NUCLEAR AND MITOCHONDRIAL DNA EVOLUTION IN ADÉLIE PENGUINS:

STUDIES OF MODERN AND ANCIENT POPULATIONS

A thesis presented in partial fulfillment of the requirements for the degree of **Doctor of Philosophy (PhD)** in **Genetics**

Allan Wilson Centre for Molecular Ecology and Evolution
Institute for Natural Sciences
Massey University
Auckland, New Zealand

GABRIELLE ANGELA BEANS PICÓN

2012

Teatro mágico - sólo para locos

La entrada cuesta la razón

- Hermann Hesse

ABSTRACT

The Adélie penguin of Antarctica (*Pygoscelis adeliae*) breeds on the Antarctic continent and on offshore islands. Its evolutionary history has been, and its current biology remains, dependent on a range of climate variables. Over geological time, glacial warming and cooling periods have resulted in Adélie penguin populations decreasing and expanding. Therefore, understanding Adélie penguin population dynamics at a genetic level can provide insights into how the species responds to changing climates, one reason why Adélie penguins are an important natural model species. In addition, sub-fossil bone deposits of this species below modern and abandoned colonies provide an excellent source of ancient DNA that can bring a temporal dimension to population studies of the species. In combination, these attributes enable us to address some fundamental questions regarding evolutionary change.

Making use of known mitochondrial DNA mutation rates and current population sizes, a positive and significant correlation between population size and modern mitochondrial control region diversity was detected. This finding supports the use of mitochondrial DNA for population inferences. Effective population sizes of breeding colonies are shown to have increased since the late Pleistocene. To extend current tools available for understanding Adélie penguins, six nuclear intron loci were recovered from a wide range of introns that can be applied to population genetics and phylogenetic studies of penguins. Five introns were used to investigate the persistence of the mitochondrial Antarctic (A) and Ross Sea (RS) lineages. No evidence for the existence of these lineages was found in the nuclear loci sequenced. A signature of historical population expansion, preceding the mitochondrial one, was detected. The utility of four introns in resolving penguin phylogenetic signals was also determined. Non-coding nuclear sequence of one intron were obtained from ancient sub-fossil remains of Adélie penguins using multiplex PCR enrichment, followed by second-generation sequencing of a barcoded library. A shift in haplotype frequencies was detected between ancient and modern intron sequences in Adélie penguins, despite a small sample size. In the future, advancing the current methodologies and extending sampling to additional introns as well as older samples, is likely to provide a new level of understanding of this remarkable species.

ACKNOWLEDGEMENTS

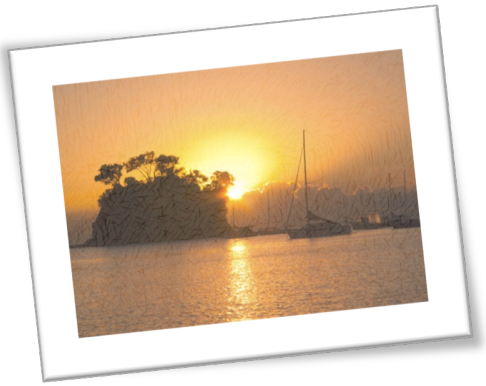
It's hard to believe so much time has gone by since I first landed in New Zealand, eager to begin an adventure on different levels. Certainly, I have experienced more than I would have anticipated on that sunny day, May 31st, 2006, when I thought the cab driver was messing with me by taking such a bizarre suburban route from the airport to the lab at Massey in Albany. It's the end of one road now, finally, and my PhD adventure is winding down as I write these words. There are, of course, an innumerable number of people I would like to thank in great detail for their support, both academically and on a personal level. I suppose everyone feels this way!

First, and foremost, I owe great thanks to my supervisor, David Lambert, for a great number of things; for offering me the chance to come to New Zealand, for letting me explore my own ideas for different projects, for encouraging me to support the All Blacks, for helping me find positive approaches to my writing when I was deep in thesis-end negativity. Thank you, as well, and Sherene and Christine, for welcoming me into your home in Brisbane and giving my thesis a jump-start when it most needed it.

Thanks to my co-supervisor Austen Ganley, for useful discussions and moral support when I needed it, among other things. Special thanks to Leon Huynen and Sankar Subramanian, who gave me invaluable help, particularly in my last panicked moments when I unreasonably needed things done quickly. You really came through for me, and I greatly appreciate it. Thanks also to my examiners for helpful comments and for giving me the PhD in the end!

Thanks to everyone who contributed to different aspects of this research, either directly or through previous work on Adélie penguins. Thanks to everyone who

collected samples before my arrival (though I do wish I could have gone and collected some myself, it's incredibly helpful to have freezers and fridges full of bones and blood when you start off), and everyone who helped produce the mitochondrial DNA sequences I analyzed for part of this thesis. Thanks also to everyone involved in radio-carbondating Adélie subfossil bones. Thanks to Phil Lyver for collaborating with us and providing demographic Adélie data, as well as for being great to chat to. Thanks to Michael Knapp for helping me get to grips with tagging a load of ancient multiplexed introns for my FLX run and for being a great host at Otago University, and of course thanks to Lisa Matisoo-Smith for welcoming me to her group for those few weeks I was visiting. Thanks to Allan Baker and Oliver Ryder for "other" penguin samples and DNA. Thanks especially to Oliver for extracting DNA for me twice due to NZ customs mysteriously keeping one DNA shipment! Thanks to Tim for showing me round the Australian bush and diving.



Thanks to the Allan Wilson Centre, Massey University, the New Zealand Postgraduate Study Abroad Award, Massey University Institute of Molecular Biosciences and Massey University Institute of Natural Sciences for funding my PhD scholarship and conference travel (Wellington, Hawaii, Christchurch, Kaikoura, Barcelona oh my!).

I can't possibly fit in here all the friends from afar that have been there for me, and the countless friends I have made since moving to New Zealand. Every one of them has played a part in making my life here memorable in every way imaginable. From hot tub parties to laid back barbecues, theme parties to Pohutukawa-filled wild Christmases, from Tongan beach escapes to South Island winter and summer adventures, from mountain biking to diving to hiking to just simply walking around the East Coast Bays, from east coast sunrises to epic west coast sunsets… I have loved making New Zealand my home and you are all the main reason it feels that way. Thank you! Special thanks to Katie, Jyothsna, Martina, Hayley, Eli, Jarod, Phil,

all flatmates past and present, AUUC, Jarod, Arapeta & Ra (for giving me my own taonga) and everyone who spoke to me in Building 11 since the beginning. Thanks to anyone who came to visit while I was here.

Yes, of course, thanks to my family, who from the beginning were nothing but supportive (though I do recall my father trying to get me a PhD position in Austria so I wouldn't leave Europe). Thanks for not minding too much that I was literally on the other side of the world! Thanks for supporting me in difficult times (there have been a few, I can't lie), and for coming all the way out here to try and understand why I like it so much (not to mention sustaining the NZ economy by purchasing every possible kind of souvenir). I am incredibly lucky to have so much love in my life, and though I may be a bit difficult sometimes, I never take it for granted. Everything I am is thanks to you.

Eric, you came into this story near the end, during the most difficult and also the most joyous moments. Thank you for putting up with me during my thesis meltdowns, thank you for keeping me going, encouraging me, and always bringing hope and happiness with you.

Finally, thanks, beautiful, green, ocean-fringed Aotearoa. No matter where I end up, I will love this quirky little country forever. Verde que te quiero verde!

Table of Contents

ABSTRACT I

ACKNOWLEDGEMENTS II

TABLE OF CONTENTS V

LIST OF FIGURES X

LIST OF TABLES XIII

CHAPTER ONE 1

INTRODUCTION

1.1 THESIS STRUCTURE 1

1.2 ADÉLIE PENGUINS 2

1.2.1 MOLECULAR ECOLOGY AND EVOLUTION IN ADÉLIE PENGUINS 4

1.3 MOLECULAR MARKERS 6

1.4 MITOCHONDRIAL DNA 8

I.4.1 THE RELATIONSHIP BETWEEN MITOCHONDRIAL DNA DIVERSITY AND POPULATION SIZE 10

I.4.2 CRITICISM OF MTDNA IN POPULATION GENETICS AND PHYLOGENETICS 13

1.5 NUCLEAR INTRONS 14

1.5.1 FOUR GROUPS OF INTRONS 15

1.5.2 SPLICEOSOMAL INTRON EVOLUTION 16

1.5.3 USING INTRONS IN PHYLOGENETICS AND POPULATION GENETICS 20

1.5.4 NUCLEAR INTRONS FOR PENGUINS 23

1.6 EVOLUTIONARY RATES 23

1.6.1 METHODS OF CALCULATING EVOLUTIONARY RATES 24

1.6.2 APPARENT TIME-DEPENDENCY OF RATES 25

1.7 INTRON EVOLUTIONARY RATES AND THE POTENTIAL OF ANCIENT DNA TECHNIQUES 26

1.8 AIMS OF THIS PHD 29

1.9 REFERENCES 30

CHAPTER TWO 41

GENETIC DIVERSITY AND EFFECTIVE POPULATION SIZE OF ADÉLIE PENGUINS

2.1 ABSTRACT 41

2.2 INTRODUCTION 42

2.3 MATERIAL AND METHODS 44

2.3.1 ANNUAL COUNT OF ADÉLIE PENGUIN BREEDING PAIRS 44

2.3.2 SAMPLES, DNA EXTRACTION, PCR AMPLIFICATION AND SEQUENCING 45

2.3.3 SUMMARY STATISTICS AND POPULATION STRUCTURE (AMOVA) 47

2.3.4 GENETIC ESTIMATES OF LONG-TERM EFFECTIVE POPULATION SIZES OF COLONIES OF ADÉLIE PENGUINS IN THE ROSS SEA 48

2.4 RESULTS 49

2.4.1 SUMMARY STATISTICS AND POPULATION STRUCTURE 49

2.4.2 ESTIMATES OF GENETIC DIVERSITY (θ) AND FEMALE EFFECTIVE POPULATION SIZE (N_{EF}) 51

2.5 DISCUSSION 55

2.6 REFERENCES 61

CHAPTER THREE 65

EFFECTIVE POPULATION SIZE OF THE EXTINCT HUIA

3.1 ABSTRACT 65

3.2 INTRODUCTION 66

3.3 MATERIAL AND METHODS 69

3.3.1 SAMPLES 69

3.3.2 ANCIENT DNA METHODS 71

3.3.3 HUIA MITOCHONDRIAL HYPERVARIABLE REGION SEQUENCES 72

3.3.4 ESTIMATING GENETIC DIVERSITY AND EFFECTIVE POPULATION SIZE IN HUIA 72

3.4 RESULTS 73

3.4.1 ESTIMATING GENETIC DIVERSITY AND POPULATION SIZE OF HUIA 73

3.5 DISCUSSION 75

3.6 REFERENCES 78

CHAPTER FOUR **81**

INTRON RECOVERY IN ADÉLIE PENGUINS

4.1 ABSTRACT **81**
4.2 INTRODUCTION **82**
4.3 METHODS **85**
4.3.1 CHOOSING INTRON MARKERS FOR ADÉLIE PENGUINS 85
4.3.2 DNA EXTRACTIONS 86
4.3.3 SCREENING INTRON MARKERS IN MODERN ADÉLIE PENGUINS 86
4.4 RESULTS **92**
4.4.1 LITERATURE AND NCBI SEARCH 92
4.4.2 PRIMER SCREEN AND INITIAL INTRON SEQUENCING RESULTS 93
4.5 DISCUSSION **97**
4.6 REFERENCES **101**

CHAPTER FIVE **107**

USING INTRONS TO ELUCIDATE ADÉLIE PENGUIN POPULATION GENETICS AND PENGUIN PHYLOGENY

5.1 ABSTRACT **107**
5.2 INTRODUCTION **108**
5.3 METHODS **110**
5.3.1 DNA EXTRACTIONS 110
5.3.2 PCR AND DIRECT SEQUENCING 112
5.3.3 SEQUENCE ANALYSIS AND PHASING OF INTRONS 113
5.3.4 ADÉLIE POPULATION GENETIC ANALYSIS 114
5.3.5 PENGUIN INTRON PHYLOGENETICS 117
5.4 RESULTS **119**
5.4.1 ADÉLIE PENGUIN INTRON ANALYSES 119
5.4.2 PENGUIN PHYLOGENETIC ANALYSES 127
5.5 DISCUSSION **131**
5.5.1 ADÉLIE PENGUIN INTRON POPULATION GENETICS 132
5.5.2 PENGUIN PHYLOGENY 135
5.5.3 CONCLUSIONS 137
5.6 REFERENCES **138**

CHAPTER SIX **143**

RECOVERING ANCIENT NUCLEAR INTRONS OF ADÉLIE PENGUINS USING SECOND-GENERATION SEQUENCING

6.1 ABSTRACT **143**
6.2 INTRODUCTION **144**
6.3 METHODS **147**
6.3.1 DNA EXTRACTIONS FROM BONE 147
6.3.2 PCR VERIFICATION OF DNA EXTRACTIONS 148
6.3.3 DESIGNING INTERNAL PRIMERS FOR ANCIENT DNA WORK 150
6.3.4 DIRECT MULTIPLEX PCR FLX SEQUENCING METHODOLOGY 151
6.3.5 ANALYTICAL METHODS 155
6.4 RESULTS **158**
6.4.1 FLX OUTPUT, ASSEMBLY AND COVERAGE 158
6.4.2 ANALYSIS OF MODERN AND ANCIENT AK1I5 SEQUENCES 161
6.4.3 BLAST RESULTS OF CONTAMINANT READS 166
6.5 DISCUSSION **171**
6.5.1 DIRECT MULTIPLEX FLX SEQUENCING 171
6.5.2 ANCIENT ADÉLIE ADENYLATE KINASE INTRON 5 SEQUENCES 173
6.5.3 CONTAMINANT SEQUENCES 174
6.5.4 CONCLUSIONS 176
6.6 REFERENCES **178**

CHAPTER SEVEN **183**

CONCLUSIONS AND PERSPECTIVES

7.1 INTRODUCTION **183**
7.2 THESIS SUMMATION **184**
7.3 FUTURE PERSPECTIVES **187**
7.4 REFERENCES **189**

I APPENDIX ONE **191**

THE MOLECULAR ECOLOGY OF THE EXTINCT NEW ZEALAND HUIA

II APPENDIX TWO **203**

SUPPLEMENTARY MATERIAL CHAPTER FIVE: INDIVIDUAL PHYLOGENETIC TREES

III APPENDIX THREE **209**

SUPPLEMENTARY MATERIAL FOR CHAPTER SIX: EXTENDED METHODS

III.1 DIRECT SEQUENCING OF ANCIENT MYELIN PROTEOLIPID PROTEIN INTRON FOUR **210**

III.2 FLX TAGGING PROTOCOL AND LIBRARY QUANTIFICATION **212**

IV APPENDIX FOUR **221**

DRC AUTHOR CONTRIBUTION FORMS

List of Figures

Figure 1.1: Phylogenetic representation of modern penguins, taken from (BAKER *et al.* 2006). 3

Figure 1.2: Schematic drawing of different marker classes, their relative variability and adequacy for different research questions. 7

Figure 1.3: The Adélie penguin mitochondrial genome. 11

Figure 1.4: U2-type spliceosomal intron splicing mechanism. 15

Figure 1.5: Different theories for the evolution of introns, a) introns late theory (L) and b) introns early (E) and first (F) theories. 17

Figure 1.6: Structure of the human beta-fibrinogen gene, including both coding (exon) and noncoding (intron) regions. 21

Figure 2.1: The distribution of Ross Sea Adélie penguin colonies (1 – 15) from which samples were used in this study, along with a photograph (courtesy of K. Barton) used for counting number of breeding pairs from the colony at Cape Bird. 46

Figure 2.2: Schematic representation of the late Pleistocene migration of Adélie penguins into the Ross Sea of Antarctica, together with estimation times for the common ancestor of the current populations. 58

Figure 3.1: Extreme reverse sexual bill dimorphism in Huia. 67

Figure 3.2: Provenance of Huia samples used in this study. 69

Figure 3.3: DNA nucleotide variation in 199bp of the mitochondrial hypervariable region (HVRI) among 21 Huia. 73

Figure 3.4: Most probable mitochondrial diversity (θ) estimate from LAMARC. 74

Figure 3.5: Effect of generation time and mutation rate on the population size estimate 77

Figure 4.1: Distribution of Adélie penguin samples used in this study 87

Figure 4.2 Hypothetical length variant heterozygote electropherogram read showing mixed peaks as a result of a 1bp indel (marked in red). 92

