability to determine genetic clade of *H. castro* and *H. monteiroi* samples at the gel electrophoresis stage. Phylogenetic analysis of available mitochondrial DNA for *H. castro* and *H. monteiroi* revealed fixed base differences between clades, which in previous studies has demonstrated low levels of mismatch between sampling location and genetic clade assignment (Taylor et al. 2019). The primers targeted these fixed base differences, and lab testing of the primers on samples from Azorean *H. castro* and *H. monteiroi* did not detect any such mismatches between expected and observed species identities. Despite being unable to detect mismatches in the current dataset, the method would be ideal for such detection, singling out samples for further genetic characterisation (i.e., Sanger sequencing, ddRAD or microsatellite analysis). This method appears promising for future use in studies involving mitochondrial DNA and is particularly useful when used in the Azores for *H. castro* and *H. monteiroi.*

In Chapter 3, the ancient duplication of MHC genes (Burri et al. 2010) was targeted, aiming to detect and amplify both DAB lineages in *H. castro* and *H. monteiroi.* The existence of the DAB2 lineage for both species was confirmed, and primers designed to allow short-read amplicon sequencing of both lineages, using Illumina MiSeq. Phylogenetic analysis confirmed that the primers discriminated between DAB lineages, and sequencing confirmed the existence of polymorphic sites, indicative that primers were amplifying multiple alleles and were capturing variability in MHC sequences. This chapter demonstrates a novel discovery for *H. castro* and *H. monteiroi*, with the description and reliable amplification of a second DAB lineage.

In Chapter 4, the primers developed in Chapter 3 were used to create MID-tag primers, capable of assigning individual genotype when used in HTS. A pipeline to reduce reads into alleles was developed and utilised, revealing increased numbers of DAB1 alleles (compared to those found by Burri et al. 2014), and novel characterisation of DAB2 alleles for both *H. castro* and *H. monteiroi.*

Patterns of diversity between DAB lineages differed from previous studies in another species, *H. leucorhous*, where DAB2 was more diverse than DAB1 (Dearborn et al. 2014); in the present study, the opposite was observed with higher diversity in DAB1 alleles. Phylogenetics revealed MHC alleles alone segregated into distinct, DAB-specific clades (as seen in other species, see Dearborn et al. 2014) but did not separate according to species. Comparisons of MHC diversity between species revealed higher genetic diversity in *H. monteiroi* individuals, suggesting MHC diversity is not being constrained by low population size (Aguilar et al. 2004) which may be due to a much larger past population size, combined with strong selection favouring continued high MHC diversity (Landry et al. 2001; Aguilar et al. 2004; Huchard et al. 2010).

Chapter 4 also revealed that whilst individual MHC alleles do not separate according to species, when the alleles possessed by an individual are combined into a single genotype, there is a clear divergence of MHC into species-specific groups. This reflects the divergence observed at other encoding regions such as mitochondrial DNA (Taylor et al. 2019), suggesting selection at MHC loci may even contribute to speciation (Eizaguirre et al. 2009a). It is also possible that the ability to detect MHC genotype may influence species recognition between storm petrels, and act as an additional pre-mating isolation mechanism (Ritchie 2007) to that already described (i.e. vocalisations as described in Bolton et al. 2004). The findings here serve as a useful contribution to our knowledge about selection, MHC variability and speciation in *H. castro* and *H. monteiroi.*

### 5.2.3: Patterns of Variability in *H. castro* and *H. monteiroi,* and Implications for Speciation, Mate Choice and MHC Heterozygosity

Characterising MHC diversity and genotyping of individual MHC profile reveals interesting patterns of variability across different markers, species and DAB lineages. Diversity statistics show that *H. monteiroi* demonstrates slightly higher genetic diversity than *H. castro*, which could be considered surprising for an island endemic with a small population size. Small population sizes in island endemics has been shown to contribute to low genetic diversity and could result in inbreeding (Frankham 1995; Furlan et al. 2012). However, levels of general genetic diversity are not always the same for MHC loci, and variability patterns between the two can vary wildly (Aguilar et al. 2004). High MHC variability in *H.*

*monteiroi* could demonstrate the strong selection pressure on MHC loci to maintain heterozygosity for individual fitness (Penn et al. 2002; Sommer 2005b; Huchard et al. 2010). Prior estimates of population size suggest *H. monteiroi* individuals were once much more numerous (Bolton et al. 2008), with higher diversity reflected by this larger population size; the long life span and generation time of storm petrels (Bolton et al. 2008b; Nava et al. 2017) may mean that diversity loss is slow to occur despite lower populations. Previous studies suggest variability could be retained due to a historical population bottleneck, but genetic signatures of such a bottleneck have not been observed in either *H. castro* or *H. monteiroi* (Nava et al. 2017).

Phylogenetic analysis shows MHC alleles alone are not divergent between species, but they do diverge according to DAB lineage. Assessing MHC alleles alone demonstrates a level of incomplete lineage sorting (ILS) for MHC loci, which is typical of recently diverged species (Welch et al. 2011). However, this pattern is not observed when alleles are combined in an individual's multilocus genotype, and divergence between individuals is instead assessed. Here, individuals show distinct divergence and sort into species-specific clusters, indicating that incomplete lineage sorting is far less evident than when considering single MHC alleles in isolation. The presence of a low level of outliers suggests that species could still be in the process of diverging or could even indicate gene flow or hybridisation (also raised in Silva et al. 2016), but more genetic information is needed to confirm this. Other markers such as mtDNA have also shown strong species divergence with small numbers of haplotype sharing between species, reflecting the patterns uncovered here with MHC loci (Friesen et al. 2007c; Smith et al. 2007a; Taylor et al. 2019). In comparison, neutral markers such as microsatellites show strong levels of ILS between *H. castro* and *H. monteiroi*, and ddRAD sequencing of unspecified genes also show less definitive separation (Silva et al. 2016; Taylor et al. 2019). These represent neutral markers that are typically not under selection and so may not exhibit strong signals of sequence divergence, compared to MHC loci (Manlik et al. 2019). This disparity in divergence suggests that MHC loci in *H. castro* and *H. monteiroi* may be under strong selection to maintain heterozygosity within and between species, which may in turn have implications for mate choice or response to environmental pathogens (Piertney and Oliver 2006; Biedrzycka et

al. 2018). It may also demonstrate the need to focus on MHC diversity rather than neutral diversity when assessing population fitness and future survival prospects (Manlik et al. 2019).

Divergence of MHC genotype according to species further supports research that *H. castro* and *H. monteiroi* are reproductively isolated (Friesen et al. 2007b; Taylor and Friesen 2017). It is understood that MHC is detectable by scent (Boehm and Zufall 2006; Caro and Balthazart 2010; Leclaire et al. 2017; Grogan et al. 2018), and assessment of MHC by scent has been documented in petrels in previous studies (Steiger et al. 2008; Leclaire et al. 2017). The MHC forms a strong part of mate choice in some species, and also serves as a way to recognise individuals (Eizaguirre et al. 2009b; Huchard et al. 2010). The discovery here that MHC genotypes are largely divergent between species, indicates that assessment of MHC provides a mechanism for species recognition. Whether it be in the context of mate choice or species recognition, divergent MHC genotypes could be informing and contributing to the maintenance of reproductive isolation, and therefore also speciation (Ritchie 2007; Eizaguirre et al. 2009a; Brock and Wagner 2018).

## 5.3: Future Research Directions and Limitations

### 5.3.1: Limitations of Research

Potential limitations of the clade-specific primers have been touched upon in Chapter 2. In the lab, samples for testing the clade-specific primers were limited to those from *H. castro* and *H. monteiroi* collected in the Azores, with a very small number (circa 4 samples) of *H. castro* from Ascension and St Helena islands. Whilst testing on these samples was widely successful, demonstrating a 100% success rate of matching in-field assessment to genetic assignment of species, it is unclear how the primers would perform when tested on *H. castro* samples taken from other locations across their geographic range. *In silico* testing used data from GenBank to test the possible annealing patterns of the primers and appears to show some issues with applying the primers to samples from other locations. In some cases, primers fail to bind at all, or amplify with primers intended for the opposite clade - most notably samples from Hawaii belong to a *H. castro* clade, however *in silico* they appear to amplify with the primers designed around the *H. monteiroi* clade. In addition, lab testing revealed no

mismatches between expected and observed clade, which, whilst it proves the method works, did not allow for any analysis of mismatch detection and how mismatches can be further explored using more in-depth analysis (e.g., Sanger or ddRAD sequencing). Overall, this method appears robust, but it would benefit from additional physical testing in the lab with real samples taken from other geographic locations. This would determine whether the method can be used in a wider context, and if not, it could be adapted to suit other locations.

Whilst every effort was made to capture whole MHC variability, there is still the possibility that some alleles that exist in the population may be unrecorded here. During primer design phases, it is possible that some MHC variability was missed, possibly because individuals with rare alleles have not yet been sampled, or primer design could unintentionally exclude alleles. Specifically, primers placed too close to the exon could miss regions that encode divergent loci, should those regions be found further away from the exon (Worley et al. 2008; Strand et al. 2013). Recommendations for primer design made by Burri et al. (2014) were followed, and the testing of multiple primer pairs should increase the chances of designing a pair to capture variability. Nevertheless, the subsequent capture of 65 DAB1 and 27 DAB2 alleles still provides a more comprehensive of MHC diversity than previous research (Burri et al. 2014).

High throughput sequencing (HTS) is invaluable for the sequencing of multiple alleles and has proved beneficial in MHC studies (Margulies et al. 2005; Sikkema-Raddatz et al. 2013; Lighten et al. 2014; O'Connor et al. 2019), and subsequent pipelines work well to reduce reads into true alleles. However, there exists multiple different approaches to categorising alleles, which could present with different results depending on the method used (Stuglik et al. 2011; Sommer et al. 2013; Lighten et al. 2014; Rekdal et al. 2018). The method described here appears reliable at reducing reads to alleles, and it appears that strict parameters on read counts, read count percentage share and removal of contamination or tag-jumping is sufficient. The inclusion of replicates in this study would further improve allele-calling, however the lack of replicates is an oversight that should not be missed in future.

It is unfortunate that copy number of MHC loci could not be reliably determined in this study, and allele numbers served only to provide a minimum estimate.

Knowing copy number would aid determination of Mendelian inheritance of alleles, and would also inform conclusions on MHC variability, selection and fitness. Copy number is difficult to determine, however, with methods currently used prone to bias (Minias et al. 2019) and efforts made here should not be underappreciated.

### 5.3.2: Future Recommendations

Several recommendations can be made for future work.

Firstly, the methodological determination of genetic clade, using clade-specific primers (Chapter 2), should be tested on samples from other locations across the geographical range of the *H. castro* species complex, to test the potential wider applications of the method. The current mismatches presented in the mitochondrial data available (Friesen et al. 2007b; Smith et al. 2007a; Taylor et al. 2019) did not present in lab testing, and it would be interesting to assess the ability of the method to detect such mismatches. Similarly, *in silico* testing identified some potential issues with the method working on samples obtained from other locations, and it would be useful to compare this with physical testing, to see if *in silico* estimations of primer use were correct. Should the primers prove unsuitable for testing in other locations, the method itself (i.e., clade-specific primers) could be developed further, to provide more specific primers for these locations, especially in areas with similarly diverging, seasonal populations, such as Madeira, Cape Verde, Ascension Island and the Galapagos (Taylor et al. 2019).

The genotyping of MHC in mated individuals here provides an exciting basis for future studies involving mate choice. The selection pressures on MHC genes have been extensively described, and this selection can extend into the decisions made by individuals in selecting partners for mating, to improve offspring fitness (Piertney and Oliver 2006). Individuals may choose to target mates that are dissimilar at MHC loci, to improve offspring heterozygosity (Santos et al. 2017; Hoover et al. 2018), or in systems where heterozygosity does not provide a fitness advantage, individuals may choose to mate with similar individuals to avoid excessive heterozygosity (Slade et al. 2019). Alternatively, MHC genotypes may prove so diverse between individuals, that random mating alone can provide enough fitness advantage without specifically targeting MHC (as

described by Dearborn et al. 2016 for Leach's storm petrels). Using the data provided here in mate choice studies provides an exciting opportunity to explore the mechanisms underlying mate selection and MHC genotype selection more deeply. In order to truly quantify if mate choice is targeting MHC, wider genome testing of other loci could also be undertaken, to assess if other genes alongside MHC are also targeted (Hume et al. 2018). Further to this, expression of MHC genes could also be explored, to assess both the advantage the genes provide, and their suitability to detection by others (Wegner et al. 2006; Dearborn et al. 2015c).

Future studies could also assess the parasites or pathogens present amongst *H. castro* and *H. monteiroi*, which in turn may provide further information on how environmental selection pressure, in the form of selection for effective immune responses, may shape MHC divergence between the two species (Spurgin and Richardson 2010)

Finally, the determination of copy number would prove useful in future for MHC-based studies on *H. castro* and *H. monteiroi*, providing better insight to MHC diversity and selection. This could be achieved in future through long-read sequencing (He et al. 2021).

## 5.4: Implications for Storm Petrels and Conservation

The genetic profiles and behaviour of the two species could have different implications for their futures, based on their differing population dynamics, and the findings presented here could contribute to ongoing research in conservation of the two species.

The research presented here provides the most detailed characterisation of MHC Class IIB genes in *H. castro* and *H. monteiroi* thus far. With its implications for fitness, characterising MHC diversity can provide researchers with insights into how *H. castro* and *H. monteiroi* are equipped to deal with pathogens, especially for the endemic, population of *H. monteiroi*, which owing to its much smaller population size, may be vulnerable to sudden population decline due to new pathogens or parasites (Landry et al. 2001; Weber et al. 2013; Biedrzycka et al. 2018). Characterising MHC diversity could enable predictions to be made about

population-level impacts, should a new threat to fitness be identified in populations. Further to this, characterising MHC genotypes could also allow for individual fitness to be assessed, and whether MHC genotype of mated pairs is linked to successful breeding or offspring fitness (Forsberg et al. 2007; Huchard et al. 2010). Genotyping of individuals could also prove useful in any future mate choice studies, especially in regard to the sustainability of viable genetic diversity (Manlik et al. 2019)d. The genotyping of individual MHC profiles provided here is the first stage in identifying if MHC-based mate choice occurs in either population and characterising whether individuals target the MHC profile of potential mates, could inform any potential translocation and breeding programmes.

## 5.5: Acknowledgements

At this final chapter, I'd like to thank Frank, Rob, Renata and Carsten, my entire supervisory team who got me through this, and I wouldn't have got there without their support.

This project was funded by KESS II

*Knowledge Economy Skills Scholarships 2 (KESS2) is a pan-Wales higher level skills initiative led by Bangor University on behalf of the Higher Education sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.*

*Ysgoloriaeth Sgiliau Economi Gwybodaeth 2 (KESS2) yn Gymru gyfan sgiliau lefel uwch yn fenter a arweinir gan Brifysgol Bangor ar ran y sector AU yng Nghymru. Fe'i cyllidir yn rhannol gan Gronfeydd Cymdeithasol Ewropeaidd (ESF) cydgyfeirio ar gyfer Gorllewin Cymru a'r Cymoedd.*

Ysgoloriaethau Sgiliau Economi Gwybodaeth
Knowledge Economy Skills Scholarships

UNDEB EWROPEAIDD
EUROPEAN UNION

Llywodraeth Cymru
Welsh Government

Cronfa Gymdeithasol Ewrop
European Social Fund

# Appendices

## Appendix for Chapter 2

### A 2.1: Comparing the use of DMSO and BSA in PCR

When testing clade-specific primers, multiple non-specific bands of non-target length were consistently visible when viewed on agarose gel. The target bands of around 90 and 200bp were also weak, making it difficult to determine if clade-specific primers were successful or not. The use of DMSO and BSA PCR additives were trialled, to improve amplification strength and specificity. BSA and DMSO were tested both separately and together, on the same 3 samples and one negative control. The additives were tested in one PCR reaction, with the same cycling conditions applied to all 3 experiments. The PCR reagent mix remained consistent, only changing the amount of BSA or DMSO, according to which mix was being tested. PCR results were viewed on agarose gel (Figure A2.1). Whilst some non-specific bands were still present, the target amplicon (~100 and ~200bp) strengths were much brighter and non-specific bands were reduced. In conclusion, using a combination of DMSO and BSA yielded the clearest results, and so these additives were included in all PCR reactions moving forwards.