Figure 4.3 2% agarose gel showing the six intron markers that produced high quality sequence well in Adélie penguins. 94

Figure 4.4 Adenylate kinase intron 5 length variant heterozygote. 95

Figure 4.5 Primer positions and amplicon sizes for Adenylate kinase intron 5 external and internal primers (not to scale). 96

Figure 4.6 Variable sites of introns sequenced in the Adélie penguin as compared to available sequences from other penguin species. 96

Figure 5.1 Phylogenetic representation of extant penguin genera and divergence times. 109

Figure 5.2 Haplotypes found for four introns in Adélie penguins. 120

Figure 5.3 Haplotypes found for Ak1i5 in Adélie penguins. 121

Figure 5.4 Mismatch distributions for five intron loci in Adélie penguins. 123

Figure 5.5 Haplotype networks constructed using sequence data from four introns of Adélie penguins. 126

Figure 5.6 Haplotype network constructed using sequence data from *AK1i5* of Adélie penguins. 127

Figure 5.7 Unrooted Bayesian modern penguin phylogenetic consensus tree for the concatenated four intron dataset (1926 bp incl. gaps). 129

Figure 5.8 Rooted Bayesian modern penguin phylogenetic consensus tree for the concatenated *AK1i5*/*ODC6* dataset (1216 bp incl. gaps). 130

Figure 6.1: Overview of direct multiplex sequencing of ancient Adélie penguin nuclear intron products. 146

Figure 6.2: The age and geographical distribution of subfossil Adélie penguin bones used in this study. 149

Figure 6.3: Multiplex primer positions and groupings. 153

Figure 6.4: FLX read distribution across tags and replicates. 159

Figure 6.5: Defining nucleotide changes for haplotypes of modern and ancient *AK1i5* sequences. 164

Figure 6.6: Temporal haplotype network for modern and ancient Adélie *AK1i5* sequences. 165

Figure 6.7: Distribution of BLAST hit results for FLX sequencing reads. 168

Figure 6.8: Distribution of significant BLAST hits for FLX sequencing reads. 169

Figure II.1 Unrooted Bayesian modern penguin phylogenetic consensus tree for locus *UCHL3*. 204

Figure II.2: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *AK1i5*. 205

Figure II.3: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *MPP4*. 206

Figure II.4: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *MPP4*.. 207

Figure II.5: UPGMA phylogenetic tree for the four concatenated intron-only dataset.208

Figure III.1: Schematic representation of positions of *MPP4* primers. 211

Figure III.2: *MPP4* direct sequencing results in four fragments from ancient Adélie subfossil bone samples. 212

Figure III.3: Example of a tagged target sequence prior to the adapter fill-in step. 213

Figure III.4: 3% agarose gel showing the result of a ligation test of one adapter. 215

Figure III.5: Consensus ancient Adélie *AK1i5* sequences obtained from FLX sequencing. 220

List of Tables

Table 2.1 Summary statistics for *HVRI* in fifteen Adélie penguin colonies 50

Table 2.2: AMOVA results. 51

Table 2.3: Pairwise ϕ_{st} results. 52

Table 2.4: Estimates of population sizes from population counts together with genetic diversity in colonies of Adélie penguins. 53

Table 3.1: Huia samples used in this study. 70

Table 3.2: Long-term population size estimates of Huia based on mitochondrial hypervariable region diversity. Reproduced with permission from Lambert *et al* (2009). 75

Table 4.1 Details of primer pairs tested in Adélie penguin samples from the Ross Sea, Antarctica. 84

Table 4.2 PCR conditions and annealing temperature ranges tested during optimization runs for 26 primer pairs in modern samples. 88

Table 4.3 PCR program details for modern samples. 88

Table 4.4 Results of PCR condition testing intron screen in modern Adélie penguins. 90

Table 5.1 Adélie penguin sample provenance, together with mtDNA lineage and intron haplotypes 111

Table 5.2 Penguin species sequenced for four introns 112

Table 5.3 Summary statistics for five intron loci in Adélie penguins. 122

Table 5.4 Results of the mismatch analysis for five intron loci in Adélie penguins and θ estimates. 125

Table 5.5 Effective population size estimates for five intron loci in Adélie penguins 125

Table 6.1: Multiplex primer groupings and information for the three different introns that were the subject of this study. 152

Table 6.2 Coverage and distribution of FLX reads for adenylate kinase intron 5 160

Table 6.3: Modeltest results for modern and ancient AK1i5 Adélie datasets 163

Table 6.4: Summary statistics and neutrality tests for modern and ancient AK1i5 Adélie datasets 163

Table 6.5: Ancient and Modern Samples included in Analyses 164

Table III.1 Internal primers for *MPP4* 210

Table III.2: Barcoded Adapters used for DMPS FLX Titanium Sequencing 216

1 Chapter One

INTRODUCTION

1.1 Thesis Structure

This thesis covers a series of manuscripts that are published, submitted or in preparation for publication. As a result some of the information provided in the introduction chapter will be repeated in subsequent chapter introductions. Some methodological details will, therefore, also be repeated. In this thesis, current knowledge of Adélie penguin population history is extended using mitochondrial and nuclear non-coding intron markers.

Chapter Two investigates the relationship between mitochondrial genetic diversity and effective population size, using mitochondrial hypervariable 1 (*HVR1*) sequences from a number of Adélie penguin colonies. This chapter has been submitted for. **Chapter Three** uses *HVR1* sequences of the extinct New Zealand Huia to infer the species' population size prior to their decline. The work carried out in Chapter Three was included in the publication 'The Molecular Ecology of the Extinct New Zealand Huia', (2009) PLoS ONE **4**(11), pg e8019, and is presented in Appendix One.

The focus of the chapters shifts from utilizing mitochondrial DNA markers for the elucidation of population history to developing and utilizing nuclear intron markers in

Adélie penguins. In **Chapter Four**, candidate nuclear intron loci for Adélie penguin phylogenetics and population genetics are recovered and described. A subset of these intron loci has been applied to Adélie penguin population genetics and broader penguin phylogenetics in **Chapter Five**. In **Chapter Six** second-generation sequencing technology is applied to recover nuclear intron sequence from ancient Adélie penguins.

Chapter Seven presents a short discussion and summary of the conclusions of the thesis as a whole, including potential directions for future work.

1.2 Adélie penguins

Adélie penguins (Sphenisciformes: Spheniscidae: *Pygoscelis adeliae*) are an excellent model species for studying molecular evolution and ecology, as our understanding of this species is extensive in both traditional ecological areas and in more recent molecular studies. Given the depth of the research on Adélie penguin ecology and behaviour (AINLEY 2002; WOEHLER 1993), only areas relevant to this work will be outlined.

Adélie penguins belong to the genus *Pygoscelis*, along with the Chinstrap (*P. antarcticus*) and Gentoo penguins (*P. papua*). All three species breed and reside in Antarctica and the Antarctic Convergence, and are phylogenetically basal to nearly all other penguins, sharing a common ancestor approximately 38 mya, while Adélie penguins split from this lineage approximately 19 mya (Fig. 1.1) (BAKER *et al.* 2006); but see (CLARKE *et al.* 2007).

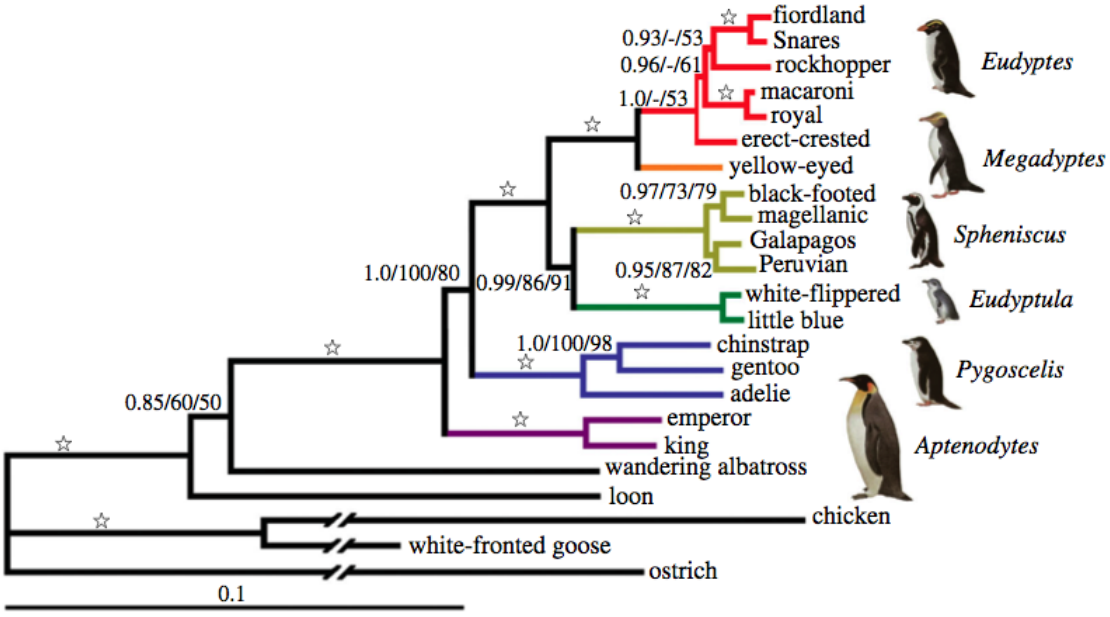


Figure 1.1: Phylogenetic representation of modern penguins, taken from (BAKER *et al.* 2006).

2802 bp of *RAG-1* and 2889 bp of mitochondrial 12S and 16S rDNA, *cytb* and *COI* were used to reconstruct this phylogeny, and Bayesian/maximum likelihood/maximum parsimony support for each node is shown, with a white star=0.1/100/100.

At present there are approximately 2.47 million breeding pairs of Adélie penguins in Antarctica. Nesting in this species occurs during the summer months and an estimated 177 colonies are known around the Antarctic (WOEHLER 1993). Adélie penguins have inhabited the Ross Sea coast continuously for at least the past 7000 years (BARONI and OROMBELLI 1994) with colony sizes varying greatly. Adélie penguins typically exhibit a high degree of natal philopatry (AINLEY 2002), though during times of great environmental instability increased movement between nearby colonies has been detected (DUGGER *et al*. 2010; SHEPHERD *et al*. 2005). The number of breeding pairs present in the various colonies has been accurately documented since the 1960s, particularly in the Ross Sea (MCNEILL *et al*. 2011; WILSON *et al*. 2001).

Adélie penguins occupy a very narrow ecological niche, breeding on ice-free areas of the Antarctic coastline, and construct their nests from pebbles to keep eggs and chicks dry. They are adapted to a narrow optimum of sea-ice presence and as a result Adélie penguins are highly susceptible to changes in climate (AINLEY 2002) with past ice ages and warming events resulting in waves of colony abandonment and

recolonization. Radiocarbon-dating of Adélie penguin remains from abandoned colonies of the Victoria Land coast indicate a Late Pleistocene occupation from 45,000 to around 25,000 years ago, and a later Holocene recolonization when coastline emerged after the Last Glacial Maximum approximately 20-26 kya (BARONI and OROMBELLI 1994). As a result of fluctuations in sea-ice coverage during the Holocene period, Adélie penguin populations have expanded and contracted repeatedly (HALL *et al*. 2006), with a major expansion occurring along the Victoria Land coast in the last 1,000 years.

1.2.1 Molecular ecology and evolution in Adélie penguins

Nearly all studies investigating the ecology and evolution of Adelie penguins have been based on molecular studies of the Adelie mitochondrial genome. The mitochondrial control region of Adélie penguins was characterized and found to contain a repeat complex which made the region much longer than most avian control regions (RITCHIE and LAMBERT 2000). HVRI sequences obtained from a large number of modern and ancient Adélie penguins revealed the presence of two distinct mitochondrial lineages, referred to as *A* and *RS*. The *RS* lineage is located mainly in the Ross Sea area, while the *A* lineage is present at all other sampled locations around Antarctica. The existence of these two lineages suggests that Adélie penguins expanded out of two ancestral refugia during the last glacial maximum (RITCHIE *et al*. 2004).

For the estimation of evolutionary rates, Adélie penguins have proven important. Utilizing an ancient DNA approach, in which molecular changes are measured from serially preserved samples, the first estimate of this kind for the mitochondrial hypervariable control region was obtained (LAMBERT *et al*. 2002). Not every species is amenable to this technique, as there should be large numbers of living individuals and large amounts of well preserved, dated, ancient tissue samples. This is the case for Adélie penguins. Penguin remains from egg and chick predation lie on the ground at their colonies and are eventually buried by new nests. Penguin guano accumulates at the bases of the pebbly nests, creating ornithogenic soils in layers from successive

occupations beneath extant and abandoned colonies. The thicker these layers are, the older a colony is (BARONI and OROMBELLI 1994). Stratigraphic analysis of these soils provides a record of Adélie penguin colony occupation, and penguin remains can be radiocarbon-dated. These remains are well-preserved thanks to the extremely cold and dry conditions in the Antarctic.

Using numerous serially preserved subfossil remains of Adélie penguins, evolutionary rates have been estimated for the *HVRI* region (LAMBERT *et al*. 2002; MILLAR *et al*. 2008a; RITCHIE *et al*. 2004). The rate obtained for this region, from 162 subfossil bones spanning 37,000 years, was 0.86 substitutions per site per million years (0.53-1.17 95% confidence interval), higher than many previous estimates. From complete mitochondrial genomes of Adélie penguins up to 44,000 years old, evolutionary rates for the mitochondrial genome as a whole, as well as for individual regions of the genome, have been estimated. (SUBRAMANIAN *et al*. 2009). These rates are also generally high. Whether these high evolutionary rates were linked to high mutation rates was investigated by extensive sequencing of the *HVRI* region for penguin families at Cape Bird. Germline heteroplasmic mutations, detected in mothers and chicks, were used to generate a mutation rate of 0.55 mutations per site per million years (0.29 – 0.88 95% confidence interval) (MILLAR *et al*. 2008a). The two rates obtained for the HVRI region (from the serially sampled and pedigree approaches) were not statistically different (MILLAR *et al*. 2008a). This amount of knowledge available on the Adélie penguin makes it an unparalleled natural model species in which to test aspects of population genetics and molecular evolution.

Ancient nuclear DNA studies, though still technologically challenging, should be feasible in Adélie penguins thanks to the excellent preservation conditions of their subfossil remains. Nine nuclear microsatellite loci (allele lengths 78-132bp) were genotyped for an ancient population of Adélie penguins from Inexpressible Island aged approximately 6000 years B.P. (SHEPHERD *et al*. 2005). The ancient and modern populations showed that a shift in allele frequencies had occurred, and that allele sizes had increased, demonstrating microevolutionary change (SHEPHERD *et al*. 2005).

1.3 Molecular Markers

The application of molecular markers as tools in molecular ecology and evolution can only be as good as our understanding of their biology. If applying molecular markers for population genetics and phylogenetics, assumptions regarding the mutational processes and evolution governing these are generally made. First and foremost, nucleotide substitutions are largely assumed to be neutral to selection. Genes and genetic regions not under selective pressure are deemed neutral, and the loss or maintenance of mutations in these is thought to be governed by drift and purifying selection (KIMURA 1983; OHTA 1992). The identity of the marker being used and mechanisms of its molecular evolution should be known, and substitutions should generally fit a molecular clock with a predictable rate (FRIESEN 2000).

As technology has improved, molecular markers that once monopolized studies have been steadily replaced by others able to provide another dimension to a study question. During the mid-1960s protein electrophoretic approaches were common, but since the 1970s methods began introducing DNA analysis (particularly through the use of restriction enzymes). With the advent of polymerase chain reaction (PCR) amplification in the 1980s and DNA sequencing technologies, it became much easier and cheaper to perform more direct DNA studies. These techniques have been widely applied to the study of population genetics and phylogenetics of living organisms through the use of molecular markers (AVISE 1994). The mitochondrial genome has supplied the most frequently used markers for the elucidation of species and population histories for decades, and is still widely popular. The characteristics of the mitochondrial genome are detailed in the next section and include reasons for the popularity of this molecule. A non-trivial reason mtDNA has been so often used, aside from those detailed below, is the availability of nearly universal primers for PCR amplification of mitochondrial markers, excluding the control region (eg across all vertebrates) (SORENSON *et al*. 1999), and the wide scope of research questions that can be addressed using it (Fig. 1.2).

The application of nuclear molecular markers, however, often requires costly marker development strategies for non-model organisms. Recent advances in genomic

biology are increasing available resources for marker development, and new sequencing technologies are contributing to these resources at a rapid rate. This implies that applying a suite of nuclear markers either independently or in conjunction with mitochondrial markers is now much more feasible. Among nuclear markers, depending on the scope of the question being studied, different genomic regions are favored (Fig. 1.2). Overall the most commonly used nuclear markers include microsatellite repeat regions, single-nucleotide polymorphisms (SNPs), nuclear non-coding regions (introns, for the most part, described as 'EPICs' below), nuclear protein coding loci (NPCL), and anonymous nuclear markers (ANMs).

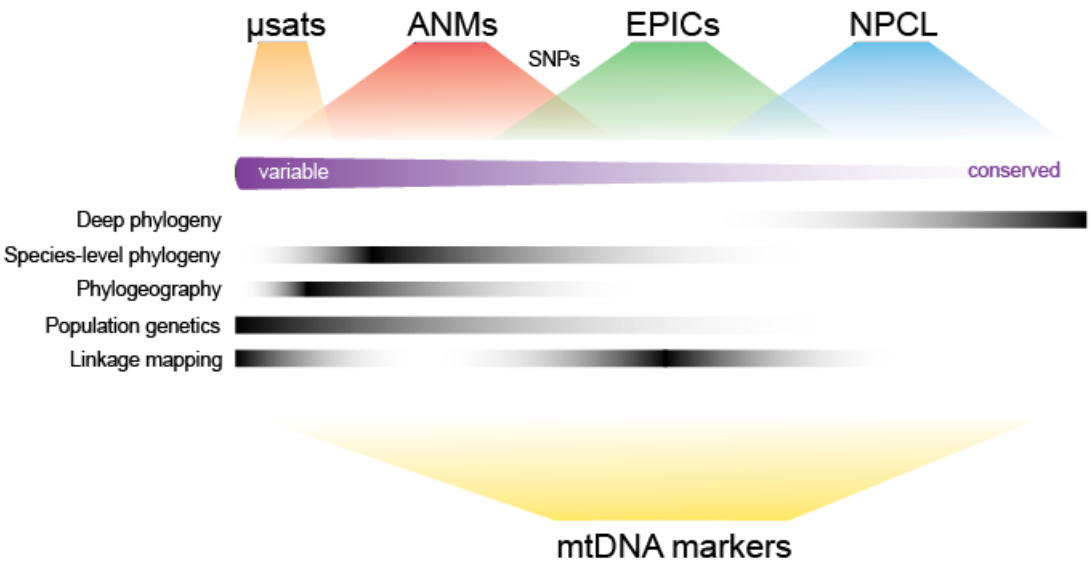


Figure 1.2: Schematic drawing of different marker classes, their relative variability and adequacy for different research questions.

Nuclear markers are indicated above. Different mitochondrial genes and regions are adequate to different research questions; however, for simplicity these have been combined into a joint class. Only nuclear markers are used for linkage mapping. Markers shown: microsatellites (μsats), anonymous nuclear markers (ANMs), single nucleotide polymorphisms (SNPs), exon-primed, intron-crossing markers (EPICs), nuclear protein coding loci (NCPL), and mitochondrial DNA markers (mtDNA markers). Figure adapted from (THOMSON *et al*. 2010).

For deep phylogenetics, which requires highly conserved sequences, NPCLs are often chosen. They are functionally constrained and so accumulate changes much more slowly than non-coding regions, making them easy to align over large phylogenetic distances (TOWNSEND *et al*. 2008). Intron markers are non-coding and therefore presumably not under the functional constraint that protein-coding regions are under. As a result they are more variable and thus suited to more shallow phylogenetics or population genetics. Primers to amplify introns are generally designed within the

flanking exons, conserved enough that the primer sites should not accumulate substitutions across a shallow phylogenetic span (BACKSTRÖM *et al*. 2008). These markers are described in more depth in later sections of this introduction. ANMs are designed from random extracts of the nuclear genome (KARL and AVISE 1993). Most ANMs are found in non-coding genomic regions because most of the genome is non-coding. Due to being located in non-coding regions these markers often evince a high variability that makes them informative for shallow divergence level analyses (JENNINGS and EDWARDS 2005). However their very anonymity can be a problem, particularly due to unknown paralogs and copy-number (THOMSON *et al*. 2010). Microsatellites are short DNA sequence regions in which a one to six base pair motif is tandemly repeated. They are highly variable and multi-allelic and are almost exclusively used in population genetics and parentage analyses, though their random distribution has also helped facilitate the construction of genetic maps (SCHLÖTTERER 2000). SNP markers are simply single base changes in a DNA sequence, usually with only two alternative nucleotides due to the low frequency of single nucleotide substitutions at the origin of SNPs and a transition bias. They can be obtained from direct sequencing of genomic PCR products or by genotyping, and are located in both coding and non-coding regions. The lower mutation rates of nuclear loci when compared to mitochondrial DNA implies that gene trees for individual loci are not always well resolved; however, incompletely resolved gene trees, if summed over multiple loci, can still provide strong signals for phylogenetics (HARE *et al*. 2002).

1.4 Mitochondrial DNA

Mitochondrial DNA (mtDNA) has been by far the most widely used source of molecular markers in molecular ecology and evolution (AVISE 1994). This is in large part due to several defining characteristics of mtDNA. The biology of this molecule is quite different from that of the nuclear genome. It is small, but present in high copy numbers within the cell, is double-stranded, replicates independently from the nuclear genome (RANDI 2000) and usually does not recombine (but see (BARR *et al*. 2005; TSAOUSIS *et al*. 2005)). It is inherited clonally and usually through the maternal line, though episodes of paternal leakage have been observed (BARR *et al*. 2005).

Mutation rates within the molecule are also generally high and very often variability within and between species and populations is high as a result (BROWN *et al*. 1979).

The mitochondrial genome's two strands are referred to as the Heavy (H) and Light (L) strands as a result of their differing base compositions (KASAMATSU and VINOGRAD 1974). In general in vertebrates, it is composed of thirteen protein-coding genes, two ribosomal RNA (12SrRNA and 16SrRNA), and twenty-two transfer RNAs (tRNA). The heavy strand contains the genes encoding two rRNAs, fourteen tRNAs and twelve of the protein-coding genes, while the light strand encodes eight tRNAs and one protein-coding gene (PEREIRA 2000). The cytochrome *b* gene is the most widely used gene for phylogenetic work, as it is thought to be variable enough for some population work yet still conserved enough for phylogenetics. The mitochondrial genome also contains a highly variable control region (CR) that functions mainly as a regulatory region. This region in particular has been highly useful for evolutionary studies and is composed of two hyper variable regions (*HVR-I* and *HVR-II*). The control region is frequently used in population studies thanks to its high variability, while certain protein coding genes such as the cytochrome b gene are more often used for phylogenetic analysis above the species level.

Mitochondrial gene order is usually conserved within different phyla and orders but varies between them (RANDI 2000). The order of mitochondrial genes of placental mammals, turtles, fish, some lizards and *Xenopus* are among the most conserved, that is, there is little variation between taxonomic groups. Less conserved gene orders are found among birds, other lizard species, crocodilians, marsupial mammals, snakes, tuatara, lamprey, some amphibians and one fish species. The first complete avian mitochondrial genome to be published was from chicken (DESJARDINS and MORAIS 1990), and showed highly conserved features when compared to other vertebrate mtDNAs, though the gene order can be quite different. New gene rearrangements appear due to tandem duplication and deletion events associated with tRNA sequences (PEREIRA 2000). Length variation of the mitochondrial genome is in large part due to the presence of large repeat regions within the control region, as is the case in Adélie penguins (RITCHIE and LAMBERT 2000) or entire duplications of the control region, as occurs in Tuatara (*Sphenodon* sp.) (PEREIRA 2000). The chicken, tuatara

and Adélie penguin mitochondrial genomes are shown in Figure 1.3 to illustrate varying gene order.

Knowledge of mtDNA organization is essential for polymerase chain reaction (PCR) primer design for phylogenetics and population genetics. Mitochondrial DNA markers have been used to estimate variability within populations, construct phylogenies, detect population bottlenecks and expansions, determine population structure, estimate effective population sizes, estimate gene flow, introgression and hybridization, to name the most common applications (AVISE 1994; FRANKHAM *et al*. 2005; HARRISON 1989; RANDI 2000). As a result of uni-parental inheritance in mtDNA, mtDNA has one quarter the effective population size of a nuclear gene, and the maternal lineage of populations can be traced through time without the added complications that arise from paternal inheritance or recombination (PAKENDORF and STONEKING 2005). The elevated mutation rate of mtDNA, generally one order of magnitude higher than the nuclear genome (BROWN *et al*. 1979; WILSON *et al*. 1985), is another reason this marker has proven popular in molecular studies in animals. Mitochondrial DNA markers have also been frequently used in ancient DNA studies. The high copy number of mtDNA relative to nuclear DNA makes recovery of mitochondrial markers easier in samples that are often too degraded to amplify nuclear markers (RITCHIE *et al*. 2004).

1.4.1 The relationship between mitochondrial DNA diversity and population size

Under the neutral theory of molecular evolution (KIMURA 1983), levels of neutral genetic variation are expected to correlate with population size within species as well as across different species (FRANKHAM 1996). More specifically, heterozygosity at neutral loci increases as the product of population size and the mutation rate increases (KIMURA 1983; OHTA 2003). Hence, measuring the amount of genetic variation present in a population has uses for conservation biology and the elucidation of population history.

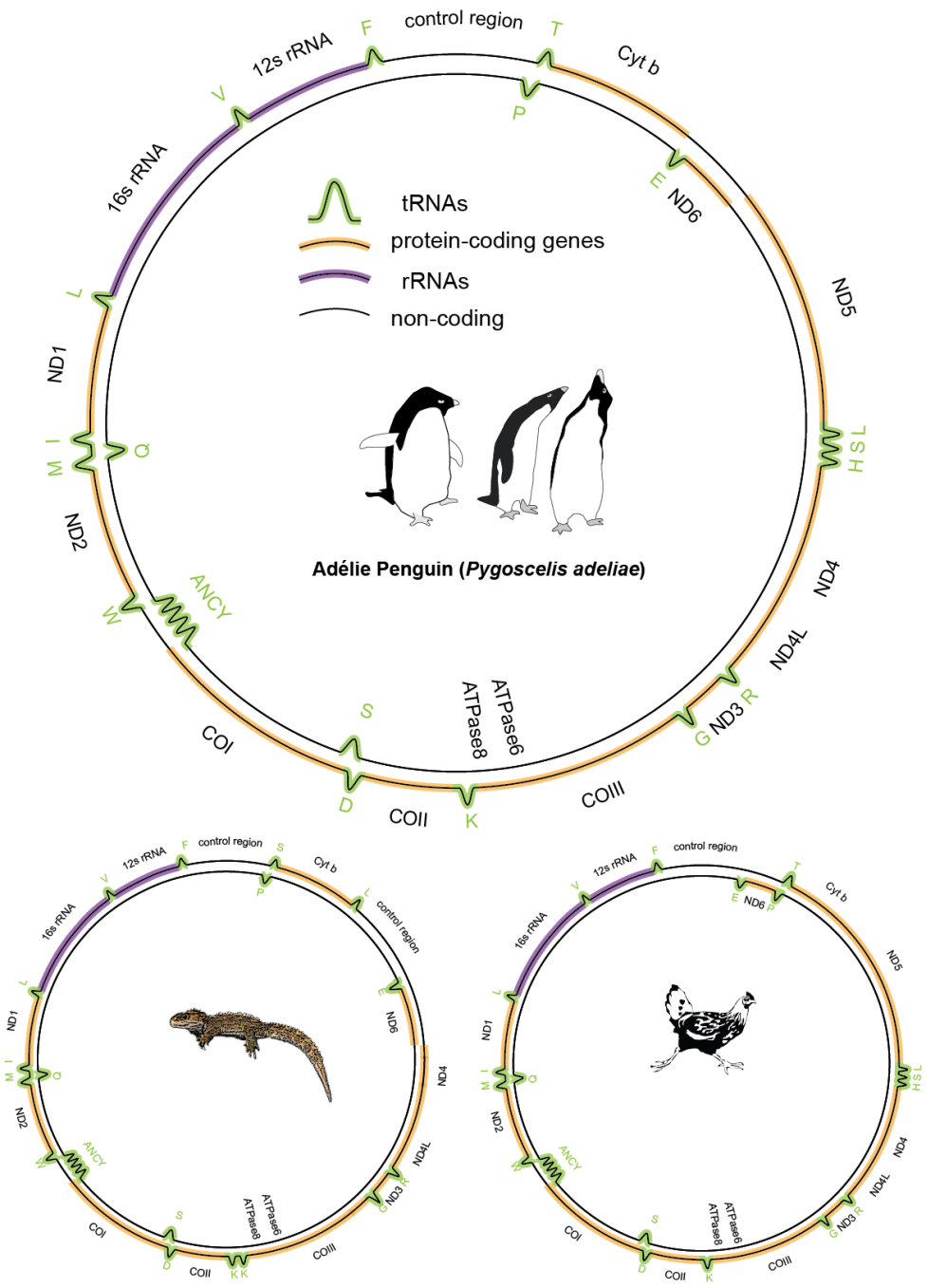


Figure 1.3: The Adélie penguin mitochondrial genome.

The tuatara (*Sphenodon* sp.) and chicken (*Gallus* sp.) mitochondrial genomes are shown below to illustrate gene order differences (DESJARDINS and MORAIS 1990; REST *et al*. 2003; SUBRAMANIAN *et al*. 2009).

Previous empirical studies have shown that a positive linear relationship exists between the logarithm of population size and allozymic diversity (SOULÉ 1976), non-coding nuclear genetic diversity (FRANKHAM 1996) and mitochondrial diversity (AVISE 1992). As a result of these studies it has been assumed that mitochondrial DNA diversity serves as a good proxy for population size, and can be used to estimate effective population size (N_e), which is the size of an "ideal population" with the same rate of genetic change as the population being studied (FRANKHAM 1996; WRIGHT 1931). An ideal population is characterized by having randomly mating individuals, equal sex ratios, discrete non-overlapping generations and random variation in reproductive success. The two most commonly estimated measures of effective population size are variance (N_{eV}) and inbreeding (N_{eI}) effective population size. The first describes the rate of allele-frequency change and represents the current effective population size, while the second is measured as the probability that two alleles can be traced back to the same gene in a common ancestor and represents the historical effective population size. In populations of fluctuating size current and historical effective population sizes will be different; stable populations should in theory have a constant effective size through time. N_e is a critical parameter in evolutionary biology but collecting enough demographic data to estimate it is often difficult. As a result, genetic methods of estimating N_e have enjoyed much popularity in evolutionary and conservation biology, despite concerns regarding precision and bias (WAPLES 2002; WAPLES 2005). N_{eV} is the most commonly estimated parameter, generally calculated by temporal methods, which measure allele frequency changes over a few generations, most frequently using microsatellite DNA (LEBERG 2005; SCHWARTZ *et al*. 1999).

Long-term or inbreeding effective population size (N_{eI}) can be estimated through coalescent analysis of contemporary genetic variation (WANG 2005). This is especially useful when a species does not have detailed historical records and a recent reduction or expansion in population size is suspected. In these cases a comparison with historical population levels is of great utility (ROMAN and PALUMBI 2003; YEUNG *et al*. 2006). Mitochondrial DNA sequence data are most frequently used to estimate N_{eI}, allowing one to infer historical population changes (CRANDALL *et al*.

1999; GEMMELL *et al*. 2004; ROMAN and PALUMBI 2003; YEUNG *et al*. 2006). Mitochondrial DNA's assumed maternal inheritance and absence of recombination, as well as its lower effective population size compared with nuclear DNA, means it can be used to trace recent evolutionary history, such as founder events, bottlenecks, and introductions, etc (HARRISON 1989). Different methods exist to estimate genetic diversity (θ) from either mitochondrial or nuclear markers, depending on the assumptions regarding gene flow and population size fluctuations primarily (BEERLI and FELSENSTEIN 1999; BEERLI and FELSENSTEIN 2001; CRANDALL *et al*. 1999; DRUMMOND *et al*. 2005; EMERSON *et al*. 2001).

1.4.2 Criticism of mtDNA in population genetics and phylogenetics

Reliance on mtDNA for phylogenetics and population genetics has been criticised in the literature, with some studies suggesting that its use should actually be discontinued (BALLARD and WHITLOCK 2004; BAZIN *et al*. 2006). Assumptions of the molecule's characteristics have been violated in several cases. For example, evidence for recombination has been detected (BALLARD and WHITLOCK 2004; EYRE-WALKER *et al*. 1999), as well as adaptive evolution (BAZIN *et al*. 2006; HURST and JIGGINS 2005). Though maternal inheritance is widely accepted as the standard model of mtDNA transmission, a number of studies have shown that paternal inheritance does occur in some species (WHITE *et al*. 2008). However, these are exceptions and for the most part mitochondrial DNA does conform to a uniparental mode of inheritance.