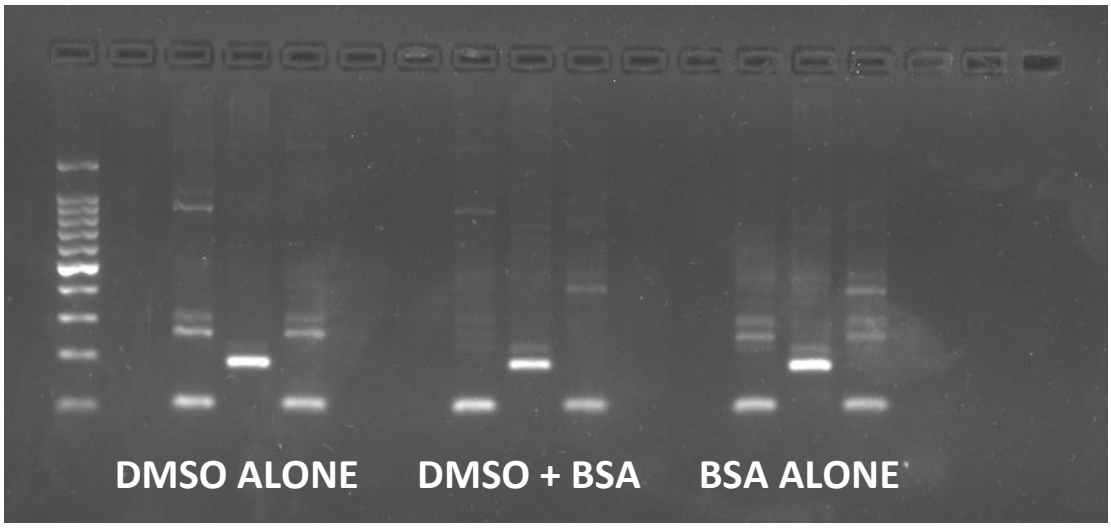
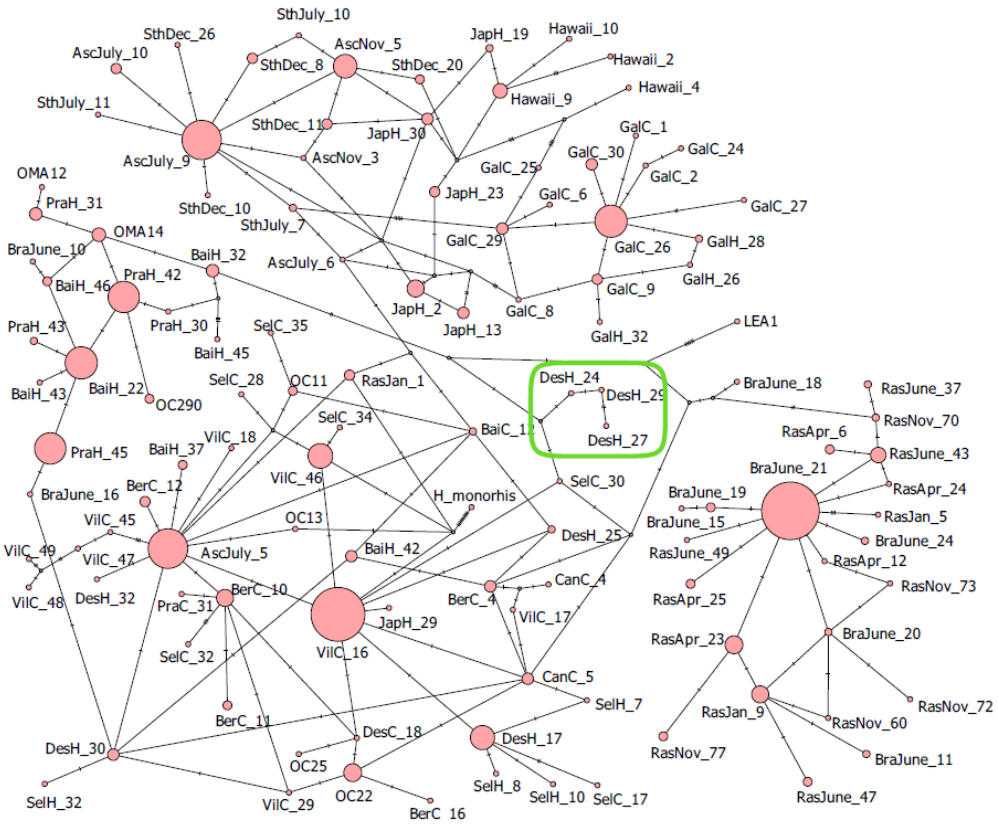



**Figure A2.1:** Gel electrophoresis image showing a direct comparison between the use of DMSO alone, BSA alone, and DMSO/BSA combined. The same 3 samples were used each time, and all reactions were run in the same PCR. PCR reagents remained the same, except for differences in BSA and DMSO amount. A combination of using BSA and DMSO produces the best results. Less ambiguous bands are present, and those that are visible are very faint in comparison to the target sequences. A 100 bp Promega ladder was used as a reference.

## A 2.2: Haplotype network of mtDNA sequences

Using DNAsp, a haplotype network of all 848 sequences was created. This demonstrated how 3 *H. castro* samples from Desertas, Madeira, may be clustering in a separate phylogenetic clade to other samples from the same region.




**S???:** This haplotype network was created in DNAsp, using an alignment of 848 mtDNA sequences. Data used included both previously published data, and new Azores samples extracted for this thesis. The separate clustering of 3 Madeiran haplotypes can be seen in the circled area. It is possible that the small number of mutation steps away from the two outgroups (LEA1 and H_monorhis_KR73325.1_reversed_) are causing this separation observed in the phylogenetic tree. Overall, haplotypes appear to cluster into groups also observed in the phylogenetic tree.

## A 2.3: Samples screened using clade-specific primers

Here, all samples screened using the clade specific primers are listed. 'Registered Species' details the species assigned to the sample and was used to determine if the primers had amplified as expected or not. *monteiroi-clade* and castro-clade refer to the clade-specific primer sets; 'N' indicates the primer set did not work for a particular sample, whilst 'Y' indicates successful amplification. Samples

beginning 'OC…' and 'OMA…' relate to samples collected from the Azores and extracted in the lab. All **lab** samples amplified as expected – samples registered as *H. castro* only amplified with castro-clade, whilst samples registered as *H. monteiroi* only amplified with *monteiroi-clade*. Leach's storm petrel (*H. leucorhous)* and European storm petrel (*H. pelagicus*) did not amplify with either primer set. *H. castro* samples from St Helena and Ascension Island (beginning 'St Helena' and 'SEBT/WEBT', respectively) did amplify with castro-clade, however some samples failed to amplify entirely. It is possible that with optimisation of PCR conditions, these primers could work better on samples from these locations.

**Table A 2.1:** Screening results from using *monteiroi-clade* and castro-clade in a multiplexed PCR reaction. Overall, the clade-specific primers are successful, amplifying the species for which each pair was designed only.

| Sample | Registered | *monteiroi-* | castro-clade |
|---|---|---|---|
| LHSP79537 | Leach | N | N |
| N06779 | European | N | N |
| OC001 | Castro | N | Y |
| OC002 | Castro | N | Y |
| OC003 | Castro | N | Y |
| OC004 | Castro | N | Y |
| OC005 | Castro | N | Y |
| OC006 | Castro | N | Y |
| OC007 | Castro | N | Y |
| OC008 | Castro | N | Y |
| OC009 | Castro | N | Y |
| OC010 | Castro | N | Y |
| OC011 | Castro | N | Y |
| OC012 | Castro | N | Y |
| OC013 | Castro | N | Y |
| OC014 | Castro | N | Y |
| OC015 | Castro | N | Y |
| OC016 | Castro | N | Y |
| OC017 | Castro | N | Y |
| OC018 | Castro | N | Y |
| OC019 | Castro | N | Y |
| OC020 | Castro | N | Y |
| OC021 | Castro | N | Y |
| OC022 | Castro | N | Y |
| OC023 | Castro | N | Y |
| OC024 | Castro | N | Y |
| OC025 | Castro | N | Y |
| OC026 | Castro | N | Y |
| OC030 | Castro | N | Y |
| OC031 | Castro | N | Y |
| OC033 | Castro | N | Y |

| | | | |
|---|---|---|---|
| OC034 | Castro | N | Y |
| OC035 | Castro | N | Y |
| OC039 | Castro | N | Y |
| OC040 | Castro | N | Y |
| OC041 | Castro | N | Y |
| OC042 | Castro | N | Y |
| OC052 | Castro | N | Y |
| OC053 | Castro | N | Y |
| OC054 | Castro | N | Y |
| OC055 | Castro | N | Y |
| OC058 | Castro | N | Y |
| OC059 | Castro | N | Y |
| OC061 | Castro | N | Y |
| OC064 | Castro | N | Y |
| OC066 | Castro | N | Y |
| OC067 | Castro | N | Y |
| OC071 | Castro | N | Y |
| OC072 | Castro | N | Y |
| OC073 | Castro | N | Y |
| OC078 | Castro | N | Y |
| OC080 | Castro | N | Y |
| OC085 | Castro | N | Y |
| OC086 | Castro | N | Y |
| OC087 | Castro | N | Y |
| OC088 | Castro | N | Y |
| OC090 | Castro | N | Y |
| OC097 | Castro | N | Y |
| OC098 | Castro | N | Y |
| OC099 | Castro | N | Y |
| OC101 | Castro | N | Y |
| OC102 | Castro | N | Y |
| OC103 | Castro | N | Y |
| OC104 | Castro | N | Y |
| OC110 | Castro | N | Y |
| OC112 | Castro | N | Y |
| OC113 | Castro | N | Y |
| OC116 | Castro | N | Y |
| OC118 | Castro | N | Y |
| OC119 | Castro | N | Y |
| OC120 | Castro | N | Y |
| OC124 | Castro | N | Y |
| OC128 | Castro | N | Y |
| OC130 | Castro | N | Y |
| OC131 | Castro | N | Y |
| OC132 | Castro | N | Y |
| OC135 | Castro | N | Y |
| OC136 | Castro | N | Y |
| OC137 | Castro | N | Y |
| OC139 | Castro | N | Y |
| OC140 | Castro | N | Y |
| OC141 | Castro | N | Y |

| | | | |
|---|---|---|---|
| OC142 | Castro | N | Y |
| OC144 | Castro | N | Y |
| OC145 | Castro | N | Y |
| OC147 | Castro | N | Y |
| OC149 | Castro | N | Y |
| OC150 | Castro | N | Y |
| OC151 | Castro | N | Y |
| OC153 | Castro | N | Y |
| OC154 | Castro | N | Y |
| OC164 | Castro | N | Y |
| OC165 | Castro | N | Y |
| OC166 | Castro | N | Y |
| OC167 | Castro | N | Y |
| OC168 | Castro | N | Y |
| OC169 | Castro | N | Y |
| OC170 | Castro | N | Y |
| OC172 | Castro | N | Y |
| OC173 | Castro | N | Y |
| OC174 | Castro | N | Y |
| OC175 | Castro | N | Y |
| OC176 | Castro | N | Y |
| OC177 | Castro | N | Y |
| OC178 | Castro | N | Y |
| OC179 | Castro | N | Y |
| OC182 | Castro | N | Y |
| OC184 | Castro | N | Y |
| OC185 | Castro | N | Y |
| OC186 | Castro | N | Y |
| OC188 | Castro | N | Y |
| OC189 | Castro | N | Y |
| OC190 | Castro | N | Y |
| OC191 | Castro | N | Y |
| OC192 | Castro | N | Y |
| OC193 | Castro | N | Y |
| OC196 | Castro | N | Y |
| OC197 | Castro | N | Y |
| OC198 | Castro | N | Y |
| OC199 | Castro | N | Y |
| OC201 | Castro | N | Y |
| OC202 | Castro | N | Y |
| OC204 | Castro | N | Y |
| OC208 | Castro | N | Y |
| OC209 | Castro | N | Y |
| OC210 | Castro | N | Y |
| OC211 | Castro | N | Y |
| OC212 | Castro | N | Y |
| OC214 | Castro | N | Y |
| OC215 | Castro | N | Y |
| OC217 | Monteiro | Y | N |
| OC224 | Monteiro | Y | N |
| OC225 | Monteiro | Y | N |

| | | | |
|---|---|---|---|
| OC228 | Monteiro | Y | N |
| OC232 | Monteiro | Y | N |
| OC233 | Monteiro | Y | N |
| OC237 | Monteiro | Y | N |
| OC238 | Monteiro | Y | N |
| OC239 | Monteiro | Y | N |
| OC242 | Monteiro | Y | N |
| OC246 | Monteiro | Y | N |
| OC247 | Monteiro | Y | N |
| OC248 | Monteiro | Y | N |
| OC249 | Monteiro | Y | N |
| OC252 | Monteiro | Y | N |
| OC255 | Monteiro | Y | N |
| OC258 | Monteiro | Y | N |
| OC259 | Monteiro | Y | N |
| OC263 | Monteiro | Y | N |
| OC265 | Monteiro | Y | N |
| OC266/348 | Monteiro | Y | N |
| OC267 | Monteiro | Y | N |
| OC268 | Monteiro | Y | N |
| OC271 | Monteiro | Y | N |
| OC273 | Monteiro | Y | N |
| OC275 | Monteiro | Y | N |
| OC278 | Monteiro | Y | N |
| OC279 | Monteiro | Y | N |
| OC286 | Monteiro | Y | N |
| OC288 | Monteiro | Y | N |
| OC289 | Monteiro | Y | N |
| OC293 | Monteiro | Y | N |
| OC294 | Monteiro | Y | N |
| OC297 | Monteiro | Y | N |
| OC299 | Monteiro | Y | N |
| OC300 | Monteiro | Y | N |
| OC301 | Monteiro | Y | N |
| OC311 | Monteiro | Y | N |
| OC315 | Monteiro | Y | N |
| OC316 | Monteiro | Y | N |
| OC317 | Monteiro | Y | N |
| OC322 | Monteiro | Y | N |
| OC325 | Monteiro | Y | N |
| OC328 | Monteiro | Y | N |
| OC329 | Monteiro | Y | N |
| OC335 | Monteiro | Y | N |
| OC347 | Monteiro | Y | N |
| OC350 | Monteiro | Y | N |
| OC365 | Monteiro | Y | N |
| OC367 | Monteiro | Y | N |
| OC369 | Monteiro | Y | N |
| OC370 | Monteiro | Y | N |
| OC371 | Monteiro | Y | N |
| OC373 | Monteiro | Y | N |

| | | | |
|---|---|---|---|
| OC375 | Monteiro | Y | N |
| OC377 | Monteiro | Y | N |
| OC378 | Monteiro | Y | N |
| OC379 | Monteiro | Y | N |
| OC379 | Monteiro | Y | N |
| OC381 | Monteiro | Y | N |
| OC382 | Monteiro | Y | N |
| OC383 | Monteiro | Y | N |
| OC384 | Monteiro | Y | N |
| OC385 | Monteiro | Y | N |
| OC389 | Monteiro | Y | N |
| OC390 | Monteiro | Y | N |
| OC391 | Monteiro | Y | N |
| OC395 | Monteiro | Y | N |
| OC397 | Monteiro | Y | N |
| OC398 | Monteiro | Y | N |
| OC403 | Monteiro | Y | N |
| OC407 | Monteiro | Y | N |
| OC409 | Monteiro | Y | N |
| OC410 | Monteiro | Y | N |
| OC413 | Monteiro | Y | N |
| OC414 | Monteiro | Y | N |
| OC418/387 | Monteiro | Y | N |
| OC420 | Monteiro | Y | N |
| OMA1 | Monteiro | Y | N |
| OMA10 | Monteiro | Y | N |
| OMA11 | Monteiro | Y | N |
| OMA12 | Monteiro | Y | N |
| OMA14 | Monteiro | Y | N |
| OMA2 | Monteiro | Y | N |
| OMA22 | Monteiro | Y | N |
| OMA23 | Monteiro | Y | N |
| OMA26 | Monteiro | Y | N |
| OMA3 | Monteiro | Y | N |
| OMA30 | Monteiro | Y | N |
| OMA33 | Monteiro | Y | N |
| OMA35 | Monteiro | Y | N |
| OMA37 | Monteiro | Y | N |
| OMA4 | Monteiro | Y | N |
| OMA41 | Monteiro | Y | N |
| OMA42 | Monteiro | Y | N |
| OMA5 | Monteiro | - | - |
| OMA6 | Monteiro | Y | N |
| OMA7 | Monteiro | Y | N |
| OMA8 | Monteiro | - | - |
| OMA9 | Monteiro | Y | N |
| OC219 | Monteiro | Y | N |
| OC220 | Monteiro | Y | N |
| OC223 | Monteiro | Y | N |
| OC229 | Monteiro | Y | N |
| OC235 | Monteiro | Y | N |

| | | | |
|---|---|---|---|
| OC243 | Monteiro | Y | N |
| OC244 | Monteiro | Y | N |
| OC245 | Monteiro | Y | N |
| OC251 | Monteiro | Y | N |
| OC253 | Monteiro | Y | N |
| OC270 | Monteiro | Y | N |
| OC272 | Monteiro | Y | N |
| OC283 | Monteiro | Y | N |
| OC285 | Monteiro | Y | N |
| St. Helena 1 | Castro | - | - |
| St. Helena 2 | Castro | N | Y |
| St. Helena 3 | Castro | - | - |
| SEBT6198038 | Castro | N | Y |
| SEBT6234040 | Castro | N | Y |
| WEBT62773 | Castro | - | - |
| WEBT62775 | Castro | N | Y |

# Appendix for Chapter 3

## A3.1: IQTree Scripts

```
### Navigate to iqtree programme folder ###
cd C:\iqtree-1.6.12-Windows


### Run ModelFinder for best-fit model ###
bin\iqtree -s FILENAME.phy -m MF

### Run a Kimura 2-parameter model with a gamma rate of 4 & 200
bootstraps ####

bin\iqtree -s FILENAME.phy -m K2P+G4 -b 200
```

# Appendix to Chapter 4

## A.4.1: Alterations Made to the SPRIselect Guide

The standard SPRISelect guide was followed, aside for the following modifications:

(i) Incubation steps were held for 5 minutes instead of the suggested 1 minute, to ensure beads mixed and settled appropriately.

(ii) The ethanol wash in step 4 was carried out twice, and at sufficient volume to fully submerge the beads present in the eppendorf. This extra wash ensured the beads were cleaned sufficiently, and the change in volume assured all beads were covered by the wash.

(iii) After the final ethanol wash, the SPRI beads were air-dried for 1 minute.

## A.4.2: Calculations for creating DAB1 and DAB2 'Super Pools'

Calculations for pooling of samples was done as follows. This example uses values from one of the pools, to make calculations clear:

(i) First, the concentration of the pool with the highest concentration was divided by the volume of that pool. Each pool volume reflected the number of samples present. This calculates the concentration of 1 µL.

*e.g.* ***59 ng/µL ÷ 27 µL = 2.18 ng in 1 µL***

(ii) These values were multiplied, to give a larger volume suitable for pipetting into a super pool:

*e.g.* ***3 × 1 = 3 µL  3 × 2.18 = 6.54 ng → In 3 µL there is 6.54 ng***

(iii) For the remaining pools, their concentration value was also divided by the volume, to calculate the concentration of 1 µL from that pool.

*e.g.* ***47.5 ng/µL ÷ 29 µL = 1.6 ng in 1 µL***

(iv) Next, the number calculated in (i) – the concentration of 1 µL from the most concentrated pool – was taken and divided by the concentration of the current pool. This would calculate the amount needed of the current pool, to match the concentration of 1 µL from the highest pool.