Whether mitochondrial DNA can be considered a strictly neutral marker has been questioned for over a decade (RAND and KANN 1996). A number of recent studies have suggested that mitochondrial genetic diversity does not in fact reflect the size and history of a population and so should not be applied to population genetics studies (BALLARD and WHITLOCK 2004; BAZIN *et al*. 2006; HAHN 2008). The high evolutionary rate of the molecule lends support to its neutrality (BROWN *et al*. 1982; BROWN *et al*. 1979), though within species there are often an excess of rare haplotypes that carry mildly deleterious mutations (FRY 1999). Selection on any part of the mitochondrial genome could potentially influence polymorphism in the whole

molecule because of the lack of recombination (BALLARD 1995). Bazin *et al*.'s (2006) study concluded that natural selection could be acting on mtDNA through genetic draft. This is defined as the recurrent fixation of advantageous mutations which then leads to a frequent loss of variability at the locus under selection and any linked loci (GILLESPIE 2000). It has been suggested that the effect of genetic draft on diversity increases with population size (GILLESPIE 2001). Bazin *et al*. (2006) arrived at this conclusion by comparing average diversity between many species of invertebrates and vertebrates, small and large organisms, and marine and terrestrial organisms. Nuclear DNA and allozyme diversity were also analyzed and found to be greater in organisms with higher average population size, while mtDNA diversity failed to reflect this difference (BAZIN *et al*. 2006). Other studies have suggested that for vertebrates in general, or animals with generally small population sizes (eg humans), mtDNA is still an adequate marker (ATKINSON *et al*. 2008; EYRE-WALKER 2006; HUGHES and HUGHES 2007; MULLIGAN *et al*. 2006). Positive, significant relationships between genetic diversity and catch size of numerous fish species, which have larger population sizes than humans and than many other species of vertebrate, were also found (MCCUSKER and BENTZEN 2010). It does appear, from these studies, that genetic drift and not draft is the prevailing force in mitochondrial evolution, at least within vertebrates. Further studies examining this question within vertebrate species (among populations) would be valuable in contributing to this interesting question.

1.5 Nuclear Introns

As molecular techniques in ecology and phylogeny studies have improved, studies embracing multi-locus approaches that incorporate data from nuclear sequences have multiplied, thus relinquishing complete dependence on mitochondrial DNA. This is important, as incongruencies have been found between individual gene trees, suggesting that inferring a species tree from only one gene tree may not be the wisest choice. Ideally, a number of unlinked loci should be used so the resultant gene trees provide independent estimates of the species tree (PAMILO and NEI 1988). Also, a range of nuclear loci, for example introns, provide phylogenetic resolution at

intermediate depths of divergence for which homoplasy becomes a problem for the mitochondrial control region and at which nuclear protein coding genes have not yet accumulated many differences (SHAPIRO and DUMBACHER 2001).

Increasingly nuclear non-coding introns are being used independently or in conjunction with mitochondrial DNA in phylogenetics and some population genetics studies. As with most molecular markers used in ecology, introns are useful primarily because, as non-coding regions of the genome, they are assumed to be essentially neutral to selection (discussed further below) (FRIESEN 2000).

1.5.1 Four groups of introns

Introns are untranslated or non-coding gene regions of genomic DNA, spliced out as mature messenger RNA (mRNA) molecules are formed (GILBERT 1978). They can be classified into four groups, depending on their splicing mechanism. **Groups I** and **II** are self-splicing and are both found in bacterial genomes, while group I introns are also found in ribosomal RNAs (rRNA) of protists and fungal nuclei (CREER 2007). The RNA structures of introns belonging to these two groups facilitate their self-splicing activity, and they contain internal open reading frames (ORFs) which facilitate intron removal and propagation via reverse transcription (ROY and GILBERT 2006).

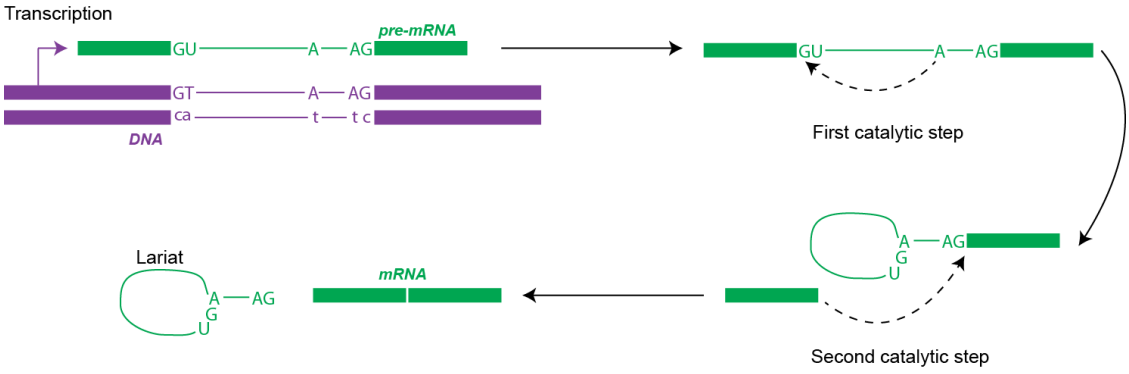


Figure 1.4: U2-type spliceosomal intron splicing mechanism.

In eukaryotes the most common insertions in pre-mRNA genes are **spliceosomal** introns. They generally lack ORFs and have nearly random sequences. Spliceosomal introns require five RNAs and hundreds of proteins (the spliceosome) to create a complex and aid the excision of introns from maturing mRNA molecules (NILSEN 2003). Two types of spliceosome are known in eukaryotic introns. The first is the most common, the U2-type that splices GT-AG introns (the intron begins with 5'GT and ends with 3'AG), which have a pyrimidine rich region preceding the 3' splice site (Fig. 1.4). The U12-type splices the much more rare AT-AC introns (BELSHAW and BENSASSON 2006). Generally spliceosomal introns contain a 'branch site' with the sequence YUNAY (Y is any pyrimidine and N is any base) found 18-40 bases from the 3' end (FRIESEN 2000). The self-splicing mechanism of Group II introns is similar to that of spliceosomal introns, in that the 2'-OH of an adenine residue within the introns acts as a nucleophile. This shared mechanism suggests that components of the spliceosome are derived from a group II intron, indicating a possible evolutionary relationship between the two (HAUGEN *et al*. 2005).

Another type, transfer RNA (**tRNA**) introns can be found in eukaryotic nuclei and in Archaea. They are enzymatically removed by a cut and join mechanism requiring ATP and an endonuclease, a completely different pathway from spliceosomal introns (HAUGEN *et al*. 2005).

1.5.2 Spliceosomal Intron Evolution

Intron sizes range from ~ 50 bp (base pairs) to tens of thousands of bp (FRIESEN 2000). Their numbers vary hugely between eukaryotic species, from under 100 introns per genome to hundreds of thousands in vertebrates and plants (ROY and GILBERT 2006). These differences imply huge intron gains or losses over evolutionary time, which still constitute a puzzle related to the evolution of genomic complexity (ROY and GILBERT 2006). Introns incur a cost during replication and transcription simply by their presence, with the cost proportional to the intron's size (DURET 2001). Are introns selfish elements that invaded the eukaryotic genome, neutral, or do they confer a selective advantage to the evolution of the genome?

Various theories support the latter hypothesis. Different theories of genome evolution suggest a prominent role for introns, in which the proliferation of introns may have facilitated the construction of the first full-length genes through recombination (GILBERT 1978; GILBERT 1987; ROY 2003). Recombination of introns would allow different exons to increase the rate at which they could be recombined as independent structures. New genes could be created by duplication and migration of exons along a gene during evolution (LIMA-DE-FARIA 1995). Alternative splicing of genes, possible due to the presence of spliceosomal introns, also contributes to protein evolution, as a single eukaryotic gene can encode many different proteins, without involving a change at the DNA level (LIMA-DE-FARIA 1995). There are several plausible hypotheses and models highlight the advantages that introns may confer to the evolution of the genome, however it is not certain whether these advantages are cause or consequence of the spread of introns (DURET 2001; ROY and GILBERT 2006). The mechanisms by which introns are gained and lost are still not completely understood and continues to be the subject of debate in the literature (RAGG 2011; ROY and IRIMIA 2009).

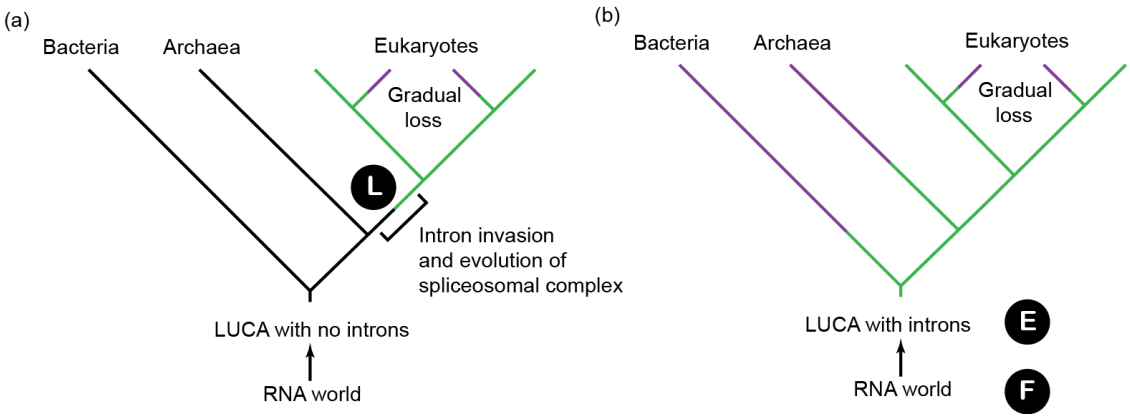


Figure 1.5: Different theories for the evolution of introns, a) introns late theory (L) and b) introns early (E) and first (F) theories.

Green branches indicate lineages containing introns, black branches show pre-intron stages and purple branches signify secondary loss of introns. Image redrawn from **(JEFFARES *et al.* 2006)**.

Another hotly debated topic regarding intron evolution is the timing and manner of spliceosomal intron emergence. All essential components of two classes of spliceosomes are present in early diverging eukaryotes, implying that functional intron excision mechanisms were present in the last common eukaryotic ancestor

(COLLINS and PENNY 2005) Traditionally, two theories exist to explain the evolution of spliceosomal introns. The first hypothesis, commonly called **introns-early** (Fig. 1.5b), postulates that introns arose in early life forms where they provided recombination sites between exons, and were subsequently lost from prokaryote genomes (GILBERT 1978). In the second hypothesis, **introns-late** (Fig. 1.5a), introns were thought to have been inserted randomly into continuous protein-coding regions in eukaryotic primordial genes (CAVALIER-SMITH 1985). A third hypothesis, **introns first** (Fig. 1.5b), is similar to the introns early theory but suggests spliceosomal introns are remnants of the RNA world (JEFFARES *et al*. 1998; POOLE *et al*. 1998). Debate continues over these hypotheses (BELSHAW and BENSASSON 2006; JEFFARES *et al*. 2006; KOONIN 2006; SOUZA 2003).

Aside from the role intron presence may play in genomic evolution, the evolution of introns themselves is an interesting subject. Intron size varies within and among genomes, and is influenced by various factors. Intron length may be a selected trait, due to its correlation to recombination; longer introns were found to be present in regions of low recombination, where the efficacy of natural selection is assumed to be lower (DURET 2001; MARAIS *et al*. 2005). However, deletions within introns have been found to be more frequent than insertions so a trend towards the collapse of intron length might be expected but is not the case. It may be therefore that long introns are advantageous, by increasing recombination rates in areas where rates are low (DURET 2001). The correlation between intron size and recombination rate could potentially also be explained by transposable element insertion events and not selective pressures (DURET 2001), though purifying selection of exons could be associated to increased intron size (MARAIS *et al*. 2005).

The evolution of intron size through insertion-deletion (indel) mutations is not yet fully understood. Further understanding of the underlying mutation rates and substitution properties is needed to resolve this issue (JOHNSON 2004). Intron size has been shown to be quite stable over evolutionary timescales in some vertebrates (WALTARI and EDWARDS 2002), while a study in pigeons and doves reported a deletion bias across the phylogeny (JOHNSON 2004). In this study deletions outnumbered insertions by 6 to 1, which the authors concluded was caused by a

selection bias. Insertions are hypothesized to be more deleterious than deletions because of reduced transcription and splicing efficiency (LYNCH 2002).

Within introns, mutations are expected to accumulate randomly and be free from selective pressure due to their non-coding nature (FRIESEN 2000). Conserved regions within introns, however, may affect eukaryotic gene expression through initial transcription, editing, polyadenylation and nuclear export of pre-mRNA, translation and decay of mRNA, and in alternative splicing of exons (CREER 2007; GAZAVE *et al*. 2007). Thus, some introns possess a structure involving conserved and RNA secondary structure regions as well as mutational hotspots, evolving under different evolutionary constraints. First introns of genes and sequences near intron-exon junctions in humans are conserved and are most likely under purifying selection, due to the presence of transcription factor binding sites in the first and splice-recognition sites in the second (CHAMARY and HURST 2004). The evidence for the assumed effective neutrality of introns stems mainly from observations of high substitution rates, similar to that of synonymous sites in exons, most of which are also assumed to evolve neutrally or nearly-neutrally (CHAMARY and HURST 2004; HUGHES and YEAGER 1997), though synonymous sites are now thought to be under some form of purifying selection as well (CHAMARY *et al*. 2006). Though some introns or conserved regions within introns exhibit functional significance, empirical studies suggest most nucleotides within introns are free to vary and their evolution generally follows the neutral or nearly-neutral model. This is especially likely to be the case within species with smaller populations for which genetic drift may have a greater effect than weak purifying selection of mildly deleterious mutations.

This is an important point, as in order for a molecular marker to be amenable to population genetic and phylogenetic studies, substitutions occurring within it ought to be selectively neutral. Empirical studies using introns so far suggest that this is the case, for the most part. Studies have found that most substitutions are distributed randomly within introns; transitions outnumber transversions; substitutions form a 'star' pattern with a few hubs; and Tajima's D statistic does not differ significantly from 0, implying no strong selective force is acting (CONGDON *et al*. 2000; FRIESEN *et al*. 1999; FRIESEN *et al*. 1997; PRYCHITKO and MOORE 1997). As a result of the

evidence for functional components within introns one should not assume neutrality of introns *a priori* and statistical tests such as those mentioned above should be applied as a matter of course to identify potential selective constraints. This however does not preclude introns providing valuable markers for population genetics and phylogenetics.

1.5.3 Using introns in phylogenetics and population genetics

Intron sequences have been used for phylogenetic resolution of relationships between closely related species for nearly twenty years (LESSA 1992; SLADE *et al*. 1993). It is in recent years, however, that they have begun to be used with any real frequency. Introns are amenable to be applied to most of the same things as mitochondrial DNA, though they will be less effective at reflecting recent population changes due to their larger effective population size (FRIESEN 2000). Despite accumulating changes more rapidly than coding DNA, when compared to the more commonly used mitochondrial DNA, diploid spliceosomal intron alleles have an effective population size four times larger and mutate at a rate one quarter the rate of animal mtDNA. MtDNA is thought to evolve 5-10 times faster than nuclear DNA, based on restriction digests of the mitochondrial genome and thermostability studies of single-copy nuclear DNA (BROWN *et al*. 1979). As a result mtDNA haplotypes coalesce (become monophyletic) more rapidly than introns and so track recent speciation events more effectively. As discussed above, however, there are uncertainties regarding complete dependence on mtDNA and so developing intron markers and understanding their evolution is highly desirable. Aside from recent discussion regarding mtDNA's applicability for inferring population history, a gene tree obtained from mtDNA is still technically from only one locus and so may not be representative of the entire population history. By using more than one locus, we increase our ability to distinguish between the effects of selection and population demography, as the first will act locally and the second should present a common signature across many loci (HARE 2001).

Intron-exon structure of a gene is generally conserved over wide evolutionary expanses (PRYCHITKO and MOORE 1997). As a result, conserved exon sites flanking introns offer ideal sites to place primers that may cross-amplify across a range of species (FRIESEN 2000). These primers are commonly referred to as exon-primed, intron-crossing primers (EPIC) (Fig. 1.6). This strategy was introduced over 15 years ago (LESSA 1992; SLADE *et al*. 1993) but, unlike universal primers used in mitochondrial DNA which amplify successfully across different animal species, primers have yet to prove widely applicable (ZHANG and HEWITT 2003), and so a certain amount of empirical testing is required before applying them. EPIC primers are generally placed such that there are stretches of exon sequence obtained large enough to positively identify the amplification product (PRYCHITKO and MOORE 1997). There are different strategies one can take when amplifying intron markers. The first is to select previously used primers and markers and test them, preferentially choosing those which have been shown to work in your study species or which worked in a closely related species. Once a marker is successfully amplified, however, it is generally a good idea to create taxon-specific primers from the obtained sequence, as well as internal intron primers. The second approach is to utilize the abundant sequence data on GenBank (URL) and create EPIC primers from genomic and mRNA cDNA sequences, or generate genomic data for the species being studied and design primers from there.

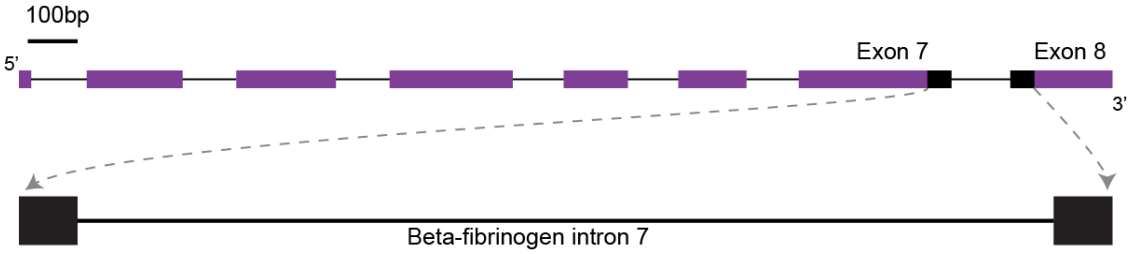


Figure 1.6: Structure of the human beta-fibrinogen gene, including both coding (exon) and noncoding (intron) regions.

Expanded is intron 7 with flanking segments of exons 7 and 8, to illustrate the concept of the EPIC primer method. primers Intron 7 has the same position in chickens and humans, and is regularly used in avian phylogenetics. The 100 bp scale applies to exon regions only. Adapted from Prychitko & Moore (1997).

The use of introns in phylogenetics is more complex than that of mitochondrial DNA, due to its diploid nature and the frequency of length-variant heterozygotes and insertion-deletion mutations, and the occurrence of recombination (CREER 2007).

Intragenic recombination, largely absent from animal mtDNA, occurs throughout the nuclear genome. If the recombination rate approaches the substitution rate, haplotypes will have more than one ancestor and different segments within the haplotype will have independent histories (HARE 2001; ZHANG and HEWITT 2003). This can seriously affect phylogenetic reconstruction, biasing a gene tree to show a false signature of population expansion (SCHIERUP and HEIN 2000). Methods exist, however, to estimate recombination rates and identify recombinants in multiple sequence alignments (e.g., RDP3: (MARTIN *et al*. 2010)). Recombination can then be incorporated into evolutionary models during data analysis (ZHANG and HEWITT 2003). Aside from potentially functional regions within introns, the linkage of whole introns or nucleotide stretches to a functionally important gene, as can occur in regions of low recombination, will affect the evolution of the intron through hitchhiking (MAYNARD SMITH and HAIGH 1974). This should be taken into account when analyzing intron data. Insertion/deletion polymorphism (indels) can also complicate data analysis, and direct sequence reads may appear superimposed if the individual is a heterozygote. They make up a large part of intron polymorphism and can potentially contain phylogenetic information, though most phylogenetic methods do not use this information efficiently (ZHANG and HEWITT 2003). Heterozygosity and allele discrimination is another problem with using nuclear regions. At a given locus, heterozygotes will present two different alleles or haplotypes, which need to be determined in order to extract the most information for genetic analyses (ZHANG and HEWITT 2003). There are experimental and analytical approaches to address this issue. For example, cloning of PCR products is a universally applicable method to phase heterozygotes, however it is costly, laborious, requires the analysis of numerous clones in order to pick up poorly represented alleles. Artifacts can also be introduced due to recombination occurring upon transformation of bacterial cells (ZHANG and HEWITT 2003). Statistical approaches such as the one implemented in PHASE (STEPHENS and DONNELLY 2003; STEPHENS *et al*. 2001) utilize the allelic information from homozygotes or heterozygotes differing only in one position to help resolve the phase of multi-site heterozygotes. Length-variant heterozygotes (the product of indel polymorphisms) can also be resolved analytically in some cases (DIXON 2010; DMITRIEV and RAKITOV 2008; FLOT 2007). These methods are not perfect, however, and incorrect phasing can affect downstream analyses (GARRICK *et al*. 2010).

1.5.4 Nuclear Introns for Penguins

A number of nuclear markers have been used previously in penguin phylogenetics. Nuclear protein coding loci used for phylogenetics include *RAG-1* (BAKER *et al*. 2006; ERICSON *et al*. 2006), and c-mos proto-oncogene (600bp exon fragment) (VAN TUINEN *et al*. 2001). Nuclear introns used for phylogenetics of penguins include 370bp fragment of intron 11 of glyceraldehyde-3-phosphodehydrogenase (*G3PDH*) (Adélie penguins) (VAN TUINEN *et al*. 2001), intron C of the gametologous avian chromo-helicase-DNA-binding protein (*CHD1Z*/*CHD1W*) (Adélie penguins) (SUNDSTRÖM *et al*. 2003), adenylate kinase intron 5 (*AK1i5*) (*Spheniscus mendiculus*) (SHAPIRO and DUMBACHER 2001), ß-fibrinogen intron 7 (*FGB7*) (*Eudyptula minor*, *Spheniscus humboldti*) (ERICSON *et al*. 2006; FAIN and HOUDE 2004), myoglobin intron 2 (*Spheniscus humboldti*, *Eudyptula minor*) (ERICSON *et al*. 2006; HACKETT *et al*. 2008), interferon regulatory factor 2 intron 2 (*Eudyptula minor*) (HACKETT *et al*. 2008), and ornithine decarboxylase intron 6-7 (*Spheniscus humboldti*) (ERICSON *et al*. 2006).

A large number of intron markers have now been developed and applied to avian phylogenetics. Of potential use for work in penguins are those developed or tested in closely related orders (e.g. Procellariformes, Ciconiiformes, Pelecaniiformes, Gaviiformes (HACKETT *et al*. 2008)). For example, four intron markers isolated in marbled murrelets (*Brachyramphus marmoratus*), which are coastal seabirds, may cross-amplify in penguins (FRIESEN *et al*. 1997). Several studies have sought to develop broadly applicable intron markers for avian phylogenetics from available genomic sources that may potentially cross-amplify in penguins as well (BACKSTRÖM *et al*. 2008; BORGE *et al*. 2005; KIMBALL *et al*. 2009; PRIMMER *et al*. 2002).

1.6 Evolutionary Rates

Evolution is change occurring across generations, that is a gradual accumulation of changes in steps. In order to understand evolution it is essential to characterize that change, and one important component of change is the rate at which that change

occurs. Understanding the rate of evolution, be it at a molecular or phenotypic level, is therefore a fundamental question in evolutionary biology. These rates are not constant across phylogenetic lineages (BROMHAM and PENNY 2003; KUMAR 2005). For example, evolutionary rates in avian species appear slower than most rates in mammals (MINDELL *et al*. 1996). Evolutionary rates are affected by different population processes, including selection and genetic drift. When a gene region is selectively neutral, it will accumulate mutations in a clock-like fashion (KIMURA 1983). Within phylogenetics and population genetics, rates of neutral molecular evolution are important parameters allowing researchers to investigate population processes including divergence times, population sizes, etc.

1.6.1 Methods of calculating evolutionary rates

The concept of evolutionary rate is often confused in the literature and may refer to mutation, pedigree or substitution rates (HO and LARSON 2006). Mutation rates typically refer to the rate at which instantaneous changes occur within the genome, excluding lethal or near-lethal mutations (HO and LARSON 2006). They can be estimated by mutant accumulation assays, for example. In these, the rate of change in a fraction of mutants arising from a large mutant-free popualtion is used to calculate the mutation rate (e.g., in *Caenorhabditis elegans* (DENVER *et al*. 2000)). Pedigree rates are an estimate of the mutation rate obtained by observing the number of nucleotide changes over a known genealogy of individuals (HO and LARSON 2006).

Substitution rates are difficult to measure directly, so traditionally they have been calculated using calibration methods. In these, the sequence difference among extant taxa is characterized and divided by the age of the most recent common ancestor, calibrated using fossil material or geological data (e.g., (SHIELDS and WILSON 1987). These rates estimate the frequency at which mutations are fixed within a population, as purifying selection or drift acts to remove the majority of changes in the genome. In the case of a perfectly neutral gene region, mutation rates and substitution rates will be the same (KIMURA 1983). Shields and Wilson (1987) utilized this method to estimate the rate of evolution of the mitochondrial genome in birds, using two species

of geese (*Anser* and *Branta*) and found a mean substitution rate of 0.02 s/s/Myr. (QUINN *et al*. 1991) estimated a rate of substitution of 0.208 s/s/Myr for the HVR-I of *Branta* subspecies based on the divergence between them and the overall mtDNA rate estimated by Shields and Wilson (1987). These rates have been widely used in avian phylogenetics.

Ancient DNA techniques provide an alternative method for estimating substitution rates. Genetic changes can be measured from serially preserved samples. The first study using this technique calculated an evolutionary rate of the hypervariable region I (*HVRI*) of the mitochondrial control region, from 344bps of sequence using 96 known age Adélie penguin subfossil bones (LAMBERT *et al*. 2002). A Bayesian Markov chain Monte Carlo approach was used, currently implemented in the software Bayesian Evolutionary Analysis Sampling Trees (BEAST) (DRUMMOND and RAMBAUT 2007).

1.6.2 Apparent time-dependency of rates

As a result of different approaches to estimating mitochondrial evolutionary and mutation rates, discrepancies between estimates originating from the different methods have become apparent. Rates based on phylogenetic calibrations are lower than those based on serially preserved samples (aDNA), and again, these are substantially lower than mutation rates obtained from pedigree studies and mutant accumulation studies. Neutral theory holds that the rate of mutation (μ) is equal to the rate of evolution (K) for neutrally evolving sequences (KIMURA 1983). One hypothesis that has arisen to explain these differences is the "time dependency of molecular rates". This hypothesis states that the relationship between the age of calibration of a molecular rate (short-term for mutation rates and long-term for phylogenetic rates) and the rate of change can be described by a vertically translated exponential decay curve (HO *et al*. 2005). The authors suggest using this curve to correct molecular date estimates (HO 2007; HO *et al*. 2005).

A heated debate followed. Several studies supported this concept, observing higher rates within or between closely related species compared to more distantly related species (BURRIDGE *et al*. 2008; GRATTON *et al*. 2008; HENN *et al*. 2009; HO *et al*. 2007; HOWELL *et al*. 2008). Other studies have criticized the time-dependency hypothesis for not taking into account the methodological upward bias in rate estimation produced by the Bayesian-MCMC approach (BANDELT 2008; DEBRUYNE and POINAR 2009; EMERSON 2007; NAVASCUÉS and EMERSON 2009). A recent study in Adélie penguins sought to address this question within a single species, by estimating a mutation rate following a pedigree approach, and an evolutionary rate from ancient DNA data, for the *HVR-I* region of the mitochondrial genome (MILLAR *et al*. 2008a). The pedigree data, from 508 families, was analysed using a model to correct for the effect of heteroplasmies on mutation rate estimation, and gave an estimate of μ=0.55 mutations/site/Myrs, while the ancient DNA approach, using 162 known-age subfossil bones, gave an estimate of K=0.86 substitutions/site/Myrs. These rates were not significantly different and support the assumption of the *HVR-I*'s neutrality and find no support for the time-dependency hypothesis. However, population genetics theories do predict a time dependent rate at selectively constrained sites due to the removal of slightly deleterious mutations, in contrast to the expectation at neutral sites (no time dependency). Subramanian & Lambert (2011) sought to re-examine the concept of time-dependency by estimating rates separately for neutral and constrained sites in primate mitochondrial genomes. They found no differences between rates calibrated at different evolutionary timescales for synonymous sites; however, an order of magnitude variation at constrained sites was detected. It appears that time-dependency is valid for constrained sites, but not for neutrally evolving regions (SUBRAMANIAN and LAMBERT 2011).

1.7 Intron evolutionary rates and the potential of ancient DNA techniques

No direct estimates of intron rates of evolution have been obtained using ancient DNA techniques. All available rates stem from comparisons to mitochondrial genes or by using calibration techniques. Mitochondrial DNA sequences are said, in general, to evolve on average 5-10 times faster than nuclear genes (WILSON *et al*.

1985). As a general rule, it has been suggested that substitutions in introns occur at a mean rate approximately one-sixtieth that of the mitochondrial control region (FRIESEN 2000). A range of introns amplified in murrelets (*Brachramphus marmoratus*; Charadriiormes, Alcidae) showed divergence rates one-quarter the mean of mtDNA (0.45% Ma^{-1} compared to 2% Ma^{-1}). Divergence rates were highest for the P40 intron (0.72+/-0.04% Ma^{-1}), and lowest for the tropomyosin intron (0.16+/-0.01% Ma^{-1}) (CONGDON *et al*. 2000). The substitution rate for α-enolase intron VIII was estimated as 3.7 x 10^{-6} s/s/gen and 4.5 x 10^{-6} s/s/gen for least and crested auklets (Aethia spp., generation times 6.7 and 8.3 years respectively) by comparing auklet sequences to marbled murrelet sequences and assuming a divergence time of 12 MYA (WALSH *et al*. 2005). In a comparative study of the evolution of ß-fibrinogen intron 7 (*FGB7*) and the mitochondrial cytochrome *b* (*cytb*) gene, *cytb* evolved 2.8 times as fast as *FGB7* (14 times as fast at third codon positions) (PRYCHITKO and MOORE 2000). However the phylogenetic signal between the two was comparable due to *FGB7*'s more uniform distribution of nucleotide substitutions and lower base composition bias. Homoplasy was lower, and all nucleotides appeared free of selective constraints, increasing the number of informative sites per unit sequenced (PRYCHITKO and MOORE 2000).

Ancient DNA techniques, as mentioned above, have been used to estimate evolutionary rates of mitochondrial genes, for example in Adélie penguins (LAMBERT *et al*. 2002; MILLAR *et al*. 2008a), tuatara (HAY *et al*. 2008), bison (SHAPIRO *et al*. 2004), among others. To date this has not been attempted for nuclear sequences. The reason for this is related to the technical difficulties inherent in ancient DNA studies – template sequences for analysis are generally low in number and highly degraded, which makes obtaining a large amount of sequence difficult. The presence of miscoding lesions in DNA templates complicates the use of ancient DNA further. Cytosine residues tend to become deaminated, changing to uracil residues, which are then identified as thymine by DNA polymerases, leading to an observed change and potential overestimation of sequence divergence (GILBERT *et al*. 2005; PÄÄBO *et al*. 2004).

One important point regarding ancient DNA evolutionary rate estimates is that they typically rely on Bayesian MCMC statistical methods like BEAST (DRUMMOND and RAMBAUT 2007). These methods are generally based on simple demographic models that do not take into account recombination, migration, population subdivision, bottlenecks or DNA damage. This may make such methods susceptible to upward biases in the estimation of substitution rates when the models used are severely ill-fitting to the data, or indeed if the data are not informative enough (NAVASCUÉS and EMERSON 2009).

Most ancient DNA studies work with mitochondrial DNA, as there are approximately 10^2-10^5 mitochondria per somatic cell, compared to just two copies of nuclear genes. That said, as the field has progressed and new techniques are being applied, ancient DNA studies are increasingly going nuclear (GREEN *et al*. 2006; NOONAN *et al*. 2006; NOONAN *et al*. 2005). Initially the application of the multiplex PCR technique, in which numerous primers are designed to amplify small overlapping sequences in just a few reactions, was utilized effectively to obtain a large amount of sequence data from a small amount of starting template (RÖMPLER *et al*. 2006a). This technique has been used successfully to obtain the full mitochondrial genome from fossil material of a number of animals, for example, mammoth and Adélie penguins (KRAUSE *et al*. 2006; SUBRAMANIAN *et al*. 2009). For nuclear studies using ancient DNA, this technique was successful applied to sequence the melanocortin-1 receptor gene of wooly mammoths (RÖMPLER *et al*. 2006b). This suggests that for nuclear introns, a multiplex technique could also yield a large amount of sequence data at once, without being substantially more expensive than ordinary methods for obtaining DNA sequences. While multiplex techniques are potentially useful for this objective, attempting to generate enough intron sequence data for rate estimation is still not straightforward. Intron sequence from a range of known-age samples is needed. These sequences need to be repeated independently either through cloning PCR products or through repeated amplification and sequencing rounds in order to reliably identify real substitutions (including heterozygote substitutions) and differentiate them from miscoding lesions.