*e.g.* ***2.18 ÷ 1.6 = 1.365 µL*** *from this pool would contain* ***2.18 ng***

(v) The final step was to multiply this volume by the same amount as done in step (ii), in this case 3, to pipette the appropriate volume.

*e.g.* ***1.365 × 3 = 4.095 µL*** *to pipette, for a concentration of* ***6.54ng***

These steps were done for each pool, and the end volumes calculated were combined into a super pool for each DAB lineage.

## A.4.3: Dilution Calculations for 4nm Illumina Libraries

To dilute to 4nm:

(i) The total concentration of the library was first figured out in nm. To do this, the Qubit concentration and Tapestation average bp length were inputted to the following calculation:

$$\frac{\textit{Qubit Concentration in ng/ µL}}{\textit{(660 g/mol × library bp size)}} \times 10^6 = \textit{concentration in nM}$$

For DAB1 the Qubit concentration was 9.88 ng/µL and fragment length was 496 bp

The molarity of the pool worked out at 30.2 nm

For DAB2, the Qubit concentration was 16.3 ng/µL and fragment length was 483 bp

Here, the molarity worked out at 51.1 nm

(ii) Next, the molarity of each DAB library was divided by 4 nm, to calculate the dilution factor and therefore volume of water needed to dilute each library. For DAB1 the dilution factor was 7.55 and for DAB 2, it was 12.76.

(iii) It was decided that 5 µL of each library would be used in the final dilution. To calculate water needed for dilution, 5 µL was multiplied by the dilution factor calculated in (ii). This value then had 5 subtracted (to account for the 5 µL library volume) to give the exact volume of water needed to dilute to 4 nm. For DAB1 a 4nm Illumina sample would contain 5 µL library and 32.75 µL water. For DAB2, to achieve a 4nm dilution it would require 5 µL library and 58.8 µL water.

## A.4.4: Bioinformatics Script for Illumina data

**### SCRIPT WRITTEN BY LORNA DRAKE, CARDIFF UNIVERSITY ###**

```
### Gunzip to get fastq files

gunzip Locus1F.fastq.gz
gunzip Locus1R.fastq.gz


gunzip Locus2F.fastq.gz
gunzip Locus2R.fastq.gz


### GREP TO EXTRACT AND QUANTIFY READS WITH 8BP MID-TAGS
PRESENT ###

grep "GACCTCTCCATGTYTGCACGAR" Locus1F.fastq > L1Read1_For_all.txt
grep "^........GACCTCTCCATGTYTGCACGAR" L1Read1_For_all.txt >
L1Read1_For_8.txt
grep -v "^........GACCTCTCCATGTYTGCACGAR" L1Read1_For_all.txt >
L1Read1_For_trn.txt
wc – l L1Read1_For_*


grep "GCAATGTTCTGCCMAGCACT" Locus1F.fastq > L1Read1_Rev_all.txt
grep "^........GCAATGTTCTGCCMAGCACT" L1Read1_Rev_all.txt > L1Read1_Rev_8.txt
grep -v "^........GCAATGTTCTGCCMAGCACT" L1Read1_Rev_all.txt >
L1Read1_Rev_trn.txt
wc – l L1Read1_Rev_*


grep "GACCTCTCCATGTYTGCACGAR" Locus1R.fastq > L1Read2_For_all.txt
grep "^........GACCTCTCCATGTYTGCACGAR" L1Read2_For_all.txt >
L1Read2_For_8.txt
grep -v "^........GACCTCTCCATGTYTGCACGAR" L1Read2_For_all.txt >
L1Read2_For_trn.txt
wc – l L1Read1_For_*


grep "GCAATGTTCTGCCMAGCACT" Locus1R.fastq > L1Read2_Rev_all.txt
grep "^........GCAATGTTCTGCCMAGCACT" L1Read2_Rev_all.txt > L1Read2_Rev_8.txt
grep -v "^........GCAATGTTCTGCCMAGCACT" L1Read2_Rev_all.txt >
L1Read2_Rev_trn.txt
wc – l L1Read2_Rev_*



grep "GACCTGCCTCCCTGCACAAACA" Locus2F.fastq > L2Read1_For_all.txt
grep "^........GACCTGCCTCCCTGCACAAACA" L2Read1_For_all.txt >
L2Read1_For_8.txt
grep -v "^........GACCTGCCTCCCTGCACAAACA" L2Read1_For_all.txt >
L2Read1_For_trn.txt
wc – l L2Read1_For_*


grep "GCAATGTTCTGCCMAGCACT" Locus2F.fastq > L2Read1_Rev_all.txt
grep "^........GCAATGTTCTGCCMAGCACT" L2Read1_Rev_all.txt > L2Read1_Rev_8.txt
```

```
grep -v "^........GCAATGTTCTGCCMAGCACT" L2Read1_Rev_all.txt >
L2Read1_Rev_trn.txt
wc – l L2Read1_Rev_*


grep "GACCTGCCTCCCTGCACAAACA" Locus2R.fastq > L2Read2_For_all.txt
grep "^........GACCTGCCTCCCTGCACAAACA" L2Read2_For_all.txt >
L2Read2_For_8.txt
grep -v "^........GACCTGCCTCCCTGCACAAACA" L2Read2_For_all.txt >
L2Read2_For_trn.txt
wc – l L2Read1_For_*


grep "GCAATGTTCTGCCMAGCACT" Locus2R.fastq > L2Read2_Rev_all.txt
grep "^........GCAATGTTCTGCCMAGCACT" L2Read2_Rev_all.txt > L2Read2_Rev_8.txt
grep -v "^........GCAATGTTCTGCCMAGCACT" L2Read2_Rev_all.txt >
L2Read2_Rev_trn.txt
wc – l L2Read2_Rev_*
```

### FASTP TO TRIM, ALIGN AND QUALITY CHECK READS ###

### SHELL SCRIPT USED TO CARRY OUT FASTP; THE SHELL SCRIPT WAS AS FOLLOWS:

```
#!/bin/bash

##LOCUS 1
#SBATCH --partition=mammoth
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=1000000
#SBATCH --error=FastPL1.err
#SBATCH --output=FastPL1.out
#SBATCH --job-name=FastPL1

## First do a FastQC quality check, merge the paired end reads and trim
sequences using fastp

/mnt/scratch/amc281/L1/fastp -i Locus1F.fastq -I Locus1R.fastq -l 250 -m --
discard_unmerged -o mergedL1.fastq

## Next convert the fastq file to a fasta format

module load fastx_toolkit/0.0.14

fastq_to_fasta -i mergedL1.fastq -Q 30 -o mergedL1.fa


#!/bin/bash

##LOCUS 2
#SBATCH --partition=mammoth
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=1000000
#SBATCH --error=FastPL2.err
#SBATCH --output=FastPL2.out
#SBATCH --job-name=FastPL2
```

```
## First do a FastQC quality check, merge the paired end reads and trim
sequences using fastp

/mnt/scratch/amc281/L2/fastp -i Locus2F.fastq -I Locus2R.fastq -l 250 -m --
discard_unmerged -o mergedL2.fastq

## Next convert the fastq file to a fasta format

module load fastx_toolkit/0.0.14

fastq_to_fasta -i mergedL2.fastq -Q 30 -o mergedL2.fa
```

### A.4.5: BLAST Script for Confirming Species

All BLAST analyses were carried out using shell scripts on the Cardiff University Bioinformatics servers. All shell scripts took on the following format, with changes made to the parts highlighted in yellow and formatted **bold**.

```
#!/bin/bash
#SBATCH --partition=defq
#SBATCH --error=blastL285_titles.err
#SBATCH --output=blastL285_titles.out
#SBATCH --job-name=BL285titles
#SBATCH --mail-type=end
#SBATCH --mail-type=fail


# blast the clusters from usearch

module load blast/2.10.0
export BLASTDB=/mnt/scratch/amc281/BLASTDBALL   (changed to 'MYDB' for DAB2)


blastn
-query L2_AL.fasta  -db nt   -num_threads 4    -evalue 0.00001
-perc_identity 85   -outfmt "7 qseqid sacc staxids sscinames
stitle"      -max_target_seqs 3  -out L2_85_titles.txt
```

The '**-query**' was changed to the relevant DAB fasta file each time. DAB1 and DAB2 were each run at 85, 90 and 90% percent identity, with '**-perc_identity**' changed each time to reflect this. The '**-out**' file name was changed each time according to the relevant DAB lineage and percentage ID used.

## A.4.6: Samples and Genotypes

The genotypes for each sample can be seen in the tables below

Table A4.6.1: DAB1 genotypes for all storm petrel samples

| DAB Name | Genotype | | | | | |
|---|---|---|---|---|---|---|
| **OC003a** | DAB1*02 | DAB1*16 | DAB1*32 | DAB1*38 | | |
| **OC004a** | DAB1*02 | DAB1*16 | DAB1*38 | | | |
| **OC005a** | DAB1*02 | | | | | |
| **OC007a** | DAB1*17 | DAB1*12 | | | | |
| **OC008a** | DAB1*02 | DAB1*10 | DAB1*25 | | | |
| **OC009a** | DAB1*02 | DAB1*28 | | | | |
| **OC010a** | DAB1*05 | DAB1*08 | DAB1*62 | | | |
| **OC011a** | DAB1*02 | DAB1*03 | DAB1*17 | | | |
| **OC012a** | DAB1*05 | DAB1*08 | | | | |
| **OC013a** | DAB1*02 | DAB1*10 | DAB1*17 | | | |
| **OC014a** | DAB1*02 | DAB1*03 | | | | |
| **OC016a** | DAB1*02 | DAB1*17 | DAB1*12 | | | |
| **OC017a** | DAB1*02 | DAB1*05 | DAB1*08 | | | |
| **OC018a** | DAB1*05 | DAB1*10 | | | | |
| **OC021a** | DAB1*10 | DAB1*25 | DAB1*28 | | | |
| **OC022a** | DAB1*02 | DAB1*05 | DAB1*08 | | | |
| **OC023a** | DAB1*03 | DAB1*34 | | | | |
| **OC024a** | DAB1*05 | DAB1*08 | | | | |
| **OC025a** | DAB1*03 | DAB1*05 | DAB1*08 | | | |
| **OC026a** | DAB1*05 | DAB1*32 | DAB1*38 | | | |
| **OC028a** | DAB1*02 | DAB1*14 | | | | |
| **OC030a** | DAB1*10 | | | | | |
| **OC031a** | DAB1*05 | | | | | |
| **OC033a** | DAB1*02 | DAB1*10 | DAB1*25 | DAB1*28 | | |
| **OC034a** | DAB1*02 | DAB1*17 | DAB1*40 | | | |
| **OC035a** | DAB1*02 | DAB1*10 | DAB1*17 | DAB1*25 | | |
| **OC036a** | DAB1*02 | DAB1*10 | DAB1*17 | DAB1*25 | DAB1*51 | |
| **OC039a** | DAB1*02 | DAB1*10 | DAB1*25 | DAB1*28 | | |
| **OC040a** | DAB1*02 | DAB1*10 | | | | |
| **OC041a** | DAB1*02 | DAB1*10 | DAB1*17 | DAB1*25 | | |
| **OC042a** | DAB1*02 | DAB1*17 | DAB1*51 | | | |
| **OC045a** | DAB1*03 | DAB1*22 | | | | |
| **OC050a** | DAB1*02 | DAB1*03 | DAB1*22 | | | |
| **OC052a** | DAB1*02 | DAB1*03 | DAB1*05 | DAB1*08 | | |
| **OC053a** | DAB1*02 | DAB1*03 | | | | |
| **OC055a** | DAB1*02 | DAB1*05 | DAB1*08 | | | |
| **OC058a** | DAB1*02 | DAB1*03 | | | | |
| **OC059a** | DAB1*02 | DAB1*03 | DAB1*16 | | | |
| **OC061a** | DAB1*02 | DAB1*03 | DAB1*05 | | | |
| **OC064a** | DAB1*02 | DAB1*10 | DAB1*17 | DAB1*25 | DAB1*58 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **OC066a** | DAB1*10 | | | | | |
| **OC067a** | DAB1*02 | DAB1*30 | DAB1*40 | | | |
| **OC071a** | DAB1*03 | DAB1*05 | DAB1*08 | DAB1*17 | | |
| **OC072a** | DAB1*02 | DAB1*14 | | | | |
| **OC073a** | DAB1*03 | DAB1*10 | DAB1*34 | | | |
| **OC078a** | DAB1*10 | DAB1*22 | | | | |
| **OC080a** | DAB1*02 | DAB1*03 | | | | |
| **OC082a** | DAB1*03 | DAB1*22 | | | | |
| **OC085a** | DAB1*05 | DAB1*11 | DAB1*14 | | | |
| **OC086a** | DAB1*02 | DAB1*05 | | | | |
| **OC088a** | DAB1*02 | DAB1*10 | DAB1*14 | | | |
| **OC090a** | DAB1*02 | DAB1*10 | DAB1*30 | | | |
| **OC091a** | DAB1*02 | DAB1*03 | | | | |
| **OC097a** | DAB1*03 | DAB1*34 | | | | |
| **OC098a** | DAB1*05 | DAB1*12 | | | | |
| **OC099a** | DAB1*03 | DAB1*10 | DAB1*16 | DAB1*30 | | |
| **OC101a** | DAB1*03 | | | | | |
| **OC102a** | DAB1*02 | DAB1*03 | DAB1*32 | | | |
| **OC104a** | DAB1*03 | DAB1*20 | | | | |
| **OC110a** | DAB1*20 | DAB1*58 | | | | |
| **OC112a** | DAB1*02 | | | | | |
| **OC113a** | DAB1*12 | | | | | |
| **OC116a** | DAB1*02 | DAB1*17 | | | | |
| **OC118a** | DAB1*03 | DAB1*17 | | | | |
| **OC119a** | DAB1*03 | | | | | |
| **OC120a** | DAB1*20 | DAB1*63 | | | | |
| **OC124a** | DAB1*03 | DAB1*05 | | | | |
| **OC128a** | DAB1*03 | | | | | |
| **OC130a** | DAB1*05 | DAB1*53 | | | | |
| **OC131a** | DAB1*02 | DAB1*03 | | | | |
| **OC132a** | DAB1*02 | DAB1*14 | | | | |
| **OC135a** | DAB1*12 | DAB1*16 | | | | |
| **OC136a** | DAB1*02 | DAB1*16 | | | | |
| **OC137a** | DAB1*05 | DAB1*14 | | | | |
| **OC139a** | DAB1*02 | DAB1*20 | | | | |
| **OC140a** | DAB1*03 | | | | | |
| **OC141a** | DAB1*02 | DAB1*32 | | | | |
| **OC142a** | DAB1*03 | DAB1*09 | DAB1*21 | DAB1*29 | | |
| **OC145a** | DAB1*02 | DAB1*05 | DAB1*11 | | | |
| **OC147a** | DAB1*05 | | | | | |
| **OC149a** | DAB1*02 | DAB1*20 | | | | |
| **OC150a** | DAB1*03 | | | | | |
| **OC151a** | DAB1*02 | DAB1*12 | DAB1*16 | DAB1*32 | | |
| **OC153a** | DAB1*03 | DAB1*05 | | | | |
| **OC154a** | DAB1*03 | DAB1*05 | | | | |
| **OC164a** | DAB1*02 | DAB1*03 | | | | |
| **OC165a** | DAB1*03 | DAB1*12 | DAB1*16 | | | |
| **OC166a** | DAB1*03 | DAB1*32 | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **OC167a** | DAB1*02 | DAB1*20 | DAB1*38 | | | |
| **OC168a** | DAB1*03 | | | | | |
| **OC170a** | DAB1*05 | | | | | |
| **OC172a** | DAB1*02 | DAB1*20 | | | | |
| **OC173a** | DAB1*03 | | | | | |
| **OC174a** | DAB1*12 | DAB1*16 | DAB1*49 | | | |
| **OC175a** | DAB1*02 | DAB1*12 | | | | |
| **OC176a** | DAB1*02 | DAB1*51 | DAB1*40 | | | |
| **OC178a** | DAB1*20 | DAB1*53 | | | | |
| **OC179a** | DAB1*05 | DAB1*11 | DAB1*32 | DAB1*33 | | |
| **OC182a** | DAB1*03 | | | | | |
| **OC184a** | DAB1*03 | DAB1*34 | | | | |
| **OC185a** | DAB1*03 | | | | | |
| **OC186a** | DAB1*10 | DAB1*12 | DAB1*16 | DAB1*25 | | |
| **OC188a** | DAB1*03 | | | | | |
| **OC189a** | DAB1*03 | DAB1*34 | | | | |
| **OC190a** | DAB1*03 | | | | | |
| **OC191a** | DAB1*02 | DAB1*03 | DAB1*12 | DAB1*64 | | |
| **OC192a** | DAB1*03 | DAB1*14 | | | | |
| **OC193a** | DAB1*03 | DAB1*22 | DAB1*30 | | | |
| **OC196a** | DAB1*03 | DAB1*34 | | | | |
| **OC199a** | DAB1*02 | DAB1*05 | DAB1*08 | | | |
| **OC202a** | DAB1*02 | DAB1*17 | DAB1*14 | DAB1*51 | | |
| **OC214a** | DAB1*05 | DAB1*12 | | | | |
| **OC216a** | DAB1*01 | DAB1*47 | | | | |
| **OC217a** | DAB1*01 | DAB1*04 | | | | |
| **OC218a** | DAB1*01 | DAB1*09 | DAB1*39 | DAB1*41 | | |
| **OC219a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*27 | | |
| **OC220a** | DAB1*01 | DAB1*04 | DAB1*55 | | | |
| **OC221a** | DAB1*01 | DAB1*24 | | | | |
| **OC223a** | DAB1*01 | | | | | |
| **OC224a** | DAB1*07 | DAB1*15 | | | | |
| **OC225a** | DAB1*09 | DAB1*08 | DAB1*31 | DAB1*41 | | |
| **OC228a** | DAB1*04 | DAB1*26 | DAB1*33 | | | |
| **OC229a** | DAB1*01 | DAB1*06 | DAB1*07 | DAB1*44 | DAB1*42 | |
| **OC231a** | DAB1*06 | DAB1*27 | DAB1*24 | | | |
| **OC232a** | DAB1*01 | DAB1*06 | DAB1*07 | DAB1*45 | | |
| **OC233a** | DAB1*01 | DAB1*39 | | | | |
| **OC235a** | DAB1*01 | DAB1*04 | DAB1*09 | | | |
| **OC237a** | DAB1*01 | DAB1*04 | DAB1*27 | | | |
| **OC238a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*60 | | |
| **OC239a** | DAB1*52 | DAB1*54 | DAB1*56 | | | |
| **OC242a** | DAB1*04 | | | | | |
| **OC243a** | DAB1*01 | DAB1*11 | | | | |
| **OC244a** | DAB1*01 | DAB1*07 | DAB1*44 | DAB1*42 | DAB1*48 | |
| **OC245a** | DAB1*04 | DAB1*08 | DAB1*21 | DAB1*26 | DAB1*33 | |
| **OC246a** | DAB1*08 | DAB1*06 | DAB1*07 | DAB1*21 | DAB1*45 | |
| **OC247a** | DAB1*26 | DAB1*18 | DAB1*61 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **OC248a** | DAB1*04 | DAB1*06 | DAB1*13 | DAB1*35 | | |
| **OC249a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*13 | | |
| **OC250a** | DAB1*01 | DAB1*09 | | | | |
| **OC251a** | DAB1*01 | DAB1*04 | DAB1*05 | DAB1*09 | | |
| **OC252a** | DAB1*06 | DAB1*27 | DAB1*36 | | | |
| **OC253a** | DAB1*06 | DAB1*27 | DAB1*23 | DAB1*19 | | |
| **OC255a** | DAB1*04 | DAB1*26 | DAB1*33 | | | |
| **OC258a** | DAB1*08 | DAB1*11 | DAB1*36 | DAB1*37 | | |
| **OC261a** | DAB1*04 | DAB1*06 | DAB1*13 | DAB1*35 | | |
| **OC262a** | DAB1*09 | DAB1*06 | DAB1*07 | DAB1*21 | DAB1*29 | |
| **OC263a** | DAB1*04 | DAB1*13 | DAB1*26 | DAB1*33 | | |
| **OC265a** | DAB1*01 | | | | | |
| **OC266a** | DAB1*01 | | | | | |
| **OC267a** | DAB1*07 | DAB1*44 | DAB1*42 | DAB1*48 | | |
| **OC268a** | DAB1*01 | | | | | |
| **OC269a** | DAB1*05 | DAB1*09 | | | | |
| **OC270a** | DAB1*04 | DAB1*09 | | | | |
| **OC271a** | DAB1*01 | DAB1*09 | | | | |
| **OC272a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*11 | | |
| **OC273a** | DAB1*11 | DAB1*26 | | | | |
| **OC275a** | DAB1*04 | DAB1*55 | | | | |
| **OC276a** | DAB1*01 | DAB1*05 | DAB1*08 | | | |
| **OC277a** | DAB1*01 | DAB1*08 | DAB1*31 | | | |
| **OC278a** | DAB1*01 | DAB1*04 | DAB1*27 | | | |
| **OC279a** | DAB1*01 | DAB1*23 | | | | |
| **OC283a** | DAB1*07 | DAB1*15 | DAB1*24 | | | |
| **OC284a** | DAB1*01 | DAB1*39 | | | | |
| **OC285a** | DAB1*06 | DAB1*35 | | | | |
| **OC286a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*06 | DAB1*13 | |
| **OC287a** | DAB1*01 | DAB1*07 | | | | |
| **OC288a** | DAB1*07 | DAB1*23 | DAB1*15 | DAB1*19 | | |
| **OC289a** | DAB1*04 | DAB1*09 | DAB1*13 | DAB1*50 | | |
| **OC290a** | DAB1*08 | DAB1*06 | DAB1*11 | DAB1*43 | | |
| **OC293a** | DAB1*04 | DAB1*09 | DAB1*18 | | | |
| **OC295a** | DAB1*04 | DAB1*09 | DAB1*21 | DAB1*29 | | |
| **OC296a** | DAB1*01 | DAB1*04 | DAB1*09 | DAB1*46 | | |
| **OC297a** | DAB1*01 | | | | | |
| **OC298a** | DAB1*09 | DAB1*07 | DAB1*21 | DAB1*29 | DAB1*44 | DAB1*42 |
| **OC299a** | DAB1*11 | DAB1*23 | DAB1*19 | | | |
| **OC300a** | DAB1*01 | DAB1*04 | DAB1*29 | DAB1*65 | | |
| **OC301a** | DAB1*01 | | | | | |
| **OC310a** | DAB1*01 | DAB1*23 | DAB1*19 | | | |
| **OC314a** | DAB1*01 | DAB1*06 | DAB1*07 | | | |
| **OC316a** | DAB1*01 | DAB1*11 | | | | |
| **OC337a** | DAB1*04 | DAB1*09 | DAB1*06 | DAB1*13 | DAB1*23 | |
| **OC338a** | DAB1*01 | DAB1*07 | DAB1*27 | DAB1*15 | DAB1*59 | |
| **OC340a** | DAB1*01 | | | | | |
| **OC341a** | DAB1*07 | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **OC343a** | DAB1*04 | DAB1*09 | DAB1*06 | DAB1*07 | | |
| **OC351a** | DAB1*06 | DAB1*43 | | | | |
| **OC367a** | DAB1*01 | DAB1*04 | DAB1*13 | | | |
| **OC369a** | DAB1*01 | | | | | |
| **OC373a** | DAB1*08 | DAB1*31 | DAB1*60 | | | |
| **OC378a** | DAB1*04 | DAB1*13 | | | | |
| **OC379a** | DAB1*07 | DAB1*15 | DAB1*18 | | | |
| **OC383a** | DAB1*07 | DAB1*15 | | | | |
| **OC385a** | DAB1*27 | DAB1*18 | | | | |
| **OC389a** | DAB1*01 | DAB1*27 | DAB1*18 | DAB1*36 | DAB1*37 | |
| **OC390a** | DAB1*09 | DAB1*08 | DAB1*21 | DAB1*29 | DAB1*31 | |
| **OC391a** | DAB1*01 | DAB1*06 | DAB1*27 | | | |
| **OC392a** | DAB1*08 | DAB1*06 | DAB1*21 | DAB1*35 | | |
| **OC397a** | DAB1*04 | DAB1*09 | DAB1*26 | DAB1*33 | DAB1*41 | |
| **OC403a** | DAB1*01 | | | | | |
| **OC408a** | DAB1*01 | | | | | |
| **OC409a** | DAB1*08 | DAB1*26 | DAB1*31 | | | |
| **OC410a** | DAB1*01 | | | | | |
| **OC413a** | DAB1*01 | DAB1*09 | DAB1*21 | DAB1*29 | | |
| **OC414a** | DAB1*07 | | | | | |
| **OC419a** | DAB1*06 | DAB1*07 | DAB1*13 | | | |
| **OC420a** | DAB1*04 | DAB1*09 | DAB1*13 | DAB1*50 | | |
| **OC421a** | DAB1*06 | DAB1*11 | DAB1*35 | | | |
| **OMA40a** | DAB1*01 | DAB1*06 | DAB1*43 | | | |
| **OMA41a** | DAB1*01 | DAB1*48 | DAB1*57 | | | |
| **OMA42a** | DAB1*01 | DAB1*36 | DAB1*37 | | | |