The major technological advances that have given rise to so-called second-generation sequencing, however, now allow massively parallel amplification of large amounts of the genome at one time. Two seminal ancient DNA studies introduced the potential of these new technologies to sequence ancient nuclear genomes, in Neanderthals (GREEN *et al*. 2006; NOONAN *et al*. 2006). There are a number of different technologies that fall under the umbrella of second-generation sequencing (MILLAR *et al*. 2008b), but two in particular have risen to the forefront and are the most frequently used at present. The first is the FLX (Roche) sequencer, based on emulsion PCR of bead-anchored oligos (NAKANO *et al*. 2003), clonal plate amplification and pyrosequencing using light emission and detection (MARGULIES *et al*. 2005; RONAGHI *et al*. 1998). The second is the Solexa (Illumina) sequencer, based on solid-phase-anchored oligo bridge amplification (ADAMS and KRON 1997) and cluster sequencing using reversible fluorescent dNTP terminators (JU *et al*. 2006). FLX Titanium sequencers currently deliver up to 400 bp read lengths and approximately 400-600 million bases per run, while Illumina sequencers produce up to 48 gigabases of DNA up to two times 100 bp in paired end reads (KNAPP and HOFREITER 2010). While initially these sequencers have been of great use for whole-genome sequencing, the cost of a run was too prohibitive to engage in population level studies, until sample or primer barcoding protocols were introduced (BINLADEN *et al*. 2007; MEYER *et al*. 2008). The potential of combining a multiplex approach to target a large number of loci with a massively-parallel barcoded sequencing technology (STILLER *et al*. 2009) means that the time is now ripe to take on population studies utilizing ancient nuclear DNA.

1.8 Aims of this PhD

Overall, this thesis aims to shed further light on the population genetics of Adélie penguins by analysing available and new mitochondrial *HVRI* DNA sequences as well as completely novel intron sequences, and simultaneously utilise the Adélie penguin as a model species to investigate more general evolutionary questions.

The specific aims for each chapter of this PhD thesis are described below. In Chapter One (this chapter), literature pertinent to each of the following chapters is summarized and research questions are indicated. In Chapter Two, I sought to utilise the Adélie penguin as a model species with which to address the question of whether mitochondrial DNA diversity serves as a proxy of population size. In Chapter Three I aimed to use the methods identified in Chapter Two to obtain an estimate of past population sizes of the extinct New Zealand Huia. In Chapter Four I aimed to identify candidate intron markers for population and phylogenetic studies of Adélie penguins. In Chapter Five I sought to investigate broad population structure and variability at five intron loci, with the null hypothesis that intron population structure patterns will not match mitochondrial patterns, and in parallel assess the utility of four intron loci for phylogenetics of penguins. In Chapter Six I sought to obtain ancient intron sequences from Adélie penguin subfossil bones in order to gauge allele frequency changes and to test a method used successfully for mitochondrial DNA from ancient samples.

1.9 References

ADAMS, C. P., and S. J. KRON, 1997 Method for performing amplification of nucleic acid with two primers bound to a single solid support, pp. 658, edited by W. I. F. B. RESEARCH, U.S.A.

AINLEY, D. G., 2002 *The Adélie Penguin: Bellwether of Climate Change*. Columbia University Press.

ATKINSON, Q. D., R. D. GRAY and A. J. DRUMMOND, 2008 mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory. Molecular Biology and Evolution **25:** 468-474.

AVISE, J. C., 1992 Molecular Population Structure and the Biogeographic History of a Regional Fauna: A Case History with Lessons for Conservation Biology. Oikos **63:** 62-76.

AVISE, J. C., 1994 *Molecular markers, natural history and evolution*. Chapman & Hall.

BACKSTRÖM, N., S. FAGERBERG and H. ELLEGREN, 2008 Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Molecular Ecology **7:** 964-980.

BAKER, A. J., S. L. PEREIRA, O. P. HADDRATH and K.-A. EDGE, 2006 Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. Proceedings of the Royal Society Series B **273:** 11-17.

BALLARD, J. W. O., 1995 Is mitochondrial DNA a strictly neutral marker? Trends in Ecology and Evolution **10:** 485-488.

BALLARD, J. W. O., and M. C. WHITLOCK, 2004 The incomplete natural history of mitochondria. Molecular Ecology **13:** 729-744.

BANDELT, H.-J., 2008 Time dependency of molecular rate estimates: tempest in a teacup. Heredity **100:** 1-2.

BARONI, C., and G. OROMBELLI, 1994 Abandoned penguin rookeries as Holocene paleoclimatic indicators in Antarctica. Geology **22:** 23-26.

BARR, C. M., M. NEIMAN and D. R. TAYLOR, 2005 Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytologist **168:** 39-50.

BAZIN, E., S. GLÉMIN and N. GALTIER, 2006 Population size does not influence mitochondrial genetic diversity in animals. Science **312:** 570-571.

BEERLI, P., and J. FELSENSTEIN, 1999 Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations Using a Coalescent Approach. Genetics **152:** 763-774.

BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. PNAS **98:** 4563-4568.

BELSHAW, R., and D. BENSASSON, 2006 The rise and falls of introns. Heredity **96:** 208-213.

BINLADEN, J., M. T. P. GILBERT, J. P. BOLLBACK, F. PANITZ, C. BENDIXEN *et al.*, 2007 The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. PLoS ONE **2:** e197.

BORGE, T., M. T. WEBSTER, G. ANDERSSON and G.-P. SAETRE, 2005 Contrasting Patterns of Polymorphism and Divergence on the Z Chromosome and Autosomes in Two Ficedula Flycatcher Species. Genetics **171:** 1861-1873.

BROMHAM, L., and D. PENNY, 2003 The modern molecular clock. Nat Rev Genet **4:** 216 - 224.

BROWN, W., E. PRAGER, A. WANG and A. WILSON, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol **18:** 225 - 239.

BROWN, W. M., M. GEORGE and A. C. WILSON, 1979 Rapid Evolution of Animal Mitochondrial DNA. PNAS **76:** 1967-1971.

BURRIDGE, C. P., D. CRAW, D. FLETCHER and J. M. WATERS, 2008 Geological Dates and Molecular Rates: Fish DNA Sheds Light on Time-Dependency. Molecular Biology and Evolution **25:** 624-633.

CAVALIER-SMITH, T., 1985 Selfish DNA and the origin of introns. Nature **315:** 283-284.

CHAMARY, J.-V., and L. D. HURST, 2004 Similar Rates but Different Modes of Sequence Evolution in Introns and at Exonic Silent Sites in Rodents: Evidence for Selectively Driven Codon Usage. Molecular Biology and Evolution **21:** 1014-1023.

CHAMARY, J. V., J. L. PARMLEY and L. D. HURST, 2006 Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7:** 98-108.

CLARKE, J. A., D. T. KSEPKA, M. STUCCHI, M. URBINA, N. GIANNINI *et al.*, 2007 Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. Proceedings of the National Academy of Sciences of the United States of America **104:** 11545-11550.

COLLINS, L., and D. PENNY, 2005 Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. Molecular Biology and Evolution **22:** 1053-1066.

CONGDON, B. C., J. F. PIATT, K. MARTIN and V. L. FRIESEN, 2000 Mechanisms of Population Differentiation in Marbled Murrelets: Historical versus Contemporary Processes. Evolution **54:** 974-986.

CRANDALL, K. A., D. POSADA and D. VASCO, 1999 Effective population sizes: missing measures and missing concepts. Animal Conservation **2:** 317-320.

CREER, S., 2007 Choosing and using introns in molecular phylogenetics. Evolutionary Bioinformatics **3:** 99-108.

DEBRUYNE, R., and H. N. POINAR, 2009 Time Dependency of Molecular Rates in Ancient DNA Data Sets, A Sampling Artifact? Systematic Biology **58:** 348-360.

DENVER, D. R., K. MORRIS, M. LYNCH, L. L. VASSILIEVA and W. K. THOMAS, 2000 High Direct Estimate of the Mutation Rate in the Mitochondrial Genome of Caenorhabditis elegans. Science **289:** 2342-2344.

DESJARDINS, P., and R. MORAIS, 1990 Sequence and gene organization of the chicken mitochondrial genome : A novel gene order in higher vertebrates. Journal of Molecular Biology **212:** 599-634.

DIXON, C. J., 2010 OLFinder--a program which disentangles DNA sequences containing heterozygous indels. Molecular Ecology Resources **10:** 335-340.

DMITRIEV, D. A., and R. A. RAKITOV, 2008 Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels. PLoS Computational Biology **4:** e1000113.

DRUMMOND, A., and A. RAMBAUT, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology **7:** 214.

DRUMMOND, A. J., A. RAMBAUT, B. SHAPIRO and O. G. PYBUS, 2005 Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Molecular Biology and Evolution **22:** 1185-1192.

DUGGER, K. M., D. G. AINLEY, P. O. B. LYVER, K. BARTON and G. BALLARD, 2010 Survival differences and the effect of environmental instability on breeding dispersal in an Adélie penguin meta-population. Proceedings of the National Academy of Sciences **107:** 12375-12380.

DURET, L., 2001 Why do genes have introns? Recombination might add a new piece to the puzzle. Trends in Genetics **17:** 172-175.

EMERSON, B. C., 2007 Alarm Bells for the Molecular Clock? No Support for Ho et al.'s Model of Time-Dependent Molecular Rate Estimates. Systematic Biology **56:** 237-246.

EMERSON, B. C., E. PARADIS and C. THÉBAUD, 2001 Revealing the demographic histories of species using DNA sequences. Trends in Ecology and Evolution **16:** 707-716.

ERICSON, P. G. P., C. L. ANDERSON, T. BRITTON, A. ELZANOWSKI, U. S. JOHANSSON *et al.*, 2006 Diversification of Neoaves: integration of molecular sequence data and fossils. Biology Letters **2:** 543-547.

EYRE-WALKER, A., 2006 Size Does Not Matter for Mitochondrial DNA. Science **312:** 537-538.

EYRE-WALKER, A., N. H. SMITH and J. M. SMITH, 1999 How clonal are human mitochondria? Proceedings of the Royal Society of London. Series B: Biological Sciences **266:** 477-483.

FAIN, M. G., and P. HOUDE, 2004 Parallel Radiations in the Primary Clades of Birds. Evolution **58:** 2558-2573.

FLOT, J.-F., 2007 CHAMPURU 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Molecular Ecology Notes **7:** 974-977.

FRANKHAM, R., 1996 Relationship of Genetic Variation to Population Size in Wildlife. Conservation Biology **10:** 1500-1508.

FRANKHAM, R., J. D. BALLOU and D. A. BRISCOE, 2005 *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge.

FRIESEN, V. L., 2000 Introns, pp. 274-294 in *Molecular Methods in Ecology*, edited by A. J. BAKER. Blackwell Science Ltd, Oxford.

FRIESEN, V. L., B. C. CONGDON, M. G. KIDD and T. P. BIRT, 1999 Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. Molecular Ecology **8:** 2147-2149.

FRIESEN, V. L., B. C. CONGDON, H. E. WALSH and T. P. BIRT, 1997 Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. Molecular Ecology **6:** 1047-1058.

FRY, A. J., 1999 Mildly Deletrious Mutations in Avian Mitochondrial DNA: Evidence from Neutrality Tests. Evolution **53:** 1617-1620.

GARRICK, R., P. SUNNUCKS and R. DYER, 2010 Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evolutionary Biology **10:** 118.

GAZAVE, E., T. MARQUES-BONET, O. FERNANDO, B. CHARLESWORTH and A. NAVARRO, 2007 Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol **8:** R21.

GEMMELL, N. J., M. K. SCHWARTZ and B. C. ROBERTSON, 2004 Moa were many. Proceedings of the Royal Society Series B (Suppl) **271:** S430-S432.

GILBERT, M. T. P., H.-J. BANDELT, M. HOFREITER and I. BARNES, 2005 Assessing ancient DNA studies. Trends in Ecology and Evolution **20:** 541-544.

GILBERT, W., 1978 Why genes in pieces? Nature **271:** 501.

GILBERT, W., 1987 The Exon Theory of Genes. Cold Spring Harbor Symposia on Quantitative Biology **52:** 901-905.

GILLESPIE, J. H., 2000 Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. Genetics **155:** 909-919.

GILLESPIE, J. H., 2001 Is the Population Size of a Species Relevant to its Evolution? Evolution **55:** 2161-2169.

GRATTON, P., M. K. KONOPINSKI and V. SBORDONI, 2008 Pleistocene evolutionary history of the Clouded Apollo (Parnassius mnemosyne): genetic signatures of climate cycles and a 'time-dependent' mitochondrial substitution rate. Molecular Ecology **17:** 4248-4262.

GREEN, R. E., J. KRAUSE, S. E. PTAK, A. W. BRIGGS, M. T. RONAN *et al.*, 2006 Analysis of one million base pairs of Neanderthal DNA. Nature **444:** 330-336.

HACKETT, S. J., R. T. KIMBALL, S. REDDY, R. C. K. BOWIE, E. L. BRAUN *et al.*, 2008 A Phylogenomic Study of Birds Reveals Their Evolutionary History. Science **320:** 1763-1768.

HAHN, M. W., 2008 Toward a Selection Theory of Molecular Evolution. Evolution **62:** 255-265.

HALL, B. L., A. R. HOELZEL, C. BARONI, G. H. DENTON, B. J. L. BOEUF *et al.*, 2006 Holocene elephant seal distribution implies warmer-than-present climate in the Ross Sea. PNAS **103:** 10213-10217.

HARE, M. P., 2001 Prospects for nuclear gene phylogeography. Trends in Ecology and Evolution **15:** 700-706.

HARE, M. P., F. CIPRIANO and S. R. PALUMBI, 2002 Genetic Evidence on the Demography of Speciation in Allopatric Dolphin Species. Evolution **56:** 804-816.

HARRISON, R. G., 1989 Animal Mitochondrial DNA as a Genetic Marker in Population and Evolutionary Biology. Trends in Ecology and Evolution **4:** 6-11.

HAUGEN, P., D. M. SIMON and D. BHATTACHARYA, 2005 The natural history of group I introns. Trends in Genetics **21:** 111-119.

HAY, J. M., S. SUBRAMANIAN, C. D. MILLAR, E. MOHANDESAN and D. M. LAMBERT, 2008 Rapid molecular evolution in a living fossil. Trends in Genetics **24:** 106-109.

HENN, B. M., C. R. GIGNOUX, M. W. FELDMAN and J. L. MOUNTAIN, 2009 Characterizing the Time Dependency of Human Mitochondrial DNA Mutation Rate Estimates. Molecular Biology and Evolution **26:** 217-230.

HO, S. Y. W., 2007 Calibrating molecular estimates of substitution rates and divergence times in birds. Journal of Avian Biology **38:** 409-414.

HO, S. Y. W., and G. LARSON, 2006 Molecular clocks: when times are a-changin'. Trends in Genetics **22:** 79-83.

HO, S. Y. W., M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2005 Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Molecular Biology and Evolution **22:** 1561-1568.

HO, S. Y. W., B. SHAPIRO, M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2007 Evidence for Time Dependency of Molecular Rate Estimates. Systematic Biology **56:** 515-522.

HOWELL, N., C. HOWELL and J. L. ELSON, 2008 Molecular clock debate: Time dependency of molecular rate estimates for mtDNA: this is not the time for wishful thinking. Heredity **101:** 107-108.

HUGHES, A. L., and M. A. K. HUGHES, 2007 Coding sequence polymorphism in avian mitochondrial genomes reflects population histories. Molecular Ecology **16:** 1369-1376.

HUGHES, A. L., and M. YEAGER, 1997 Comparative Evolutionary Rates of Introns and Exons in Murine Rodents. Journal of Molecular Evolution **45:** 125-130.

HURST, G. D. D., and F. M. JIGGINS, 2005 Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society Series B **272:** 1525-1534.

JEFFARES, D. C., T. MOURIER and D. PENNY, 2006 The biology of intron gain and loss. Trends in Genetics **22:** 16-22.

JEFFARES, D. C., A. M. POOLE and D. PENNY, 1998 Relics from the RNA World. Journal of Molecular Evolution **46:** 18-36.