Table A4.6.2: DAB2 genotypes for all storm petrel samples

| DAB Name | Genotype | | |
|---|---|---|---|
| **OC003** | DAB2*11 | | |
| **OC004** | DAB2*05 | | |
| **OC005** | DAB2*11 | DAB2*15 | |
| **OC007** | DAB2*03 | | |
| **OC008** | DAB2*03 | | |
| **OC009** | DAB2*06 | DAB2*14 | |
| **OC010** | DAB2*06 | | |
| **OC011** | DAB2*03 | | |
| **OC012** | DAB2*11 | | |
| **OC013** | DAB2*04 | | |
| **OC014** | DAB2*05 | DAB2*08 | DAB2*26 |
| **OC016** | DAB2*08 | | |
| **OC017** | DAB2*11 | | |
| **OC018** | DAB2*02 | DAB2*13 | |
| **OC021** | DAB2*03 | DAB2*14 | |
| **OC022** | DAB2*03 | | |
| **OC023** | DAB2*02 | DAB2*05 | |
| **OC024** | DAB2*11 | | |
| **OC025** | DAB2*02 | | |

| | | | |
|---|---|---|---|
| **OC026** | DAB2*02 | | |
| **OC028** | DAB2*03 | DAB2*05 | |
| **OC030** | DAB2*13 | | |
| **OC031** | DAB2*02 | | |
| **OC033** | DAB2*03 | DAB2*14 | |
| **OC034** | DAB2*02 | | |
| **OC035** | DAB2*03 | | |
| **OC036** | DAB2*03 | | |
| **OC039** | DAB2*03 | DAB2*02 | |
| **OC040** | DAB2*03 | DAB2*03 | |
| **OC041** | DAB2*03 | | |
| **OC042** | DAB2*06 | | |
| **OC045** | DAB2*03 | DAB2*21 | |
| **OC050** | DAB2*05 | DAB2*21 | |
| **OC052** | DAB2*05 | | |
| **OC053** | DAB2*03 | DAB2*04 | |
| **OC055** | DAB2*02 | | |
| **OC058** | DAB2*02 | DAB2*05 | |
| **OC059** | DAB2*04 | | |
| **OC061** | DAB2*02 | DAB2*05 | |
| **OC064** | DAB2*04 | | |
| **OC066** | DAB2*02 | DAB2*13 | |
| **OC067** | DAB2*02 | DAB2*03 | |
| **OC072** | DAB2*05 | DAB2*06 | |
| **OC073** | DAB2*02 | DAB2*04 | |
| **OC078** | DAB2*04 | DAB2*21 | |
| **OC080** | DAB2*02 | DAB2*03 | |
| **OC082** | DAB2*03 | DAB2*21 | |
| **OC085** | DAB2*05 | DAB2*15 | |
| **OC086** | DAB2*02 | | |
| **OC088** | DAB2*04 | DAB2*05 | |
| **OC090** | DAB2*13 | DAB2*14 | |
| **OC091** | DAB2*03 | DAB2*05 | |
| **OC097** | DAB2*02 | DAB2*03 | |
| **OC098** | DAB2*14 | DAB2*22 | |
| **OC099** | DAB2*08 | | |
| **OC101** | DAB2*07 | | |
| **OC102** | DAB2*06 | DAB2*07 | |
| **OC104** | DAB2*08 | DAB2*13 | |
| **OC110** | DAB2*13 | DAB2*14 | |
| **OC112** | DAB2*04 | | |
| **OC113** | DAB2*14 | | |
| **OC116** | DAB2*02 | | |
| **OC118** | DAB2*02 | | |
| **OC119** | DAB2*03 | DAB2*06 | |
| **OC120** | DAB2*13 | DAB2*22 | |
| **OC124** | DAB2*04 | DAB2*05 | |
| **OC128** | DAB2*04 | DAB2*05 | |
| **OC130** | DAB2*02 | DAB2*05 | |
| **OC131** | DAB2*02 | DAB2*06 | |
| **OC132** | DAB2*05 | | |
| **OC135** | DAB2*23 | | DAB2*23 only here |

| | | | |
|---|---|---|---|
| **OC136** | DAB2*02 | | |
| **OC137** | DAB2*05 | | |
| **OC139** | DAB2*08 | DAB2*13 | |
| **OC140** | DAB2*07 | DAB2*08 | |
| **OC141** | DAB2*05 | DAB2*06 | |
| **OC142** | DAB2*01 | DAB2*07 | |
| **OC145** | DAB2*05 | DAB2*15 | |
| **OC147** | DAB2*05 | DAB2*11 | |
| **OC149** | DAB2*06 | DAB2*13 | |
| **OC150** | DAB2*03 | DAB2*04 | |
| **OC151** | DAB2*06 | | |
| **OC153** | DAB2*02 | DAB2*06 | |
| **OC154** | DAB2*03 | DAB2*05 | |
| **OC164** | DAB2*04 | | |
| **OC165** | DAB2*07 | | |
| **OC166** | DAB2*06 | DAB2*22 | |
| **OC167** | DAB2*02 | DAB2*13 | |
| **OC168** | DAB2*07 | | |
| **OC170** | DAB2*02 | DAB2*05 | |
| **OC172** | DAB2*08 | DAB2*13 | |
| **OC173** | DAB2*07 | DAB2*08 | |
| **OC174** | DAB2*11 | | |
| **OC175** | DAB2*03 | DAB2*14 | |
| **OC176** | DAB2*05 | | |
| **OC178** | DAB2*02 | DAB2*13 | |
| **OC179** | DAB2*02 | DAB2*06 | DAB2*13 |
| **OC182** | DAB2*04 | DAB2*08 | |
| **OC184** | DAB2*02 | DAB2*05 | |
| **OC185** | DAB2*03 | DAB2*07 | |
| **OC186** | DAB2*02 | | |
| **OC188** | DAB2*07 | DAB2*08 | |
| **OC189** | DAB2*02 | DAB2*04 | |
| **OC190** | DAB2*02 | DAB2*03 | |
| **OC191** | DAB2*04 | DAB2*06 | |
| **OC192** | DAB2*03 | DAB2*05 | |
| **OC193** | DAB2*03 | DAB2*07 | |
| **OC196** | DAB2*02 | DAB2*08 | |
| **OC199** | DAB2*05 | | |
| **OC202** | DAB2*05 | | |
| **OC214** | DAB2*05 | DAB2*14 | |
| **OC216** | DAB2*01 | | |
| **OC217** | DAB2*01 | | |
| **OC218** | DAB2*09 | | |
| **OC219** | DAB2*01 | | |
| **OC220** | DAB2*01 | | |
| **OC221** | DAB2*01 | DAB2*10 | |
| **OC223** | DAB2*01 | | |
| **OC224** | DAB2*01 | | |
| **OC225** | DAB2*09 | | |
| **OC228** | DAB2*01 | | |
| **OC231** | DAB2*01 | DAB2*10 | |
| **OC232** | DAB2*01 | | |

| | | | |
|---|---|---|---|
| **OC233** | DAB2*01 | | |
| **OC235** | DAB2*18 | | |
| **OC237** | DAB2*01 | | |
| **OC238** | DAB2*20 | | |
| **OC239** | DAB2*15 | DAB2*25 | |
| **OC242** | DAB2*19 | | |
| **OC243** | DAB2*01 | DAB2*17 | |
| **OC244** | DAB2*01 | | |
| **OC247** | DAB2*01 | | |
| **OC248** | DAB2*09 | | |
| **OC250** | DAB2*01 | | |
| **OC251** | DAB2*17 | DAB2*24 | |
| **OC252** | DAB2*12 | | |
| **OC253** | DAB2*01 | DAB2*16 | |
| **OC255** | DAB2*01 | | |
| **OC258** | DAB2*01 | | |
| **OC261** | DAB2*09 | | |
| **OC262** | DAB2*01 | | |
| **OC263** | DAB2*12 | | |
| **OC265** | DAB2*01 | | |
| **OC266** | DAB2*01 | | |
| **OC267** | DAB2*01 | | |
| **OC268** | DAB2*01 | | |
| **OC269** | DAB2*01 | | |
| **OC270** | DAB2*10 | | |
| **OC271** | DAB2*01 | | |
| **OC272** | DAB2*17 | | |
| **OC273** | DAB2*17 | DAB2*19 | |
| **OC275** | DAB2*01 | | |
| **OC276** | DAB2*01 | | |
| **OC277** | DAB2*01 | | |
| **OC278** | DAB2*01 | | |
| **OC279** | DAB2*01 | DAB2*16 | |
| **OC283** | DAB2*10 | | |
| **OC284** | DAB2*12 | | |
| **OC285** | DAB2*01 | DAB2*09 | |
| **OC286** | DAB2*01 | | |
| **OC287** | DAB2*12 | | |
| **OC288** | DAB2*16 | | |
| **OC290** | DAB2*01 | | |
| **OC293** | DAB2*10 | | |
| **OC295** | DAB2*01 | | |
| **OC296** | DAB2*15 | | |
| **OC297** | DAB2*01 | | |
| **OC298** | DAB2*01 | | |
| **OC299** | DAB2*15 | DAB2*16 | |
| **OC300** | DAB2*01 | | |
| **OC301** | DAB2*01 | DAB2*12 | |
| **OC310** | DAB2*01 | DAB2*16 | |
| **OC314** | DAB2*18 | | |
| **OC316** | DAB2*01 | DAB2*17 | |
| **OC337** | DAB2*16 | | |

| | | | |
|---|---|---|---|
| **OC338** | DAB2*12 | | |
| **OC340** | DAB2*01 | | |
| **OC341** | DAB2*01 | | |
| **OC351** | DAB2*01 | | |
| **OC367** | DAB2*01 | | |
| **OC369** | DAB2*01 | | |
| **OC373** | DAB2*20 | | |
| **OC379** | DAB2*10 | | |
| **OC383** | DAB2*01 | | |
| **OC385** | DAB2*01 | DAB2*10 | |
| **OC389** | DAB2*01 | DAB2*18 | |
| **OC390** | DAB2*01 | | |
| **OC391** | DAB2*01 | | |
| **OC392** | DAB2*09 | | |
| **OC397** | DAB2*09 | | |
| **OC403** | DAB2*01 | | |
| **OC408** | DAB2*12 | DAB2*16 | |
| **OC409** | DAB2*19 | | |
| **OC410** | DAB2*01 | | |
| **OC413** | DAB2*01 | | |
| **OC414** | DAB2*01 | | |
| **OC419** | DAB2*01 | | |
| **OC421** | DAB2*09 | DAB2*15 | |
| **OM40** | DAB2*01 | | |
| **OM41** | DAB2*01 | | |
| **OM42** | DAB2*01 | DAB2*18 | |