JENNINGS, W. B., and S. V. EDWARDS, 2005 Speciational History of Australian Grass Finches (*Poephila*) inferred from Thirty Gene Trees. Evolution **59:** 2033-2047.

JOHNSON, K. P., 2004 Deletion Bias in Avian Introns over Evolutionary Timescales. Molecular Biology and Evolution **21:** 599-602.

JU, J., D. H. KIM, L. BI, Q. MENG, X. BAI *et al.*, 2006 Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proceedings of the National Academy of Sciences **103:** 19635-19640.

KARL, S. A., and J. C. AVISE, 1993 PCR-based Assays of Mendelian Polymorphisms from Anonymous Single-Copy Nuclear DNA: Techniques and Applications for Population Genetics. Molecular Biology and Evolution **10:** 342-361.

KASAMATSU, H., and J. VINOGRAD, 1974 Replication of Circular DNA in Eukaryotic Cells. Annual Review of Biochemistry **43:** 695-719.

KIMBALL, R. T., E. L. BRAUN, F. K. BARKER, R. C. K. BOWIE, M. J. BRAUN *et al.*, 2009 A well-tested set of primers to amplify regions spread across the avian genome. Molecular Phylogenetics and Evolution **50:** 654-660.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KNAPP, M., and M. HOFREITER, 2010 Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. Genes **1:** 227-243.

KOONIN, E. V., 2006 The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biology Direct **1:** 22.

KRAUSE, J., P. H. DEAR, J. L. POLLACK, M. SLATKIN, H. SPRIGGS *et al.*, 2006 Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature **439:** 724-727.

KUMAR, S., 2005 Molecular clocks: four decades of evolution. Nat Rev Genet **6:** 654-662.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of Evolution in Ancient DNA from Adélie Penguins. Science **295:** 2270-2273.

LEBERG, P., 2005 Genetic approaches for estimating the effective size of populations. Journal of Wildlife Management **69:** 1385-1399.

LESSA, E. P., 1992 Rapid Surveying of DNA Sequence Variation in Natural Populations. Molecular Biology and Evolution **9:** 323-330.

LIMA-DE-FARIA, A., 1995 The formation of new mosaoc proteins and the sudden reappearance of old proteins, pp. 261-270 in *Biological Periodicity: Its Molecular Mechanism and Evoltionary Implications*, edited by A. LIMA-DE-FARIA.

LYNCH, M., 2002 Intron evolution as a population-genetic process. PNAS **99:** 6118-6123.

MARAIS, G., P. NOUVELLET, P. D. KEIGHTLEY and B. CHARLESWORTH, 2005 Intron Size and Exon Evolution in Drosophila. Genetics **170:** 481-486.

MARGULIES, M., M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376-380.

MARTIN, D. P., P. LEMEY, M. LOTT, V. MOULTON, D. POSADA *et al.*, 2010 RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26:** 2462-2463.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res., Camb. **23:** 23-36.

MCCUSKER, M. R., and P. BENTZEN, 2010 Positive relationships between genetic diversity and abundance in fishes. Molecular Ecology **19:** 4852-4862.

MCNEILL, S., K. BARTON, P. LYVER and D. PAIRMAN, 2011 Semi-automated penguin counting from digital aerial photographs, pp. 4312-4315 in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*.

MEYER, M., U. STENZEL and M. HOFREITER, 2008 Parallel tagged sequencing on the 454 platform. Nature Protocols **3:** 267-278.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al.*, 2008a Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

MILLAR, C. D., L. HUYNEN, S. SUBRAMANIAN, E. MOHANDESAN and D. M. LAMBERT, 2008b New developments in ancient genomics. Trends in Ecology & Evolution **23:** 386-393.

MINDELL, D., A. KNIGHT, C. BAER and C. HUDDLESTON, 1996 Slow Rates of Molecular Evolution in Birds and the Metabolic Rate and Body Temperature Hypotheses. Molecular Biology and Evolution **13:** 422.

MULLIGAN, C. J., A. KITCHEN and M. M. MIYAMOTO, 2006 Comment on "Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals". Science **314:** 1390a.

NAKANO, M., J. KOMATSU, S.-I. MATSUURA, K. TAKASHIMA, S. KATSURA *et al.*, 2003 Single-molecule PCR using water-in-oil emulsion. Journal of Biotechnology **102:** 117-124.

NAVASCUÉS, M., and B. C. EMERSON, 2009 Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. Molecular Ecology **18:** 4390-4397.

NILSEN, T. W., 2003 The spliceosome: the most complex macromolecular machine in the cell? BioEssays **25:** 1147-1149.

NOONAN, J. P., G. COOP, S. KUDARAVALLI, D. SMITH, J. KRAUSE *et al.*, 2006 Sequencing and Analysis of Neanderthal Genomic DNA. Science **314:** 1113-1119.

NOONAN, J. P., M. HOFREITER, D. SMITH, J. R. PRIEST, N. ROHLAND *et al.*, 2005 Genomic Sequencing of Pleistocene Cave Bears. Science **309:** 597-600.

OHTA, T., 1992 The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Evol. Syst. **23:** 263-286.

OHTA, T., 2003 Origin of the neutral and nearly neutral theories of evolution. Journal of Biosciences **28:** 371-377.

PÄÄBO, S., H. POINAR, D. SERRE, V. JAENICKE-DESPRÉS, J. HEBLER *et al.*, 2004 Genetic Analyses from Ancient DNA. Annu. Rev. Genet. **38:** 645-681.

PAKENDORF, B., and M. STONEKING, 2005 Mitochondrial DNA and human evolution. Annual Review of Genomics and Human Genetics **6:** 165-183.

PAMILO, P., and M. NEI, 1988 Relationships between gene trees and species trees. Molecular Biology and Evolution **5:** 568-583.

PEREIRA, S. L., 2000 Mitochondrial genome organization and vertebrate phylogenetics. Genetics and Molecular Biology **23:** 745-752.

POOLE, A. M., D. C. JEFFARES and D. PENNY, 1998 The path from the RNA world. Journal of Molecular Evolution **46:** 1-17.

PRIMMER, C. R., T. BORGE, J. LINDELL and G.-P. SAETRE, 2002 Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. Molecular Ecology **11:** 603-612.

PRYCHITKO, T. M., and W. S. MOORE, 1997 The Utility of DNA Sequences of an Intron from the b-Fibrinogen Gene in Phylogenetic Analysis of Woodpeckers (Aves: Picidae). Molecular Phylogenetics and Evolution **8:** 193-204.

PRYCHITKO, T. M., and W. S. MOORE, 2000 Comparative Evolution of the Mitochondrial Cytochrome *b* Gene and Nuclear b-Fibrinogen Intron 7 in Woodpeckers. Molecular Biology and Evolution **17:** 1101-1112.

QUINN, T. W., G. F. SHIELDS and A. C. WILSON, 1991 Affinities of the Hawaiian Goose Based on Two Types of Mitochondrial DNA Data. The Auk **108:** 585-593.

RAGG, H., 2011 Intron creation and DNA repair. Cellular and Molecular Life Sciences **68:** 235-242.

RAND, D. M., and L. M. KANN, 1996 Excess Amino Acid Polymorphism in Mitochondrial DNA: Contrasts Among Genes from *Drosophila*, Mice, and Humans. Molecular Biology and Evolution **13:** 735-748.

RANDI, E., 2000 Mitochondrial DNA, pp. 136-167 in *Molecular Methods in Ecology*, edited by A. J. BAKER. Blackwell Publishing Ltd, Oxford.

REST, J. S., J. C. AST, C. C. AUSTIN, P. J. WADDELL, E. A. TIBBETTS *et al.*, 2003 Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. Molecular Phylogenetics and Evolution **29:** 289-297.

RITCHIE, P. A., and D. M. LAMBERT, 2000 A repeat complex in the mitochondrial control region of Adélie penguins from Antarctica. Genome **43:** 613-618.

RITCHIE, P. A., C. D. MILLAR, G. C. GIBB, C. BARONI and D. M. LAMBERT, 2004 Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. Molecular Biology and Evolution **21:** 240-248.

ROMAN, J., and S. R. PALUMBI, 2003 Whales Before Whaling in the North Atlantic. Science **301:** 508-510.

RÖMPLER, H., P. H. DEAR, J. KRAUSE, M. MEYER, N. ROHLAND *et al.*, 2006a Multiplex amplification of ancient DNA. Nature Protocols **1:** 720-728.

RÖMPLER, H., N. ROHLAND, C. LALUEZA-FOX, E. WILLERSLEV, T. KUZNETSOVA *et al.*, 2006b Nuclear Gene Indicates Coat-Color Polymorphism in Mammoths. Science **313:** 62.

RONAGHI, M., M. UHLÉN and P. NYRÉN, 1998 A Sequencing Method Based on Real-Time Pyrophosphate. Science **281:** 363-365.

ROY, S. W., 2003 Recent Evidence for the Exon Theory of Genes. Genetica **118:** 251-266.

ROY, S. W., and W. GILBERT, 2006 The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet **7:** 211-221.

ROY, S. W., and M. IRIMIA, 2009 Mystery of intron gain: new data and new models. Trends in genetics **25:** 67-73.

SCHIERUP, M. H., and J. HEIN, 2000 Consequences of Recombination on Traditional Phylogenetic Analysis. Genetics **156:** 879-891.

SCHLÖTTERER, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109:** 365-371.

SCHWARTZ, M. K., D. A. TALLMON and G. LUIKART, 1999 Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. Animal Conservation **2:** 321-323.

SHAPIRO, B., A. J. DRUMMOND, A. RAMBAUT, M. C. WILSON, P. E. MATHEUS *et al.*, 2004 Rise and Fall of the Beringian Steppe Bison. Science **306:** 1561-1565.

SHAPIRO, L. H., and J. P. DUMBACHER, 2001 Adenylate Kinase Intron 5: A New Nuclear Locus for Avian Systematics. The Auk **118:** 248-255.

SHEPHERD, L. D., C. D. MILLAR, G. BALLARD, D. G. AINLEY, P. R. WILSON *et al.*, 2005 Microevolution and mega-icebergs in the Antarctic. PNAS **102:** 16717-16722.

SHIELDS, G. F., and A. C. WILSON, 1987 Calibration of the mitochondrial DNA evolution in geese. Journal of Molecular Evolution **24:** 212-217.

SLADE, R. W., C. MORITZ, A. HEIDEMAN and P. T. HALE, 1993 Rapid assessment of single-copy nuclear DNA variation in diverse species. Molecular Ecology **2:** 359-373.

SORENSON, M. D., J. C. AST, D. E. DIMCHEFF, T. YURI and D. P. MINDELL, 1999 Primers for a PCR-Based Approach to Mitochondrial Genome Sequencing in Birds and Other Vertebrates. Molecular Phylogenetics and Evolution **12:** 105-114.

SOULÉ, M. E., 1976 Allozyme variation, its determinants in space and time. , pp. 60-77 in *Molecular Evolution*, edited by F. J. AYALA. Sinauer Associates, Sunderland, Massachusetts.

SOUZA, S. J. D., 2003 The emergence of a synthetic theory of intron evolution. Genetica **118:** 117-122.

STEPHENS, M., and P. DONNELLY, 2003 A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet **73:** 1162 - 1169.

STEPHENS, M., N. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68:** 978 - 989.

STILLER, M., M. KNAPP, U. STENZEL, M. HOFREITER and M. MEYER, 2009 Direct multiplex sequencing (DMPS) -- a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Research **19:** 1843-1848.

SUBRAMANIAN, S., D. R. DENVER, C. D. MILLAR, T. HEUPINK, A. ASCHRAFI *et al.*, 2009 High mitogenomic evolutionary rates and time dependency. Trends in Genetics **25:** 482-468.

SUBRAMANIAN, S., and D. M. LAMBERT, 2011 Time dependency of molecular evolutionary rates? Yes and No. Genome Biology and Evolution **3:** 1324-1328.

SUNDSTRÖM, H., M. T. WEBSTER and H. ELLEGREN, 2003 Is the Rate of Insertion and Deletion Mutation Male Biased?: Molecular Evolutionary ANalysis of Avian and Primate Sex Chromosome Sequences. Genetics **164:** 259-268.

THOMSON, R. C., I. J. WANG and J. R. JOHNSON, 2010 Genome-enabled development of DNA markers for ecology, evolution and conservation. Molecular Ecology **19:** 2184-2195.

TOWNSEND, T., R. ALEGRE, S. KELLEY, J. WIENS and T. REEDER, 2008 Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. Mol Phylogenet Evol **47:** 129 - 142.

TSAOUSIS, A. D., D. P. MARTIN, E. D. LADOUKAKIS, D. POSADA and E. ZOUROS, 2005 Widespread Recombination in Published Animal mtDNA Sequences. Molecular Biology and Evolution **22:** 925-933.

VAN TUINEN, M., D. B. BUTVILL, J. A. W. KIRSCH and S. B. HEDGES, 2001 Convergence and divergence in the evolution of aquatic birds. Proceedings of the Royal Society of London Series B **268:** 1345-1350.

WALSH, H. E., I. L. JONES and V. L. FRIESEN, 2005 A Test of Founder Effect Speciation Using Multiple Loci in the Auklets (Aethia spp.). Genetics **171:** 1885-1894.

WALTARI, E., and S. V. EDWARDS, 2002 Evolutionary Dynamics of Intron Size, Genome Size, and Physiological Correlates in Archosaurs. The American Naturalist **160:** 539-552.

WANG, J., 2005 Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society of London Series B **360:** 1395-1409.

WAPLES, R. S., 2002 Definition and Estimation of Effective Population Size in the Conservation of Endangered Species, pp. 147-168 in *Population Viability Analysis*, edited by S. R. BEISSENGER and D. R. MCCULLOUGH. The University of Chicago Press.

WAPLES, R. S., 2005 Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? Molecular Ecology **14:** 3335-3352.

WHITE, D. J., J. N. WOLFF, M. PIERSON and N. J. GEMMELL, 2008 Revealing the hidden complexities of mtDNA inheritance. Molecular Ecology **17:** 4925-4942.

WILSON, A. C., R. L. CANN, S. M. CARR, M. GEORGE, U. B. GYLLENSTEN *et al.*, 1985 Mitochondrial DNA and two perspectives on evolutionary genetics. Biological Journal of the Linnean Society **26:** 375-400.

WILSON, P. R., D. G. AINLEY, N. NUR, S. S. JACOBS, K. J. BARTON *et al.*, 2001 Adélie penguin population change in the pacific sector of Antarctica: relation to sea-ice extent and the Antarctic Circumpolar Current. Marine Ecology Progress Series **213:** 301-309.

WOEHLER, E. J., 1993 The distribution and abundance of Antarctic and subantarctic penguins. , pp. Cambridge: Scott Polar Research Institute (SCAR).

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97-165.

YEUNG, C. K.-L., C.-T. YAO, Y.-C. HSU, J.-P. WANG and S.-H. LI, 2006 Assessment of the historical population size of an endangered bird, the black-faced spoonbill (*Platalea minor*) by analysis of mitochondrial DNA diversity. Animal Conservation **9:** 1-10.

ZHANG, D.-X., and G. M. HEWITT, 2003 Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Molecular Ecology **12:** 563-584.

2 Chapter Two

GENETIC DIVERSITY AND EFFECTIVE POPULATION SIZE OF ADÉLIE PENGUIN COLONIES

2.1 Abstract

Neutral genetic diversity is generally thought to vary with the size of a population - larger populations being characterised by high levels of genetic variation and smaller populations by lower levels of variation. However, whether mitochondrial genetic variation can represent a proxy for population size has been questioned. Most studies to date have compared different species, rather than different populations within the same species. Our aim was to investigate the relationship between mitochondrial DNA diversity and population size within a single vertebrate species in which colony sizes are known to vary between 1,700 and more than 200,000 breeding pairs. Using 528 mitochondrial control region sequences of individuals from 15 colonies of Adélie penguins from the Ross Sea, Antarctica, we detected significant genetic structure among colonies. We also recorded a positive correlation between mitochondrial control region diversity (θ) within colonies and the actual number of breeding pairs (N_b). When we used θ to estimate effective population sizes, the results suggest that population sizes have not been constant over time. This is consistent with previous evidence that Adélie penguin populations have expanded into Antarctica since the late Pleistocene. Our results suggest that mitochondrial diversity can be used to determine changes in population size over time. Hence, this paper provides support for many studies that have used mitochondrial diversity to infer population history, at least in the case of populations of the approximate size studied here.