## A.4.7: R script for Boxplots and Histograms

```
##############################################
######### LOCUS 1 READ COUNT BOXPLOTS ########
##############################################


L1READS <- read.csv(file.choose(), header=TRUE)

names (L1READS)


#################
### BOX PLOT ###
#################

boxplot (L1READS$FAIL.NA, L1READS$UNCHOSEN.NA, L1READS$UNUSED.NA,
     L1READS$EX.NEG, L1READS$PCR.NEG, L1READS$SAMPLES,
     varwidth = TRUE,
     notch = TRUE,
     ylab = "Read Count",
     xlab = "Groups",
     col=c("gray", "coral", "lightgoldenrod1", "plum", "palegreen",
"cadetblue3"),
     cex.lab = 1.0,
     cex.axis = 1.0,
     names = c("Failed Samples", "Unpooled MID-tags", "Unused MID-Tags",
     "Extraction negatives", "PCR negatives", "Samples"),
     las = 0)


##############################################
######### LOCUS 2 READ COUNT BOXPLOTS #########
##############################################


L2READS <- read.csv(file.choose(), header=TRUE)

names (L2READS)


#################
### BOX PLOT ###
#################

boxplot (L2READS$FAIL.NA, L2READS$UNCHOSEN.NA, L2READS$UNUSED.NA,
     L2READS$EX.NEG, L2READS$PCR.NEG, L2READS$SAMPLES,
     varwidth = TRUE,
     notch = TRUE,
     ylab = "Read Count",
     xlab = "Groups",
     col=c("gray", "coral", "lightgoldenrod1", "plum", "palegreen",
"cadetblue3"),
     cex.lab = 1.0,
     cex.axis = 1.0,
     names = c("Failed Samples", "Unpooled MID-tags", "Unused MID-Tags",
     "Extraction negatives", "PCR negatives", "Samples"),
     las = 0)



#####################################
####### L2 ALLELES AS DECIMALS #######
#####################################

L2ReadPercT <- read.csv(file.choose(), header=TRUE)

names(L2ReadPercT)
```

```
L2ReadPercT$READ.PERC <- as.numeric(L2ReadPercT$READ.PERC)

summary(L2ReadPercT)

hist(L2ReadPercT$READ.PERC,
   xlab = "Read Percentages",
   ylab = "Value",
   col = "light blue",
   cex.lab = 1.6,
   cex.axis = 1.6,
   ylim = c(0, 120),
   xlim = c(0, 1),
   breaks = 50,
   las = 1)


####################################
####### L1 ALLELES AS DECIMALS ######
####################################


L1ReadPercT <- read.csv(file.choose(), header=TRUE)

names(L1ReadPercT)

L1ReadPercT$READ.PERC <- as.numeric(L1ReadPercT$READ.PERC)

summary(L1ReadPercT)

hist(L1ReadPercT$READ.PERC,
   xlab = "Read Percentages",
   ylab = "Value",
   col = "orange",
   cex.lab = 1.6,cex.axis = 1.6,
   ylim = c(0, 40),
   xlim = c(0, 1),
   breaks = 50,
   las = 1)
```

## A4.8: Divergence Analyses in R

```
################## Phylogenetic clustering and NMNDS of Azores
storm petrel MHC DAB1 and DAB2 genotypes

# FH, 2021-04-02

#################

if(!require("data.table")) {install.packages("data.table");
library(data.table)}


setwd("~/Dropbox/Documents/Seabirds/Alex McCubbin PhD/Alex
McCubbin PhD/Chapters & manuscripts/MHC Illumina chapter/")


allele.table <- read.table("DAB1&2alleles.indlist.wo.na.txt",
sep= "\t", header=T)

#allele.table <- fread("DAB1&2alleles.indlist.wo.na.txt")


indslist <- colnames(allele.table) #columns of input file are
the ind names


### calling allele names manually

DAB1.alleles = c("DAB1.01", "DAB1.02", "DAB1.03", "DAB1.04",
"DAB1.05", "DAB1.06", "DAB1.07", "DAB1.08", "DAB1.09",
"DAB1.10", "DAB1.11", "DAB1.12", "DAB1.13", "DAB1.14",
"DAB1.15", "DAB1.16", "DAB1.17", "DAB1.18", "DAB1.19",
"DAB1.20", "DAB1.21", "DAB1.22", "DAB1.23", "DAB1.24",
"DAB1.25", "DAB1.26", "DAB1.27", "DAB1.28", "DAB1.29",
"DAB1.30", "DAB1.31", "DAB1.32", "DAB1.33", "DAB1.34",
"DAB1.35", "DAB1.36", "DAB1.37", "DAB1.38", "DAB1.39",
"DAB1.40", "DAB1.41", "DAB1.42", "DAB1.43", "DAB1.44",
"DAB1.45", "DAB1.46", "DAB1.47", "DAB1.48", "DAB1.49",
"DAB1.50", "DAB1.51", "DAB1.52", "DAB1.53", "DAB1.54",
"DAB1.55", "DAB1.56", "DAB1.57", "DAB1.58", "DAB1.59",
"DAB1.60", "DAB1.61", "DAB1.62", "DAB1.63", "DAB1.64",
"DAB1.65")

DAB2.alleles = c("DAB2.01", "DAB2.02", "DAB2.03", "DAB2.04",
"DAB2.05", "DAB2.06", "DAB2.07", "DAB2.08", "DAB2.09",
"DAB2.10", "DAB2.11", "DAB2.12", "DAB2.13", "DAB2.14",
"DAB2.15", "DAB2.16", "DAB2.17", "DAB2.18", "DAB2.19",
"DAB2.20", "DAB2.21", "DAB2.22", "DAB2.23", "DAB2.24",
"DAB2.25", "DAB2.26", "DAB2.27")

all.alleles = c(DAB1.alleles, DAB2.alleles)


##### generate one vector for each ind, with TRUE/FALSE for
each allele: presence/absence info

for (inds in 1:ncol(allele.table)) {
```

```
 assign(paste("ind", names(allele.table)[inds], sep = "_"),
as.vector(all.alleles %in% allele.table[,inds]))

#  allele.table[,1] == all.alleles[allele.run]

#  all.alleles[allele.run] %in% allele.table[,1]

}

##### Combine ind vectors of presence/absence into one
combined data frame

DAB.indfile = data.frame()

for(inds in 1:ncol(allele.table)){

 indallele <- get(paste("ind", names(allele.table)[inds], sep
= "_"))

 DAB.indfile <- rbind(DAB.indfile, indallele)

}


##### rename columns (allele names) and rows (ind names) of
data frame

DAB.indfile[1:3,1:3]

names(DAB.indfile) <- all.alleles

class(DAB.indfile)

row.names(DAB.indfile) <- indslist


write.csv(DAB.indfile, file = "DAB.indfile.csv")


##### replace TRUE with 1 and FALSE with 0

DAB.indfile.presabs <- DAB.indfile

DAB.indfile.presabs[] <- lapply(DAB.indfile.presabs, gsub,
pattern = "TRUE", replacement = "1")#, fixed = TRUE)

DAB.indfile.presabs[] <- lapply(DAB.indfile.presabs, gsub,
pattern = "FALSE", replacement = "0")#, fixed = TRUE)

DAB.indfile.presabs <- data.frame(DAB.indfile.presabs) #,
quote=FALSE)

write.csv(DAB.indfile.presabs, file =
"DAB.indfile.presabs.csv")


##### Bray-Curtis & Jaccard distances applied to the presence-
absence data

library(vegan)

library(gdata)
```

```
#bray curtis & Jaccard distance calculation - doing this for
phylogenetics, and to see/read distances. NMDS incorporates
dist. calc. in the command metaMDS, so this calculation is
done again below.

mean(as.integer(DAB.indfile.presabs$DAB1.01)) #original data
frame had values not stored as numbers, so could not perform
calculations

DAB.indfile.presabs.num <- sapply(DAB.indfile.presabs,
as.numeric, USE.NAMES = T)

DAB_BC <- vegdist(x = DAB.indfile.presabs.num, method="bray",
binary=TRUE) #bray curtis distance on presence absence dataset
#look at bray curtis results

#not done: normalisation of scores.

# DAB_BC.std <-
vegdist(decostand(as.numeric(DAB.indfile.presabs), "norm"),
"jaccard")

DAB_Jac <- vegdist(x = DAB.indfile.presabs.num,
method="jaccard", binary=TRUE) #brayjaccard distance on
presence absence dataset #look at bray curtis results


DAB_BC.m <- as.matrix(DAB_BC)

DAB_Jac.m <- as.matrix(DAB_Jac)

lowerTriangle(DAB_BC.m, diag = T) <- NA #replaces lower
triangle and diagonal with NA

lowerTriangle(DAB_Jac.m, diag = T) <- NA #replaces lower
triangle and diagonal with NA

row.names(DAB_BC.m) <- indslist #add ind names

row.names(DAB_Jac.m) <- indslist #add ind names


write.csv(as.matrix(DAB_BC.m), file = "DAB_BC.csv")

write.csv(as.matrix(DAB_Jac.m), file = "DAB_Jac.csv")


##nmds####

# info: https://jonlefcheck.net/2012/10/24/nmds-tutorial-in-r/

row.names(DAB.indfile.presabs.num) <- indslist #add ind names

set.seed(1)

mds.bc1 <- metaMDS(DAB.indfile.presabs.num, distance =
"bray",binary=TRUE, k = 2, trymax = 50000) #get values for
NMDS analysis Sites<- data.frame(ASAP, scores(mds.1, display =
"sites")) #site scores for each

set.seed(1)

mds.jac1 <- metaMDS(DAB.indfile.presabs.num, distance =
"jaccard",binary=TRUE, k = 2, trymax = 50000, sratmax =
```

```
0.999999) #get values for NMDS analysis Sites<-
data.frame(ASAP, scores(mds.1, display = "sites")) #site
scores for each


mds.1 #I got the NMDS working for the jaccard distance, using
k=2, trymax=50000 and sratmax== 0.999999. Did not sent
rand.seed unfortunately.

#variable name for this result was mds.1, which is what the
script below uses for plotting.


save(mds.1, file = "DAB_Jac_NMDS.k2.rda") #save NMDS results
to file. Can take long to re-generate (here 50000 steps, =
hours?)

ls()

rm(mds.1)

load("DAB_Jac_NMDS.k2.rda")

ls()

speciesID = c("C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C",
"C", "C", "C", "C", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "M", "M", "M", "M", "M", "M", "M", "M", "M")

speciesCol = c("1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",
"1", "1", "1", "1", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2", "2",
"2", "2", "2", "2", "2", "2", "2", "2", "2", "2")
```

```
# castro = C = 1, monteiro = M = 2


#par(mfrow = c(1,1))

stressplot(mds.1)


mds.1$species

spp <- data.frame(scores(mds.1, display = "species"))


##GGplot2##

#visualise NMDS plot with GGPlot2

library(ggplot2)

data.scores = as.data.frame(scores(mds.1))

row.names(data.scores)

data.scores$taxon <- speciesID # create a column of site
names, from the rownames of data.scores

data.scores$grp <- speciesCol # add the grp variable created
earlier

head(data.scores) #look at the data


allele.scores <- as.data.frame(scores(mds.1, "species"))
#Using the scores function from vegan to extract the species
scores and convert to a data.frame

allele.scores$alleles <- rownames(allele.scores) # create a
column of species, from the rownames of species.scores

head(species.scores) #look at the data


ggplot() +


geom_text(data=data.scores,aes(x=NMDS1,y=NMDS2,label=row.names
(data.scores)),colour = speciesCol, alpha=0.9, cex=3)

 #+... add the species labels


#geom_point(data=data.scores,aes(x=NMDS1,y=NMDS2,shape=species
ID,colour=speciesID),size=1) # add the point markers
```

# Bibliography


Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D. and Wayne, R.K. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America* 101(10), pp. 3490–3494. doi: 10.1073/pnas.0306582101.

Alberts, B., Johnson, A., Lewis, J. et al. 2002a. The Adaptive Immune Response. In: *Molecular Biology of the Cell*. 4th ed. New York: Garland Science. Available at: https://www.ncbi.nlm.nih.gov/books/NBK21070/.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walters, P. 2002b. *Molecular Biology of the Cell*. 4th ed. New York and London: Garland Science. Available at: http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E14-10-1437.

Amo, L., Avilés, J.M., Parejo, D., Peña, A., Rodríguez, J. and Tomás, G. 2012. Sex recognition by odour and variation in the uropygial gland secretion in starlings. *Journal of Animal Ecology* 81(3), pp. 605–613. doi: 10.1111/j.1365-2656.2011.01940.x.

Andersson, M. 1994. The Theory of Sexual Selection. In: *Sexual Selection.* Princeton University Press, pp. 3–31. Available at: http://www.jstor.org/stable/j.ctvs32s1x.5.

Andersson, M. and Simmons, L.W. 2006. Sexual selection and mate choice. *Trends in Ecology and Evolution* 21(6), pp. 296–302. doi: 10.1016/j.tree.2006.03.015.

Balthazart, J. and Taziaux, M. 2009. The underestimated role of olfaction in avian reproduction? *Behavioural Brain Research* 200(2), pp. 248–259. doi: 10.1016/j.bbr.2008.08.036.

Bang, B.G. 1966. The olfactory apparatus of tubenosed birds (Procellariiformes). *Acta Anat* 65, pp. 391–415.

Bang, B.G. and Cobb, S. 1968. The size of the olfactory bulb in 108 species of Birds. *The Auk* 85, pp. 55–61. doi: 10.2307/4083624.

Behjati, S. and Tarpey, P.S. 2013. What is next generation sequencing? *Archives of disease in childhood*, 98(6), pp. 236 LP – 238. Available at: http://ep.bmj.com/content/98/6/236.abstract.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. 2013. GenBank. Nucleic acids research, pp. D36–D42. Available at: https://pubmed.ncbi.nlm.nih.gov/23193287.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. 2005. GenBank. Nucleic Acids Research 33(DATABASE ISS.), pp. 34–38. doi: 10.1093/nar/gki063.

Bentkowski, P. and Radwan, J. 2019. Evolution of major histocompatibility complex gene copy number. *PLoS Computational Biology* 15(5), pp. 1–15. doi: 10.1371/journal.pcbi.1007015.

Bernatchez, L. and Landry, C. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology* 16(3), pp. 363–377. Available at: http://www.blackwell-synergy.com/doi/abs/10.1046/j.1420-9101.2003.00531.x.

Biedrzycka, A. et al. 2017a. Extreme MHC class i diversity in the sedge warbler (*Acrocephalus schoenobaenus*); Selection patterns and allelic divergence suggest that

different genes have different functions. *BMC Evolutionary Biology* 17(1), pp. 1–12. doi: 10.1186/s12862-017-0997-9.

Biedrzycka, A. et al. 2018. Blood parasites shape extreme major histocompatibility complex diversity in a migratory passerine. *Molecular Ecology*, pp. 0–1. doi: 10.1111/mec.14592.

Biedrzycka, A., Sebastian, A., Migalska, M., Westerdahl, H. and Radwan, J. 2017b. Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources* 17(4), pp. 642–655. doi: 10.1111/1755-0998.12612.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R. and Willerslev, E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PloS One* 2(2), pp. e197–e197. Available at: https://pubmed.ncbi.nlm.nih.gov/17299583.

BirdLife International (2016) Species factsheet: *Hydrobates monteiroi*. Downloaded from http://www.birdlife.org on 20/12/2016

BirdLife International (2016) Species factsheet: *Hydrobates castro*. Downloaded from http://www.birdlife.org on 20/12/2016

Boehm, T. and Zufall, F. 2006. MHC peptides and the sensory evaluation of genotype. *Trends in Neurosciences* 29(2), pp. 100–107. doi: 10.1016/j.tins.2005.11.006.

Bohnet, S., Rogers, L., Sasaki, G. and Kolattukudy, P.E. 1991. Estradiol induces proliferation of peroxisome-like microbodies and the production of 3-hydroxy fatty acid diesters, the female pheromones, in the uropygial glands of male and female mallards. *The Journal of biological chemistry* 266(15), pp. 9795–9804.

Bolton, M. 2007. Playback experiments indicate absence of vocal recognition among temporally and geographically separated populations of Madeiran Storm-petrels *Oceanodroma castro*. *Ibis* 149(2), pp. 255–263. doi: 10.1111/j.1474-919X.2006.00624.x.

Bolton, M. et al. 2008. Monteiro's Storm-petrel *Oceanodroma monteiroi*: A new species from the Azores. *Ibis* 150(4), pp. 717–727. doi: 10.1111/j.1474-919X.2008.00854.x.

Bolton, M., Medeiros, R., Hothersall, B. and Campos, A. 2004. The use of artificial breeding chambers as a conservation measure for cavity-nesting procellariiform seabirds: A case study of the Madeiran storm petrel (*Oceanodroma castro*). *Biological Conservation* 116(1), pp. 73–80. doi: 10.1016/S0006-3207(03)00178-2.

Bolton, M. and Thomas, R. 2001. Moult and ageing of storm petrels *Hydrobates pelagicus*. *Ringing and Migration* 20(3), pp. 193–201. doi: 10.1080/03078698.2001.9674244.

Bonadonna, F. 2009. Olfaction in petrels: From homing to self-odor avoidance. *Annals of the New York Academy of Sciences* 1170, pp. 428–433. doi: 10.1111/j.1749-6632.2009.03890.x.

Bonadonna, F., Cunningham, G., Jouventin, P., Hesters, F. and Nevitt, G. 2003. Evidence for nest-odour recognition in two species of diving petrel. *The Journal of experimental biology* 206, pp. 3719–3722. doi: 10.1242/jeb.00610.

Bonadonna, F., Miguel, E., Grosbois, V., Jouventin, P. and Bessiere, J.M. 2007. Individual odor recognition in birds: An endogenous olfactory signature on petrels' feathers? *Journal of Chemical Ecology* 33(9), pp. 1819–1829. doi: 10.1007/s10886-007-9345-7.

Bonadonna, F. and Nevitt, G.A. 2004. Partner-specific odor recognition in an Antarctic seabird. *Science* 306(5697), p. 835. doi: 10.1126/science.1103001.

Bonadonna, F., Spaggiari, J. and Weimerskirch, H. 2001. Could osmotaxis explain the ability of blue petrels to return to their burrows at night? *The Journal of experimental biology* 204, pp. 1485–1489.

Borgia, G., Kaatz, I.M. and Condit, R. 1987. Flower choice and bower decoration in the satin bowerbird *Ptilonorhynchus violaceus*: a test of hypotheses for the evolution of male display. *Animal Behaviour* 35(4), pp. 1129–1139. Available at: https://www.sciencedirect.com/science/article/pii/S0003347287801690.

Boul, K.E., Funk, C.W., Darst, C.R., Cannatella, D.C. and Ryan, M.J. 2007. Sexual selection drives speciation in an Amazonian frog. *Proceedings of the Royal Society B: Biological Sciences* 274(1608), pp. 399–406. Available at: https://doi.org/10.1098/rspb.2006.3736.

Brambilla, A., Keller, L., Bassano, B. and Grossen, C. 2018. Heterozygosity–fitness correlation at the major histocompatibility complex despite low variation in Alpine ibex (*Capra ibex*). *Evolutionary Applications* 11(5), pp. 631–644. doi: 10.1111/eva.12575.

Braune, P., Schmidt, S. and Zimmermann, E. 2008. Acoustic divergence in the communication of cryptic species of nocturnal primates (*Microcebus* ssp.). *BMC Biology* 6(1), p. 19. Available at: https://doi.org/10.1186/1741-7007-6-19.

Bretagnolle, V. 1996. Acoustic communication in a group of nonpasserine birds, the petrels. In: Kroosdma, D. E. and Miller, E. H. eds. *Ecology and Evolution of Acoustic Communication in Birds*. Ithaca, New York: Cornell University Press, pp. 160–177.

Bried, J. et al. 2009. Seabird Habitat restoration on praia islet, Azores Archipelago. *Ecological Restoration* 27(1), pp. 27–36. doi: 10.3368/er.27.1.27.

Bried, J. and Bolton, M. 2005. An initial estimate of age at first return and breeding in Madeiran Storm-petrels *Oceanodroma castro. Atlantic Seabirds* 7(2), pp. 71–74.

Brock, C.D. and Wagner, C.E. 2018. The smelly path to sympatric speciation? *Molecular Ecology* 27(21), pp. 4153–4156. doi: 10.1111/mec.14845.

Brown, D.S., Burger, R., Cole, N., Vencatasamy, D., Clare, E.L., Montazam, A. and Symondson, W.O.C. 2014. Dietary competition between the alien Asian Musk Shrew (*Suncus murinus*) and a re-introduced population of Telfair's Skink (*Leiolopisma telfairii*). *Molecular Ecology* 23(15), pp. 3695–3705. Available at: https://doi.org/10.1111/mec.12445.

Brown, J.H., Jardetzky, T.S., Gorgat, J.C., Stern, L.J., Urban, R.G., Strominger, J.L. and Wiley, D.C. 1993. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. 364(July), pp. 1467–1468.

Brown, R.M. et al. 2010. Range expansion and hybridization in Round Island petrels (*Pterodroma* spp.): evidence from microsatellite genotypes. *Molecular Ecology* 19(15), pp. 3157–3170. Available at: https://doi.org/10.1111/j.1365-294X.2010.04719.x.

Bruford, M., Hanotte, O., Brookfield, J. and Burke, T. 1998. Single-Locus and Multilocus DNA Fingerprinting. *Molecular Genetics Analysis of Populations: A Practical Approach*

Burri, R., Hirzel, H.N., Salamin, N., Roulin, A. and Fumagalli, L. 2008a. Evolutionary patterns of MHC class II B in owls and their implications for the understanding of avian

MHC evolution. *Molecular Biology and Evolution* 25(6), pp. 1180–1191. doi: 10.1093/molbev/msn065.

Burri, R., Niculita-Hirzel, H., Roulin, A. and Fumagalli, L. 2008. Isolation and characterization of major histocompatibility complex (MHC) class II B genes in the Barn owl (Aves: *Tyto alba*). *Immunogenetics* 60(9), pp. 543–550. doi: 10.1007/s00251-008-0308-0.

Burri, R., Promerová, M., Goebel, J. and Fumagalli, L. 2014. PCR-based isolation of multigene families: Lessons from the avian MHC class IIB. *Molecular Ecology Resources* 14(4), pp. 778–788. doi: 10.1111/1755-0998.12234.

Burri, R., Salamin, N., Studer, R.A., Roulin, A. and Fumagalli, L. 2010. Adaptive divergence of ancient gene duplicates in the avian MHC class II β. *Molecular Biology and Evolution* 27(10), pp. 2360–2374. doi: 10.1093/molbev/msq120.

Campagna, S., Mardon, J., Celerier, A. and Bonadonna, F. 2012. Potential semiochemical molecules from birds: a practical and comprehensive compilation of the last 20 years studies. *Chemical senses* 37(1), pp. 3–25. doi: 10.1093/chemse/bjr067.

Carboneras, C. 1992. Family *Hydrobatidae* (storm-petrels)., pp. 258–271.

Caro, S.P. and Balthazart, J. 2010. Pheromones in birds : myth or reality ? Journal of comparative physiology A, *Neuroethology, sensory, neural, and behavioral physiology* 196(10), pp. 751–766. doi: 10.1007/s00359-010-0534-4.Pheromones.

Chen, S., Zhou, Y., Chen, Y. and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/bioinformatics/bty560.

Chen, Z. and Wiens, J.J. 2020. The origins of acoustic communication in vertebrates. *Nature Communications* 11(1), p. 369. Available at: https://doi.org/10.1038/s41467-020-14356-3.

Costello, M. et al. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms., pp. 1–10.

Coyne, J.A. and Orr, H.A. 1998. The evolutionary genetics of speciation. *Philosophical transactions of the Royal Society of London*. Series B, Biological sciences 353(1366), pp. 287–305. Available at: https://pubmed.ncbi.nlm.nih.gov/9533126.

Hamilton, D.W. and Marlene, Z. 1982. Heritable True Fitness and Bright Birds: A Role for Parasites? *Science* 218(4570), pp. 384–387. Available at: https://doi.org/10.1126/science.7123238.

Dalén, L., Götherström, A. and Angerbjörn, A. 2004. Identifying species from pieces of faeces. *Conservation Genetics* 5(1), pp. 109–111. doi: 10.1023/B:COGE.0000014060.54070.45.

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D. 2012. jModelTest 2 : more models , new heuristics and parallel computing CircadiOmics : integrating circadian genomics , transcriptomics , proteomics. *Nature Methods* 9(8), p. 2106. doi: 10.1038/nmeth.2109.

Dean, L.L., Dunstan, H.R., Reddish, A. and MacColl, A.D.C. 2021. Courtship behavior, nesting microhabitat, and assortative mating in sympatric stickleback species pairs. *Ecology and Evolution* (August 2020), pp. 1741–1755. doi: 10.1002/ece3.7164.

Dearborn, D.C., Gager, A.B., Gilmour, M.E., McArthur, A.G., Hinerfeld, D.A. and Mauck, R.A. 2015. Non-neutral evolution and reciprocal monophyly of two expressed MHC class II B

genes in Leach's storm-petrel. I*mmunogenetics* 67(2), pp. 111–123. doi: 10.1007/s00251-014-0813-2.

Dearborn, D.C., Gager, A.B., McArthur, A.G., Gilmour, M.E., Mandzhukova, E. and Mauck, R.A. 2016. Gene duplication and divergence produce divergent MHC genotypes without disassortative mating. *Molecular Ecology* 25(17), pp. 4355–4367. doi: 10.1111/mec.13747.

Doherty, P.C. and Zinkernagel, R.M. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* 256(5512), pp. 50–52. Available at: https://doi.org/10.1038/256050a0.

Driver, R.J. and Balakrishnan, C.N., 2021 Highly Contiguous Genomes Improve the Understanding of Avian Olfactory Receptor Repertoires, Integrative and *Comparative Biology*;, icab150, https://doi.org/10.1093/icb/icab150

Edelaar, P., Alonso, D., Lagerveld, S., Senar, J.C. And Björklund, M. 2012. Population differentiation and restricted gene flow in Spanish crossbills: not isolation-by-distance but isolation-by-ecology. *Journal of Evolutionary Biology* 25(3), pp. 417–430. Available at: https://doi.org/10.1111/j.1420-9101.2011.02443.x.

Eizaguirre, C., Lenz, T.L., Sommerfeld, R.D., Harrod, C., Kalbe, M. and Milinski, M. 2011. Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evolutionary Ecology* 25(3), pp. 605–622. Available at: https://doi.org/10.1007/s10682-010-9424-z.

Eizaguirre, C., Lenz, T.L., Traulsen, A. and Milinski, M. 2009a. Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. *Ecology Letters* 12(1), pp. 5–12. doi: 10.1111/j.1461-0248.2008.01247.x.

Eizaguirre, C., Yeates, S.E., Lenz, T.L., Kalbe, M. and Milinski, M. 2009b. MHC-based mate choice combines good genes and maintenance of MHC polymorphism. Molecular Ecology 18(15), pp. 3316–3329. doi: 10.1111/j.1365-294X.2009.04243.x.

Excoffier, L. and Lischer, H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* 10(3), pp. 564–567. doi: 10.1111/j.1755-0998.2010.02847.x.

Farell, E.M. and Alexandre, G. 2012. Bovine serum albumin further enhances the effects of organic solvents on increased yield of polymerase chain reaction of GC-rich templates. *BMC Research Notes* 5(1), p. 1. Available at: BMC Research Notes.

Forsberg, L.A., Dannewitz, J., Petersson, E. and Grahn, M. 2007. Influence of genetic dissimilarity in the reproductive success and mate choice of brown trout - Females fishing for optimal MHC dissimilarity. *Journal of Evolutionary Biology* 20(5), pp. 1859–1869. doi: 10.1111/j.1420-9101.2007.01380.x.

Frankham, R. 1995. Effective population size/adult population size ratios in wildlife: a review. *Genetical Research* 66(2), pp. 95–107. Available at: https://www.cambridge.org/core/article/effective-population-sizeadult-population-size-ratios-in-wildlife-a-review/59D48CF3CCD19ECE47163795F6EEA89E.

Fridolfsson, A.-K. and Ellegren, H. 1999. A Simple and Universal Method for Molecular Sexing of Non-Ratite Birds. *Journal of Avian Biology* 30(1), pp. 116–121. Available at: http://www.jstor.org/stable/3677252.

Friesen, V.L., Burg, T.M. and McCoy, K.D. 2007. Mechanisms of population differentiation in seabirds: Invited review. *Molecular Ecology* 16(9), pp. 1765–1785. doi: 10.1111/j.1365-294X.2006.03197.x.

Friesen, V.L., Smith, a L., Gómez-Díaz, E., Bolton, M., Furness, R.W., González-Solís, J. and Monteiro, L.R. 2007. Sympatric speciation by allochrony in a seabird. *Proceedings of the National Academy of Sciences of the United States of America* 104(47), pp. 18589–18594. doi: 10.1073/pnas.0700446104.

Frith, D.W. and Frith, C.B. 1988. Courtship Display and Mating of the Superb Bird of Paradise Lophorina superb. Emu - *Austral Ornithology* 88(3), pp. 183–188. Available at: https://doi.org/10.1071/MU9880183.

Fu, Y.X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147(2), pp. 915–925.

Furlan, E., Stoklosa, J., Griffiths, J., Gust, N., Ellis, R., Huggins, R.M. and Weeks, A.R. 2012. Small population size and extremely low levels of genetic diversity in island populations of the platypus, *Ornithorhynchus anatinus*. *Ecology and evolution* 2(4), pp. 844–857. Available at: https://pubmed.ncbi.nlm.nih.gov/22837830.

Gabirot, M., Buatois, B., Müller, C.T. and Bonadonna, F. 2018. Odour of King Penguin feathers analysed using direct thermal desorption discriminates between individuals but not sexes. *Ibis* 160(2), pp. 379–389. doi: 10.1111/ibi.12544.

Gagliardo, A. 2013. Forty years of olfactory navigation in birds., pp. 2165–2171. doi: 10.1242/jeb.070250.

Gaigher, A., Burri, R., Gharib, W.H., Taberlet, P., Roulin, A. and Fumagalli, L. 2016. Family-assisted inference of the genetic architecture of major histocompatibility complex variation. *Molecular Ecology Resources* 16(6), pp. 1353–1364. doi: 10.1111/1755-0998.12537.

Gaigher, A., Roulin, A., Gharib, W.H., Taberlet, P., Burri, R. and Fumagalli, L. 2018. Lack of evidence for selection favouring MHC haplotypes that combine high functional diversity. *Heredity* 120(5), pp. 396–406. Available at: http://dx.doi.org/10.1038/s41437-017-0047-9.

Gangloff, B., Zino, F., Shirihai, H., González-Solís, J., Couloux, A., Pasquet, E. and Bretagnolle, V. 2013. The evolution of north-east Atlantic gadfly petrels using statistical phylogeography. *Molecular Ecology* 22(2), pp. 495–507. doi: 10.1111/mec.12119.

Gasparini, C., Congiu, L. and Pilastro, A. 2015. Major histocompatibility complex similarity and sexual selection: Different does not always mean attractive. *Molecular Ecology* 24(16), pp. 4286–4295. doi: 10.1111/mec.13222.

Glenn, T.C. 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources* 11(5), pp. 759–769. doi: 10.1111/j.1755-0998.2011.03024.x.

Goebel, J. et al. 2017. 100 million years of multigene family evolution: Origin and evolution of the avian MHC class IIB. *BMC Genomics* 18(1), pp. 1–9. doi: 10.1186/s12864-017-3839-7.

Grant, P.R. and Grant, B.R. 2009. Sympatric Speciation, Immigration, and Hybridization in Island Birds: . In: Losos, J. B. and Ricklefs, R. E. eds. *Princeton University Press*, pp. 326–357. Available at: https://doi.org/10.1515/9781400831920.326.

Griffin, P.C., Robin, C. and Hoffmann, A.A. 2011. A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. *BMC Biology* 9(1), p. 19. Available at: https://doi.org/10.1186/1741-7007-9-19.

Griffith, S.C., Owens, I.P.F. and Thuman, K.A. 2002. Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Molecular Ecology* 11(11), pp. 2195–2212. Available at: https://doi.org/10.1046/j.1365-294X.2002.01613.x.

Grogan, K.E., Harris, R.L., Boulet, M. and Drea, C.M. 2018. MHC Genetic Variation Influences both Olfactory Signals and Scent Discrimination in Ring-Tailed Lemurs. *bioRxiv* , pp. 1–16. doi: 10.1101/337105.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52(5), pp. 696–704. doi: 10.1080/10635150390235520.

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D.C. and Shyr, Y. 2012. The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 13(1), p. 666. Available at: https://doi.org/10.1186/1471-2164-13-666.

Hailer, F. et al. 2006. Bottlenecked but long-lived: high genetic diversity retained in white-tailed eagles upon recovery from population decline. *Biology letters* 2(2), pp. 316–319. Available at: https://pubmed.ncbi.nlm.nih.gov/17148392.

Hailer, F., Kutschera, V.E., Hallström, B.M., Fain, S.R., Leonard, J.A., Arnason, U. and Janke, A. 2013. Response to Comment on "Nuclear Genomic Sequences Reveal that Polar Bears Are an Old and Distinct Bear Lineage." *Science* 339(6127), pp. 1522 LP – 1522. Available at: http://science.sciencemag.org/content/339/6127/1522.2.abstract.

Hardjasa, A., Ling, M., Ma, K. and Yu, H. 2010. Investigating the Effects of DMSO on PCR Fidelity Using a Restriction Digest-Based Method. 14(April), pp. 161–164.

Harris, M.P. 1969. The biology of storm petrels in the Galápagos Islands. *Proceedings of the California Academy of Sciences, Fourth Series* 37(4), pp. 95–166.

Harvey, M.G., Bonter, D.N., Stenzler, L.M. and Lovette, I.J. 2006. A comparison of plucked feathers versus blood samples as DNA sources for molecular sexing. *Journal of Field Ornithology* 77(2), pp. 136–140. Available at: https://doi.org/10.1111/j.1557-9263.2006.00033.x.

Hasegawa, M., Kishino, H. and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution* 22(2), pp. 160–174. doi: 10.1007/BF02101694.

Hasiniaina, A.F. et al. 2020. Evolutionary significance of the variation in acoustic communication of a cryptic nocturnal primate radiation (*Microcebus* spp.). *Ecology and Evolution* 10(8), pp. 3784–3797. Available at: https://doi.org/10.1002/ece3.6177.

He, K., Minias, P. and Dunn, P.O. 2021. Long-Read Genome Assemblies Reveal Extraordinary Variation in the Number and Structure of MHC Loci in Birds. *Genome Biology and Evolution* 13(2). Available at: https://doi.org/10.1093/gbe/evaa270.

Hedrick, P.W. 1999. Balancing selection and MHC., pp. 207–214.

Hermansen, J.O.S., Sæther, S.A., Elgvin, T.O., Borge, T., Hjelle, E. and Sætre, G.P. 2011. Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and

barriers to gene flow. *Molecular Ecology* 20(18), pp. 3812–3822. Available at: https://doi.org/10.1111/j.1365-294X.2011.05183.x.

Hess, C.M. and Edwards, S.V. 2002. The Evolution of the Major Histocompatibility. *Bioscience* 52(5), pp. 423–431.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35(2), pp. 518–522. Available at: https://doi.org/10.1093/molbev/msx281.

Hogner, S. et al. 2012. Deep sympatric mitochondrial divergence without reproductive isolation in the common redstart *Phoenicurus phoenicurus*. *Ecology and Evolution* 2(12), pp. 2974–2988. Available at: https://doi.org/10.1002/ece3.398.

Holt, C. et al. 2018. Improved Genome Assembly and Annotation for the Rock Pigeon (*Columba livia*). *G3: Genes|Genomes|Genetics* 8(5), pp. 1391 LP – 1398. Available at: http://www.g3journal.org/content/8/5/1391.abstract.

Hoover, B., Alcaide, M., Jennings, S., Sin, S.Y.W., Edwards, S. V. and Nevitt, G.A. 2018. Ecology can inform genetics: Disassortative mating contributes to MHC polymorphism in Leach's storm-petrels (*Oceanodroma leucorhoa*). *Molecular Ecology* 27(16), pp. 3371–3385. doi: 10.1111/mec.14801.

Hoseini, S.S. and Sauer, M.G. 2015. Molecular cloning using polymerase chain reaction, an educational guide for cellular engineering. Journal of biological engineering 9, p. 2. Available at: https://pubmed.ncbi.nlm.nih.gov/25745516.

Hoyo, J. del, Collar, N.J. and International., B. 2014. HBW and BirdLife International illustrated checklist of the birds of the world.

Huchard, E., Knapp, L.A., Wang, J., Raymond, M. and Cowlishaw, G. 2010. MHC, mate choice and heterozygote advantage in a wild social primate. *Molecular Ecology* 19(12), pp. 2545–2561. doi: 10.1111/j.1365-294X.2010.04644.x.

Hughes, A.L. and Yeager, M. 1998. Natural Selection At Major Histocompatibility Complex Loci of Vertebrates. *Annual Review of Genetics* 32(1), pp. 415–435. Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.32.1.415.

Hume, J.B., Recknagel, H., Bean, C.W., Adams, C.E. and Mable, B.K. 2018. RADseq and mate choice assays reveal unidirectional gene flow among three lamprey ecotypes despite weak assortative mating: Insights into the formation and stability of multiple ecotypes in sympatry. *Molecular Ecology* 27(22), pp. 4572–4590. doi: 10.1111/mec.14881.

Jaccard, P. 1912. The distribution of the flora in the alpine zone.1. New Phytologist 11(2), pp. 37–50. Available at: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Jacob, J., Balthazart, J. and Schoffeniels, E. 1979. Sex differences in the chemical composition of uropygial gland waxes in domestic ducks. *Biochemical Systematics and Ecology* 7(2), pp. 149–153. doi: 10.1016/0305-1978(79)90024-3.

Jan Ejsmond, M., Radwan, J. and Wilson, A.B. 2014. Sexual selection and the evolutionary dynamics of the major histocompatibility complex. *Proceedings of the Royal Society of London B: Biological Sciences* 281(October), p. 20141662.

Janeaway, C.J., Travers, P. and Walport, M. 2001a. The major histocompatibility complex and its functions. Available at: https://www.ncbi.nlm.nih.gov/books/NBK27156/ [Accessed: 1 May 2017].

Janeaway, C.J., Travers, P., Walport, M. and Shlomchik, M. 2001b. 3-13. The two classes of MHC molecule are expressed differentially on cells. In: *Immunobiology: The Immune System in Health and Disease*. 5th edition. New York: Garland Science. Available at: https://www.ncbi.nlm.nih.gov/books/NBK27098/#A365.

Jarvis, E.D. et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215), pp. 1320 LP – 1331. Available at: http://science.sciencemag.org/content/346/6215/1320.abstract.

Jeannerat, E. et al. 2018. Stallion semen quality depends on major histocompatibility complex matching to teaser mare. *Molecular Ecology* 27(4), pp. 1025–1035. doi: 10.1111/mec.14490.

Jordan, W., and Bruford, M. New perspectives on mate choice and the MHC. *Heredity* 81, 127–133 (1998). https://doi.org/10.1046/j.1365-2540.1998.00428.x

Jouventin, P., Charmantier, A., Dubois, M.P., Jarne, P. and Bried, J. 2007. Extra-pair paternity in the strongly monogamous Wandering Albatross *Diomedea exulans* has no apparent benefits for females. *Ibis* 149, pp. 67–78. doi: 10.1111/j.1474-919X.2006.00597.x.

Juola, F.A. and Dearborn, D.C. 2012. Sequence-based evidence for major histocompatibility complex-disassortative mating in a colonial seabird. *Proceedings of the Royal Society B: Biological Sciences* 279(1726), pp. 153–162. Available at: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.0562.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature* Methods 14(6), pp. 587–589. Available at: https://doi.org/10.1038/nmeth.4285.

Kamiya, T., O'Dwyer, K., Westerdahl, H., Senior, A. and Nakagawa, S. 2014. A quantitative review of MHC-based mating preference: The role of diversity and dissimilarity. *Molecular Ecology* 23(21), pp. 5151–5163. doi: 10.1111/mec.12934.

Khan, I. et al. 2015. Olfactory Receptor Subgenomes Linked with Broad Ecological Adaptations in *Sauropsida*. *Molecular Biology and Evolution* 32(11), pp. 2832–2843. Available at: https://doi.org/10.1093/molbev/msv155.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution* 35(6), pp. 1547–1549. doi: 10.1093/molbev/msy096.

Landry, C., Garant, D., Duchesne, P. and Bernatchez, L. 2001. "Good genes as heterozygosity": the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proceedings of the Royal Society B: Biological Sciences* 268(1473), pp. 1279–1285. Available at: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2001.1659.

Leclaire, S. et al. 2014. Preen secretions encode information on MHC similarity in certain sex-dyads in a monogamous seabird. *Scientific Reports* 4(1), pp. 2–4. doi: 10.1038/srep06920.

Leclaire, S., Strandh, M., Dell'Ariccia, G., Gabirot, M., Westerdahl, H. and Bonadonna, F. 2019. Plumage microbiota covaries with the major histocompatibility complex in blue petrels. *Molecular Ecology* 28(4), pp. 833–846. doi: 10.1111/mec.14993.

Leclaire, S., Strandh, M., Mardon, J., Westerdahl, H. and Bonadonna, F. 2017. Odour-based discrimination of similarity at the major histocompatibility complex in birds.

*Proceedings of the Royal Society B: Biological Sciences* 284(1846), p. 20162466. Available at: http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2016.2466.

Lenz, T.L. and Becker, S. 2008. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. *Gene* 427(1), pp. 117–123. Available at: https://www.sciencedirect.com/science/article/pii/S037811190800471X.

Libois, E., Gimenez, O., Oro, D., Mínguez, E., Pradel, R. and Sanz-Aguilar, A. 2012. Nest boxes: A successful management tool for the conservation of an endangered seabird. *Biological Conservation* 155(March), pp. 39–43. doi: 10.1016/j.biocon.2012.05.020.

Librado, P. and Rozas, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11), pp. 1451–1452. Available at: https://doi.org/10.1093/bioinformatics/btp187.

Lighten, J., van Oosterhout, C., Paterson, I.G., Mcmullan, M. and Bentzen, P. 2014. Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources* 14(4), pp. 753–767. doi: 10.1111/1755-0998.12225.

Locke, D.P. et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American journal of human genetics* 79(2), pp. 275–290. Available at: https://pubmed.ncbi.nlm.nih.gov/16826518.

Maddison, W.P. and Knowles, L.L. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. Systematic *Biology* 55(1), pp. 21–30. Available at: https://doi.org/10.1080/10635150500354928.

Mailund, T., Munch, K. and Schierup, M.H. 2014. Lineage Sorting in Apes. *Annual Review of Genetics* 48(1), pp. 519–535. Available at: https://doi.org/10.1146/annurev-genet-120213-092532.

Manlik, O. et al. 2019. Is MHC diversity a better marker for conservation than neutral genetic diversity? A case study of two contrasting dolphin populations. *Ecology and Evolution* 9(12), pp. 6986–6998. doi: 10.1002/ece3.5265.

Mardon, J. and Bonadonna, F. 2009. Atypical homing or self-odour avoidance? Blue petrels (*Halobaena caerulea*) are attracted to their mate's odour but avoid their own. *Behavioral Ecology and Sociobiology* 63(4), pp. 537–542. doi: 10.1007/s00265-008-0688-z.

Margulies, M. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057), pp. 376–380. Available at: https://doi.org/10.1038/nature03959.

Marmesat, E., Schmidt, K., Saveljev, A.P., Seryodkin, I. V. and Godoy, J.A. 2017. Retention of functional variation despite extreme genomic erosion: MHC allelic repertoires in the Lynx genus. *BMC Evolutionary Biology* 17(1), pp. 1–16. doi: 10.1186/s12862-017-1006-z.

Mayr, E. 2013. Animal Species and Evolution: . *Harvard University Press*. Available at: https://doi.org/10.4159/harvard.9780674865327.

McDonald, P.G. and Griffith, S.C. 2011. To pluck or not to pluck: the hidden ethical and scientific costs of relying on feathers as a primary source of DNA. *Journal of Avian*

*Biology* 42(3), pp. 197–203. Available at: https://doi.org/10.1111/j.1600-048X.2011.05365.x.

Milinski, M., Griffiths, S., Wegner, K.M., Reusch, T.B.H., Haas-Assenbaum, A. and Boehm, T. 2005. Mate choice decisions of stickleback females predictably modified by MHC peptide ligands. *Proceedings of the National Academy of Sciences* 102(12), pp. 4414–4418. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.0408264102.

Miller, H.C., Belov, K. and Daugherty, C.H. 2005. Characterization of MHC class II genes from an ancient reptile lineage, *Sphenodon* (*tuatara*). *Immunogenetics* 57(11), pp. 883–891. Available at: https://doi.org/10.1007/s00251-005-0055-4.

Miller, H.C. and Lambert, D.M. 2004. Gene duplication and gene conversion in class II MHC genes of New Zealand robins (*Petroicidae*). *Immunogenetics* 56(3), pp. 178–191. doi: 10.1007/s00251-004-0666-1.

Minh, B.Q., Nguyen, M.A.T. and Von Haeseler, A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* 30(5), pp. 1188–1195. doi: 10.1093/molbev/mst024.

Minias, P., Bateson, Z.W., Whittingham, L.A., Johnson, J.A., Oyler-Mccance, S. and Dunn, P.O. 2016. Contrasting evolutionary histories of MHC class i and class II loci in grouse - Effects of selection and gene conversion. *Heredity* 116(5), pp. 466–476. Available at: http://dx.doi.org/10.1038/hdy.2016.6.

Minias, P., Pikus, E., Whittingham, L.A. and Dunn, P.O. 2018. A global analysis of selection at the avian MHC. *Evolution* 72(6), pp. 1278–1293. doi: 10.1111/evo.13490.

Minias, P., Pikus, E., Whittingham, L.A. and Dunn, P.O. 2019. Evolution of copy number at the MHC varies across the avian tree of life. *Genome Biology and Evolution* 11(1), pp. 17–28. doi: 10.1093/gbe/evy253.

Monteiro, L.R. and Furness, R.W. 1998. Speciation through temporal segregation of Madeiran storm petrel (*Oceanodroma castro*) populations in the Azores? *Philosophical Transactions of the Royal Society B: Biological Sciences* 353(1371), pp. 945–953. doi: 10.1098/rstb.1998.0259.

Monteiro, L.R., Ramos, R.A. and Furness, R.W. 1996. Past and present status and conservation of the seabirds breeding in the Azores archipelago. *Biological Conservation* 78, pp. 319–328.

Mueller, J.C., Kuhl, H., Boerno, S., Tella, J.L., Carrete, M. and Kempenaers, B. 2018. Evolution of genomic variation in the burrowing owl in response to recent colonization of urban areas. *Proceedings of the Royal Society B: Biological Sciences* 285(1878), p. 20180206. Available at: https://doi.org/10.1098/rspb.2018.0206.

Nava, C., Neves, V.C., Andris, M., Dubois, M.P., Jarne, P., Bolton, M. and Bried, J. 2017. Reduced population size does not affect the mating strategy of a vulnerable and endemic seabird. *Die Naturwissenschaften* 104(11–12), p. 103. doi: 10.1007/s00114-017-1523-z.

Neff, B.D. and Pitcher, T.E. 2005. Genetic quality and sexual selection: An integrated framework for good genes and compatible genes. *Molecular Ecology* 14(1), pp. 19–38. doi: 10.1111/j.1365-294X.2004.02395.x.

Nei, M., Gu, X. and Sitnikova, T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National*

*Academy of Sciences* 94(15), pp. 7799–7806. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.94.15.7799.

Nei, M. and Rooney, A.P. 2005. Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics* 39(1), pp. 121–152. Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.39.073003.112240.

Neves, V., Nava, C., Monteiro, E., Monteiro, P. and Bried, J. 2017. Depredation of Monteiro's Storm-Petrel (*Hydrobates monteiroi*) Chicks by Madeiran Wall Lizards (*Lacerta dugesii*). *Waterbirds (in press)*. doi: 10.1675/063.040.0113.

Nevitt, G.A. 2008. Sensory ecology on the high seas: the odor world of the *procellariiform* seabirds. *The Journal of experimental biology* 211(Pt 11), pp. 1706–1713. doi: 10.1242/jeb.015412.

Nevitt, G.A. 2000. Olfactory Foraging by Antarctic *Procellariiform* Seabirds: Life at High Reynolds Numbers. (April), pp. 245–253.

Nevitt, G.A., Losekoot, M. and Weimerskirch, H. 2008. Evidence for olfactory search in wandering albatross, *Diomedea exulans*. *Proceedings of the National Academy of Sciences*, 105 (12) 4576-4581; DOI: 10.1073/pnas.0709047105

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32(1), pp. 268–274. Available at: https://doi.org/10.1093/molbev/msu300.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends in ecology & evolution* 16(7), pp. 358–364. doi: 10.1016/s0169-5347(01)02203-0.

Nosil, P. 2008. Speciation with gene flow could be common. *Molecular Ecology* 17(9), pp. 2103–2106. doi: 10.1111/j.1365-294X.2008.03715.x.

O'Connor, E.A., Westerdahl, H., Burri, R. and Edwards, S. V. 2019. Avian MHC Evolution in the Era of Genomics: Phase 1.0. *Cells* 8(10), pp. 1–21. doi: 10.3390/cells8101152.

Ohta, T. 1998. On the Pattern of Polymorphisms at Major Histocompatibility Complex Loci., *Journal of molecular evolution*, 46(6), pp. 633–638.

Oksanen, J. et al. 2015. Vegan: Community Ecology Package. R Package Version 2.2-1 2, pp. 1–2.

Oliveira, N. et al. 2016. Status Report for Monteiro's Storm-petrel Hydrobates monteiroi., p. 24.

Owen, J.C. 2011. Collecting, processing, and storing avian blood: A review. *Journal of Field Ornithology* 82(4), pp. 339–354. doi: 10.1111/j.1557-9263.2011.00338.x.

Turbek, S.P., Browne, M., Di Giacomo, A.S., Kopuchian, C., Hochachka, W.M., Estalles, C., Lijtmaer, D.A., Tubaro, PL, Silveira L.F., Lovette, I.J., Safran, R.J., Taylor, S.A., Campagna, L. 2021. Rapid speciation via the evolution of pre-mating isolation in the Iberá Seedeater. *Science* 371(6536), p. eabc0256. Available at: https://doi.org/10.1126/science.abc0256.

Paiva, V., Ramos, J., Nava, C., Neves, V., Bried, J. and Magalhães, M. 2017. Inter-sexual habitat and isotopic niche segregation of the endangered Monteiro's storm-petrel during breeding. *Zoology* 126. doi: 10.1016/j.zool.2017.12.006.

Palomares, F., Godoy, J.A., Piriz, A. and O'Brien, S.J. 2002. Faecal genetic analysis to determine the presence and distribution of elusive carnivores: design and feasibility for

the Iberian lynx. *Molecular Ecology* 11(10), pp. 2171–2182. Available at: https://doi.org/10.1046/j.1365-294X.2002.01608.x.

Pardal, S., Alves, J.A., Mota, P.G. and Ramos, J.A. 2018. Dressed to impress: breeding plumage as a reliable signal of innate immunity. *Journal of Avian Biology* 49(7), pp. 1–13. doi: 10.1111/jav.01579.

Pearse-Pratt, R., Schellinck, H., Brown, R., Singh, P.B. and Roser, B. 1998. Soluble MHC antigens and olfactory recognition of genetic individuality: The mechanism. *Genetica* 104(3), pp. 223–230. doi: 10.1023/A:1026489524199.

Penn, D. and Potts, W. 1998. How do major histocompatibility complex genes influence odor and mating preferences? *Advances in immunology* 69, pp. 411–436. doi: 10.1016/s0065-2776(08)60612-4.

Penn, D.J. 2002. The Scent of Genetic Compatibility: Sexual Selection and the Major Histocompatibility Complex. *Ethology* 108(1), pp. 1–21. Available at: https://doi.org/10.1046/j.1439-0310.2002.00768.x.

Penn, D.J., Damjanovich, K. and Potts, W.K. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences* 99(17), pp. 11260–11264. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.162006499.

Pereira, A.G. and Schrago, C.G. 2018. Incomplete lineage sorting impacts the inference of macroevolutionary regimes from molecular phylogenies when concatenation is employed: An analysis based on Cetacea. *Ecology and Evolution* 8(14), pp. 6965–6971. doi: 10.1002/ece3.4212.

Peters, J.L., Zhuravlev, Y., Fefelov, I., Logie, A. and Omland, K.E. 2007. Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas* spp.). *Evolution; international journal of organic evolution* 61(8), pp. 1992–2006. doi: 10.1111/j.1558-5646.2007.00149.x.

Piertney, S.B. and Oliver, M.K. 2006. The evolutionary ecology of the major histocompatibility complex., *Heredity* 96(1) (2006): pp. 7-21.

Poddar, S.K. 2000. Symmetric vs asymmetric PCR and molecular beacon probe in the detection of a target gene of adenovirus*. Molecular and cellular probes* 14(1), pp. 25–32. doi: 10.1006/mcpr.1999.0278.

Potts, W.K. and Wakeland, E.K. 1990. Evolution of diversity at the major histocompatibility complex. *Trends in ecology & evolution* 5(6), pp. 181–187. doi: 10.1016/0169-5347(90)90207-T.

Quillfeldt, P., Schmoll, T., Peter, H.-U., Epplen, J.T. and Lubjuhn, T. 2001. Genetic Monogamy in Wilson's Storm-Petrel. *The Auk* 118(1), pp. 242–248. Available at: https://doi.org/10.1093/auk/118.1.242.

Rambaut, A. 2009. FigTree v1.3.1. http://tree.bio.ed.ac.uk/software/figtree/ . Available at: http://ci.nii.ac.jp/naid/10030433668/en/.

Reche, P.A. and Reinherz, E.L. 2003. Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. *Journal of Molecular Biology* 331(3), pp. 623–641. Available at: https://www.sciencedirect.com/science/article/pii/S0022283603007502.

Rekdal, S.L., Anmarkrud, J.A., Johnsen, A. and Lifjeld, J.T. 2018. Genotyping strategy matters when analyzing hypervariable major histocompatibility complex-Experience from a passerine bird. *Ecology and Evolution* 8(3), pp. 1680–1692. doi: 10.1002/ece3.3757.

Rekdal, S.L., Anmarkrud, J.A., Lifjeld, J.T. and Johnsen, A. 2019. Extra-pair mating in a passerine bird with highly duplicated major histocompatibility complex class II: Preference for the golden mean. *Molecular Ecology* 28(23), pp. 5133–5144. doi: 10.1111/mec.15273.

Richardson, D.S. and Westerdahl, H. 2003. MHC diversity in two *Acrocephalus* species: the outbred Great reed warbler and the inbred Seychelles warbler. *Molecular ecology* 12(12), pp. 3523–3529. doi: 10.1046/j.1365-294x.2003.02005.x.

Riera Romo, M., Pérez-Martínez, D. and Castillo Ferrer, C. 2016. Innate immunity in vertebrates: An overview. *Immunology* 148(2), pp. 125–139. doi: 10.1111/imm.12597.

Ritchie, M.G. 2007. Sexual Selection and Speciation. Annual Review of Ecology, *Evolution, and Systematics* 38, pp. 79–102. Available at: http://www.jstor.org/stable/30033853.

Rivero-de Aguilar, J., Westerdahl, H., Martinez-de la Puente, J., Tomas, G., Martinez, J. and Merino, S. 2016. MHC-I provides both quantitative resistance and susceptibility to blood parasites in blue tits in the wild. *Journal of Avian Biology* 47(5), pp. 669–677. doi: 10.1111/jav.00830.

Robert, A. et al. 2014. Nest fidelity is driven by multi-scale information in a long-lived seabird. *Proceedings of the Royal Society Biological sciences* 281(September), p. 20141692. doi: 10.1098/rspb.2014.1692.

Rock, K.L., Reits, E. and Neefjes, J. 2016. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends in Immunology* 37(11), pp. 724–737. Available at: http://dx.doi.org/10.1016/j.it.2016.08.010.

Roper, T.J. 1999. Olfaction in Birds. *Academic Press*, pp. 247–332. Available at: https://www.sciencedirect.com/science/article/pii/S0065345408602193.

Ruff, J.S., Nelson, A.C., Kubinak, J.L. and Potts, W.K. 2012. MHC signaling during social communication. *Advances in experimental medicine and biology* 738, pp. 290–313. Available at: https://pubmed.ncbi.nlm.nih.gov/22399386.

Salibian, A. and Montalti, D. 2009. Physiological and Biochemical Aspects of the Avian Uropygial Gland. *Brazilian journal of biology*, 69, pp. 437–446. doi: 10.1590/S1519-69842009000200029.

Sangster, G., Collinson, J.M., Crochet, P., Knox, A.G., Parkin, D.T. and Votier, S.C. 2012. Taxonomic recommendations for British birds: eighth report., pp. 874–883.

Santos, P.S.C. et al. 2016. MHC-dependent mate choice is linked to a trace-amine-associated receptor gene in a mammal. *Scientific Reports* 6(9), pp. 1–9. doi: 10.1038/srep38490.

Santos, P.S.C., Mezger, M., Kolar, M., Michler, F.U. and Sommer, S. 2018. The best smellers make the best choosers: Mate choice is affected by female chemosensory receptor gene diversity in a mammal. Proceedings of the *Royal Society B: Biological Sciences* 285(1893). doi: 10.1098/rspb.2018.2426.

Santos, P.S.C., Michler, F.U. and Sommer, S. 2017. Can MHC-assortative partner choice promote offspring diversity? A new combination of MHC-dependent behaviours among

sexes in a highly successful invasive mammal. *Molecular Ecology* 26(8), pp. 2392–2404. doi: 10.1111/mec.14035.

Schnell, I.B., Bohmann, K. and Gilbert, M.T.P. 2015. Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* 15(6), pp. 1289–1303. doi: 10.1111/1755-0998.12402.

Seddon, J.M. and Ellegren, H. 2004. A temporal analysis shows major histocompatibility complex loci in the Scandinavian wolf population are consistent with neutral evolution. *Proceedings. Biological sciences* 271(1554), pp. 2283–2291. doi: 10.1098/rspb.2004.2869.

Segelbacher, G. 2002. Noninvasive genetic analysis in birds: testing reliability of feather samples. *Molecular Ecology Notes* 2(3), pp. 367–369. Available at: https://doi.org/10.1046/j.1471-8286.2002.00180.x-i2.

Sepil, I., Radersma, R., Santure, A.W., De Cauwer, I., Slate, J. and Sheldon, B.C. 2015. No evidence for MHC class I-based disassortative mating in a wild population of great tits. *Journal of Evolutionary Biology* 28(3), pp. 642–654. doi: 10.1111/jeb.12600.

Seutin, G., White, B.N. and Boag, P.T. 1991. Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology* 69(1), pp. 82–90. Available at: http://www.nrcresearchpress.com/doi/abs/10.1139/z91-013.

Sikkema-Raddatz, B. et al. 2013. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Human mutation* 34(7), pp. 1035–1042. doi: 10.1002/humu.22332.

Silva, M.F., Smith, A.L., Friesen, V.L., Bried, J., Hasegawa, O., Coelho, M.M. and Silva, M.C. 2016. Mechanisms of global diversification in the marine species Madeiran Storm-petrel *Oceanodroma castro* and Monteiro's Storm-petrel *O. monteiroi*: Insights from a multi-locus approach. *Molecular Phylogenetics and Evolution* 98, pp. 314–323. Available at: http://dx.doi.org/10.1016/j.ympev.2016.02.014.

Silva, M.C., Chibucos, M., Munro, J.B., Daugherty, S., Coelho, M.M. and C. Silva, J. 2020. Signature of adaptive evolution in olfactory receptor genes in Cory's Shearwater supports molecular basis for smell in *procellariiform* seabirds. *Scientific Reports* 10(1), p. 543. Available at: https://doi.org/10.1038/s41598-019-56950-6.

Singh, P.B. 1999. The present status of the 'carrier hypothesis ' for chemosensory recognition of genetic individuality., pp. 231–233.

Singh, P.B., Roser, B. and Brown, R.E. 1987. MHC antigens in urine as olfactory recognition cues. *Nature* 327(6118), pp. 161–164. doi: 10.1038/327161a0.

Slade, J.W.G., Watson, M.J. and MacDougall-Shackleton, E.A. 2019. "Balancing" balancing selection? Assortative mating at the major histocompatibility complex despite molecular signatures of balancing selection. *Ecology and Evolution* 9(9), pp. 5146–5157. doi: 10.1002/ece3.5087.

Smith, A.L. and Friesen, V.L. 2007. Differentiation of sympatric populations of the band-rumped storm-petrel in the Galapagos Islands: An examination of genetics, morphology, and vocalizations. *Molecular Ecology* 16(8), pp. 1593–1603. doi: 10.1111/j.1365-294X.2006.03154.x.

Smith, A.L., Monteiro, L., Hasegawa, O. and Friesen, V.L. 2007. Global phylogeography of the band-rumped storm-petrel (*Oceanodroma castro; Procellariiformes: Hydrobatidae*).

*Molecular Phylogenetics and Evolution* 43(3), pp. 755–773. doi: 10.1016/j.ympev.2007.02.012.

Smith, N.C., Rise, M.L. and Christian, S.L. 2019. A Comparison of the Innate and Adaptive Immune Systems in Cartilaginous Fish, Ray-Finned Fish, and Lobe-Finned Fish. *Frontiers in Immunology* 10(October). doi: 10.3389/fimmu.2019.02292.

Sommer, S. 2005a. Major histocompatibility complex and mate choice in a monogamous rodent. *Behavioral Ecology and Sociobiology* 58(2), pp. 181–189. Available at: https://doi.org/10.1007/s00265-005-0909-7.

Sommer, S. 2005b. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in zoology* 2, p. 16. Available at: https://pubmed.ncbi.nlm.nih.gov/16242022.

Sommer, S., Courtiol, A. and Mazzoni, C.J. 2013. MHC genotyping of non-model organisms using next-generation sequencing: A new methodology to deal with artefacts and allelic dropout. *BMC Genomics* 14(1). doi: 10.1186/1471-2164-14-542.

Spurgin, L.G. and Richardson, D.S. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences* 277(1684), pp. 979–988. doi: 10.1098/rspb.2009.2084.

Steiger, S.S., Fidler, A.E., Valcu, M. and Kempenaers, B. 2008. Avian olfactory receptor gene repertoires: Evidence for a well-developed sense of smell in birds? *Proceedings of the Royal Society B: Biological Sciences* 275(1649), pp. 2309–2317. doi: 10.1098/rspb.2008.0607.

Stephens, M., Smith, N.J. and Donnelly, P. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68(4), pp. 978–989. Available at: https://www.sciencedirect.com/science/article/pii/S0002929707614244.

Strand, T., Wang, B., Meyer-Lucht, Y. and Höglund, J. 2013. Evolutionary history of black grouse major histocompatibility complex class IIB genes revealed through single locus sequence-based genotyping. *BMC genetics* 14, p. 29. Available at: https://pubmed.ncbi.nlm.nih.gov/23617616.

Strandh, M. et al. 2012. Major histocompatibility complex class II compatibility, but not class I, predicts mate choice in a bird with highly developed olfaction. *Proceedings of the Royal Society B: Biological Sciences* 279(1746), pp. 4457–4463. doi: 10.1098/rspb.2012.1562.

Strandh, M., Lannefors, M., Bonadonna, F. and Westerdahl, H. 2011. Characterization of MHC class I and II genes in a subantarctic seabird, the blue petrel, *Halobaena caerulea (Procellariiformes). Immunogenetics* 63(10), pp. 653–666. doi: 10.1007/s00251-011-0534-8.

Stuglik, M.T., Radwan, J. and Babik, W. 2011. jMHC: Software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources* 11(4), pp. 739–742. doi: 10.1111/j.1755-0998.2011.02997.x.

Taberlet, P., Bonin, A., Zinger, L. and Coissac, É. 2018. Environmental DNA: For *Biodiversity Research and Monitoring*. doi: 10.1093/oso/9780198767220.001.0001.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3), pp. 585–595.

Taylor, R.S. et al. 2019. Cryptic species and independent origins of allochronic populations within a seabird species complex (*Hydrobates* spp.). *Molecular Phylogenetics and Evolution* 139(July), p. 106552. Available at: https://doi.org/10.1016/j.ympev.2019.106552.

Taylor, R.S. and Friesen, V.L. 2017. The role of allochrony in speciation. *Molecular Ecology* 26(13), pp. 3330–3342. doi: 10.1111/mec.14126.

Toews, D.P.L. and Brelsford, A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology* 21(16), pp. 3907–3930. doi: 10.1111/j.1365-294X.2012.05664.x.

Toews, D.P.L., Brelsford, A., Grossen, C., Milá, B. and Irwin, D.E. 2016. Genomic variation across the Yellow-rumped Warbler species complex. *The Auk* 133(4), pp. 698–717. Available at: https://doi.org/10.1642/AUK-16-61.1.

Tregenza, T. and Wedell, N. 2000. Genetic compatibility, mate choice and patterns of parentage: Invited review. *Molecular Ecology* 9(8), pp. 1013–1027. doi: 10.1046/j.1365-294X.2000.00964.x.

Tucker, D.O.N. 1965. Electrophysiological Evidence for Olfactory Function in Birds. *Nature* 207(4992), pp. 34–36. Available at: https://doi.org/10.1038/207034a0.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research* 40(15), pp. 1–12. doi: 10.1093/nar/gks596.

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L. and Guschanski, K. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources* 20(5), pp. 1171–1181. Available at: https://doi.org/10.1111/1755-0998.13009.

Verheyden, C. and Jouventin, P. 1994. Verheyden, C., & Jouventin, P. (1994). Olfactory behavior of foraging procellariiforms. *The Auk, 111*(2), 285–291.

Wang, K. et al. 2008. Modeling genetic inheritance of copy number variations. *Nucleic acids research* 36(21), pp. e138–e138. Available at: https://pubmed.ncbi.nlm.nih.gov/18832372.

Wang, Z., Zhou, X., Lin, Q., Fang, W. and Chen, X. 2013. Characterization, Polymorphism and Selection of Major Histocompatibility Complex (MHC) DAB Genes in Vulnerable Chinese Egret (*Egretta eulophotes*). *PLoS ONE* 8(9). doi: 10.1371/journal.pone.0074185.

Weber, D.S. et al. 2013. Low MHC variation in the polar bear: implications in the face of Arctic warming? *Animal Conservation* 16(6), pp. 671–683. Available at: https://doi.org/10.1111/acv.12045.

Wegner, K.M., Kalbe, M., Rauch, G., Kurtz, J., Schaschl, H. and Reusch, T.B.H. 2006. Genetic variation in MHC class II expression and interactions with MHC sequence polymorphism in three-spined sticklebacks. *Molecular Ecology* 15(4), pp. 1153–1164. Available at: https://doi.org/10.1111/j.1365-294X.2006.02855.x.

Weimer, E. and Sherwood, K. 2019. Point-Counterpoint Series: Confirmation of Homozygous HLA alleles: Is it a Necessity? *Human Immunology* 80. doi: 10.1016/j.humimm.2019.01.002.

Welch, A.J., Yoshida, A.A. and Fleischer, R.C. 2011. Mitochondrial and nuclear DNA sequences reveal recent divergence in morphologically indistinguishable petrels. *Molecular ecology* 20(7), pp. 1364–1377. doi: 10.1111/j.1365-294X.2011.05008.x.

Wenzel B.M. (1992) The Puzzle of Olfactory Sensitivity in Birds. In: Doty R.L., Müller-Schwarze D. (eds) Chemical Signals in Vertebrates 6. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-9655-1_68

Whittaker, D.J., Gerlach, N.M., Soini, H.A., Novotny, M. V. and Ketterson, E.D. 2013. Bird odour predicts reproductive success. *Animal Behaviour* 86(4), pp. 697–703. Available at: http://dx.doi.org/10.1016/j.anbehav.2013.07.025.

Whittaker, D.J., Rosvall, K.A., Slowinski, S.P., Soini, H.A., Novotny, M. V. and Ketterson, E.D. 2018. Songbird chemical signals reflect uropygial gland androgen sensitivity and predict aggression: implications for the role of the periphery in chemosignaling. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 204(1), pp. 5–15. Available at: http://dx.doi.org/10.1007/s00359-017-1221-5.

Whittaker, D.J., Soini, H.A., Atwell, J.W., Hollars, C., Novotny, M. V. and Ketterson, E.D. 2010. Songbird chemosignals: Volatile compounds in preen gland secretions vary among individuals, sexes, and populations. *Behavioural Ecology* 21(3), pp. 608–614. doi: 10.1093/beheco/arq033.

Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F. and Freund, C. 2017. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in immunology* 8, p. 292. Available at: https://pubmed.ncbi.nlm.nih.gov/28367149.

Wittzell, H., Bernot, A., Auffray, C. and Zoorob, R. 1999. Concerted evolution of two Mhc class II B loci in pheasants and domestic chickens. *Molecular Biology and Evolution* 16(4), pp. 479–490. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a026130.

Worley, K., Collet, J., Spurgin, L.G., Cornwallis, C., Pizzari, T. and Richardson, D.S. 2010. MHC heterozygosity and survival in red junglefowl. *Molecular Ecology* 19, pp. 3064–3075. doi: 10.1111/j.1365-294X.2010.04724.x.

Worley, K., Gillingham, M., Jensen, P., Kennedy, L.J., Pizzari, T., Kaufman, J. and Richardson, D.S. 2008. Single locus typing of MHC class I and class II B loci in a population of red jungle fowl. *Immunogenetics* 60(5), pp. 233–247. doi: 10.1007/s00251-008-0288-0.

Yamazaki, K. 1976. Control of mating preferences in mice by genes in the major histocompatibility complex. *Journal of Experimental Medicine* 144(5), pp. 1324–1335. doi: 10.1084/jem.144.5.1324.

Zavodna, M., Grueber, C.E. and Gemmell, N.J. 2013. Parallel Tagged Next-Generation Sequencing on Pooled Samples – A New Approach for Population Genetics in Ecology and Conservation. *PLOS ONE* 8(4), p. e61471. Available at: https://doi.org/10.1371/journal.pone.0061471.

Zhang, G. et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science (New York, N.Y.) 346(6215), pp. 1311–1320. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4390078/.

Zink, R.M. and Barrowclough, G.F. 2008. Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology* 17(9), pp. 2107–2121. doi: 10.1111/j.1365-294X.2008.03737.x.

Zomer, S. et al. 2009. Consensus multivariate methods in gas chromatography mass spectrometry and denaturing gradient gel electrophoresis: MHC-congenic and other strains of mice can be classified according to the profiles of volatiles and microflora in their scent-marks*. Analyst* 134(1), pp. 114–123. doi: 10.1039/b807061j.