2.2 Introduction

The popularity of the remarkable mitochondrial DNA molecule derives from a number of its well-known properties. Despite the fact that it is typically inherited maternally, does not recombine (or at best at very low levels), and some regions at least evolve in a neutral or nearly neutral manner, exceptions have been suggested in relation to each of these. In addition, the fact that mitochondrial DNA has a smaller effective population size compared with nuclear DNA, means it can be used to trace recent evolutionary history, such as founder events and genetic bottlenecks (HARRISON 1989). Mitochondrial sequences have also been widely used as genetic markers to track gene flow within and among species, to reconstruct phylogenetic relationships and to estimate historical population sizes. In relation to the latter, levels of neutral genetic variation are expected to correlate with population size within species as well as across different species (FRANKHAM 1996). More specifically, heterozygosity at neutral loci is theoretically expected to increase as the product of population size and the mutation rate increases (KIMURA 1983; OHTA 2003). Hence, measuring the amount of genetic variation present in a population has uses for the elucidation of population history.

Previous empirical studies have suggested that a positive linear relationship exists between the logarithm of population size and allozymic diversity (SOULÉ 1976), non-coding nuclear genetic diversity (FRANKHAM 1996) and mitochondrial diversity (AVISE 1992). It has been assumed that mitochondrial DNA diversity serves as a good proxy for population size, and can be used to estimate effective population size (N_e), which is the size of an "ideal population" i.e. one characterized by having randomly mating individuals, equal sex ratios, discrete non-overlapping generations, constant population size and random variation in reproductive success (FRANKHAM 1996; WRIGHT 1931).

However, a number of recent studies have suggested that mtDNA diversity does not reflect the size and history of a population and so should not be applied to population genetics studies (BALLARD and WHITLOCK 2004; BAZIN *et al*. 2006). Bazin *et al*. (2006) concluded that natural selection may be acting on mtDNA through genetic

draft, defined as the recurrent fixation of advantageous mutations which then leads to a frequent loss of variability at linked loci (GILLESPIE 2000). A consequence of this idea is that mtDNA diversity is thought to reflect the time and size of the last selective sweep event. It has been suggested that the effect of genetic draft on diversity increases with population size (GILLESPIE 2001). Bazin *et al*. (2006) arrived at this conclusion by comparing average diversity between many species of invertebrates and vertebrates, small and large organisms, and marine and terrestrial organisms. Nuclear DNA and allozyme diversity were also analyzed and found to be relatively greater in organisms with larger average population size, while mtDNA diversity failed to reflect this relationship (BAZIN *et al*. 2006). Following Bazin *et al*.'s (2006) study, it has since been suggested that the relationship between mitochondrial diversity and population size holds true within species with smaller populations, for example in humans (MULLIGAN *et al*. 2006). A significant correlation consistent with the hypothesis that mtDNA diversity is positively related to population size in animal groups with smaller populations was also recorded across forty-seven species of eutherian mammals (MULLIGAN *et al*. 2006). Moreover, many contemporary studies are of species with small populations, for example of interest in conservation biology, or groups of closely related species that have separated relatively recently. Positive, significant relationships between genetic diversity and catch sizes were also found for a meta-analysis on numerous species of fish (MCCUSKER and BENTZEN 2010).

We are in a position to test the relationship between diversity of the mitochondrial control region and population size within a single species due to the unique biology and ecology of Adélie penguins. Overall the number of breeding pairs of Adélie penguins is large (minimum 2.47 million). Nesting in this species occurs during the summer months and approximately 177 colonies are known around the Antarctic (WOEHLER and RIDDLE 1998). Colony sizes vary greatly and Adélie penguins have been shown to typically exhibit a high degree of natal philopatry (AINLEY 2002). Accurate numbers of breeding pairs for different colonies are known since the 1960s (WILSON *et al*. 2001) and there are direct estimates for the evolutionary (LAMBERT *et al*. 2002) and mutation rates (MILLAR *et al*. 2008) of the mitochondrial control region available from ancient DNA and pedigree approaches, as well as a large number of mitochondrial control region sequences (LAMBERT *et al*. 2002) from many modern

populations. Having an accurate estimate of a mutation rate is essential to this study, as estimates of N_{ef} from genetic data are greatly affected by the magnitude of this parameter.

The specific objective of this study was to examine the relationship between genetic diversity of the mitochondrial control region, a neutral region of the mitochondrial genome (Millar *et al*., 2008), and population size, using the Adélie penguin as a model species. To accomplish this we compared mitochondrial genetic diversity with data from actual breeding population ground and aerial counts. In order for the Adélie penguin to be amenable to this test, colonies need to be significantly differentiated from each other and in mutation-drift equilibrium.

Since Bazin *et al*.'s (2006) paper dealt with this issue at a larger scale, and due to the uncertainties raised by recent studies (BAZIN *et al*. 2006; GERBER *et al*. 2001; GILLESPIE 2000; GILLESPIE 2001; HURST and JIGGINS 2005), an analysis of diversity at the intra specific level appears necessary. Studies of relevance to conservation or understanding population history frequently rely on genetic estimates of long-term effective population sizes. While the recent controversy has highlighted some of the issues related to the use of mitochondrial DNA, dismissing its use entirely without investigating its applicability at the species level seems unwarranted, particularly given how useful it has been as a marker in the past and could continue to be in the future.

2.3 Material and Methods

2.3.1 Annual count of Adélie penguin breeding pairs

At most times, Adélie penguins exhibit strong natal philopatry to their breeding colonies (Ainley, 2002). Therefore during this study we have assumed that a colony is essentially a population, generally receiving or donating few migrants to neighbouring colonies. The numbers of breeding pairs of Adélie penguins in each year were

obtained by direct population counts of summer colonies using ground counts up to 1981 and aerial photographs from 1982 (see Figure 2.1 for an example aerial photograph of the colony at Northern Cape Bird), encompassing 40 years of counts (WILSON *et al*. 2001). These photographs are taken annually at the same time in summer when it is known that only a single member of the breeding pair will be on the nest (the other being at sea feeding). All non-breeding individuals (juveniles and post-breeders who typically do not even come ashore) are absent from the count. Reproductive variance among Adélie penguins is low; individuals almost universally lay only 2 eggs. We have taken into account population size fluctuations by calculating a harmonic mean of breeding pair number counts, giving us a good approximation of recent effective population size (FRANKHAM *et al*. 2005). The harmonic mean ($N_b = t / \Sigma(1/X_i)$) of effective population sizes in different generations allows the impact of population size fluctuations on the overall effective population size to be taken into account (FRANKHAM *et al*. 2005). We consider the impact of gene flow on this estimate of effective sizes for each colony to be low, as studies have shown that migrants make up fewer than 1% of breeding individuals. This does sporadically increase to ~3.25% in times of great environmental stress (DUGGER *et al*. 2010). For these reasons, rather than a census in the normal sense of the word, these data are direct counts of the size of the breeding populations and therefore comparable to the genetic effective size. In this study recent effective population size approximated from recent breeding pair counts is compared to long-term female effective population size as estimated from mitochondrial control region diversity.

2.3.2 Samples, DNA extraction, PCR amplification and sequencing

A total of 528 mitochondrial control region sequences of Adélie penguins were included in the present study (RITCHIE *et al*. 2004). These samples were collected from Adélie penguin colonies around Antarctica, principally from the Ross Sea area (Fig. 2.1). Sequences can be obtained at GenBank (accession numbers AF474792 to AF474997, AY525167 to AY525326) (Lambert *et al* 2002, Ritchie *et al* 2004). New sequences can be accessed via GenBank (to be submitted: 27 sequences from Duke of York Island).

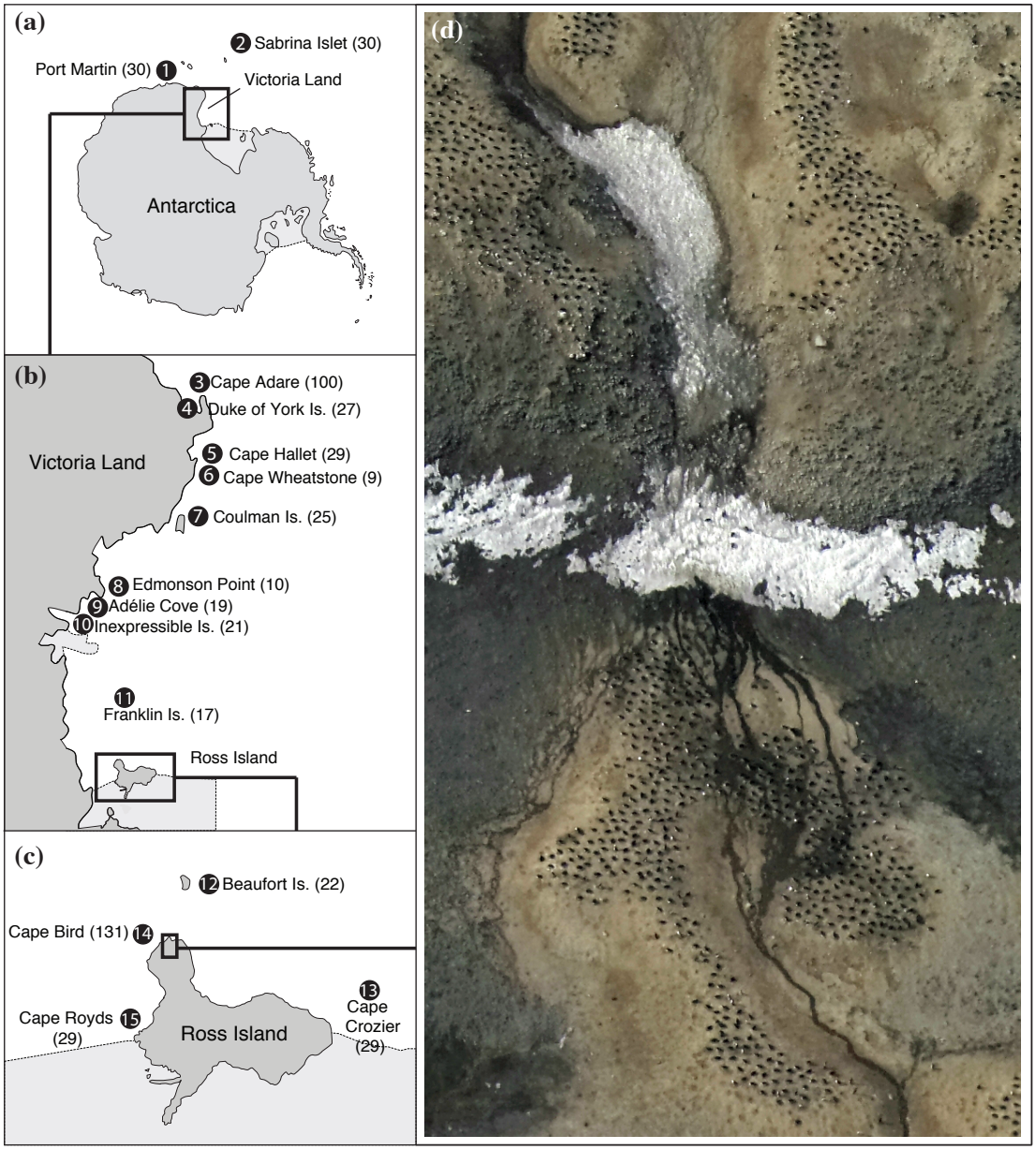


Figure 2.1: The distribution of Ross Sea Adélie penguin colonies (1 – 15) from which samples were used in this study, along with a photograph (courtesy of K. Barton) used for counting number of breeding pairs from the colony at Cape Bird.

Number of sequences used for analysis indicated in parenthesis next to colony name. Each black dot is a single penguin representing a breeding pair. The numbers refer to the colonies listed in Table 2.2.

Total genomic DNA was isolated from blood preserved in buffer (SEUTIN *et al*. 1991) using a standard phenol-chloroform extraction method (SAMBROOK *et al*. 1989). Twenty-seven mitochondrial control region DNA sequences were obtained for this study from samples collected from Duke of York Island in the Ross Sea (170,15; -

71,3) (2001-2002 field season). Five hundred and sixty base pairs of the mitochondrial 5' control region were amplified using PCR primers developed previously (Millar *et al* 2008): L-tRNA-Glu special 5'-CGCTTGGCTTYTCTCCAAGGTCTA-3' and AH 530 special 5'-GCTGATTTCACGTGAGGAGACCG-3'. Primers were heated and snap chilled prior to use to avoid primer dimer formation and eliminate secondary structure. Amplification of DNA extracts was performed in 25 μl reactions containing 60 ng of template DNA, 1x PCR Buffer, 2 mM $MgCl_2$, 200 μM dNTPs, 0.2 μM of each primer, 0.5 u of Platinum® *Taq* polymerase (Invitrogen), and sterile water.

Samples were amplified using a GeneAmp® PCR System 9700 thermocycler at 94ºC for 1 m, 55ºC for 30 s, and 72ºC for 30 s, for 32 cycles following an initial denaturation at 94ºC for 2 m. Products were purified using Agencourt® AMPure® magnetic beads, and sequenced on an ABI 3730 automated sequencer. Sequencing reactions contained 3.3μM AH530 special primer and 8-12 ng of product. Sequences were visualised in 4Peaks (Version 1.7.2, http://mekentosj.com/science/4peaks) and edited in Sequencher 4.6 (Genecodes, Ann Arbor, MI).

2.3.3 Summary Statistics and Population structure (AMOVA)

Generation of summary statistics, tests for geographic associations among HVRI sequences using AMOVA were conducted in Arlequin 3.11 and 3.5 (EXCOFFIER *et al*. 2005; EXCOFFIER and LISCHER 2010). Tajima's *D* (TAJIMA 1989) and Fu's *Fs* (FU 1997) statistics were estimated to determine whether the observed polymorphism fit with neutral model expectations. We used JModelTest 0.1.1 (POSADA 2008) to ascertain which substitution model would best fit the data among those offered in Arlequin for AMOVA. The Kimura-2-Parameter model was chosen, with a gamma value of 18, and a ti:tv of 9. Colonies represent populations but were not grouped for the analysis. Significance of AMOVA was tested with 60 000 permutations. 100 000 Markov chain steps were taken for exact test of differentiation.

2.3.4 Genetic estimates of long-term effective population sizes of colonies of Adélie penguins in the Ross Sea

Coalescent estimates of effective population size (N_e) within colonies and pairwise migration M (the dispersal rate (m) divided by the mutation rate (μ)) among populations were obtained using a Bayesian coalescent approach as implemented in MIGRATE version 2.1.3 (BEERLI 2006; BEERLI and FELSENSTEIN 1999; BEERLI and FELSENSTEIN 2001). MIGRATE simultaneously estimates genetic diversity (θ), and M (BEERLI and FELSENSTEIN 1999). The Bayesian approach was chosen as the maximum likelihood approach can fail on single locus datasets or when data are sparse and deliver non-conservative support intervals, while a Bayesian approach remedies these problems given an appropriate prior (BEERLI 2006). In a simulation study comparing the Maximum Likelihood and Bayesian approaches, estimated population size correctly reflected simulated parameters, but values of M were low compared with the simulated values (BEERLI 2006).

The long-term effective population size for maternally inherited mitochondrial DNA ($N_{e(f)}$) is estimated from genetic diversity (θ) and the mutation rate per site per generation (μ): $N_{e(f)} = \theta/2\mu$. Genetic diversity of populations is affected by migration, fluctuations in population size, population structure and selection.

Bayesian analyses of colonies of Adélie penguins were initially conducted with default starting parameters of θ and M. Successive runs were conducted using θ (median) estimated from initial runs until convergence was achieved (acceptance ratios were above 15%, posterior probability distributions were unimodal and θ estimates were robust when starting parameters were changed). Settings were optimised until a reliable estimate was obtained from the data. The robustness of the data was assessed by changing the starting number seed and prior distribution, since a failed analysis or an uninformative dataset in Bayesian analysis is identifiable when the posterior probability distribution is similar to the prior. More than 35000 samples were discarded at the beginning of each chain. Adaptive chain heating was used with four different temperatures to ensure adequate chain mixing and searching of the parameter space. Once settings were optimised, final estimates were obtained from a

run of 10 replicates of 4 long chains (9000+ long sample, 35+ sampling increment). Adélie colonies were grouped according to putative meta-populations for this study (Ainley, 2002), to allow us to gauge the magnitude of maternal gene flow between nearby colonies. θ and M estimated together. The program MIGRATE performs best with sample sizes of 20-30 sequences, but for two colonies, Cape Bird and Cape Adare, we had upwards of 100 sequences available. As a result for those two colonies subsets of the sequences were taken and then averaged together.

A mutation rate μ of 0.55 (0.33-0.99) mutations per site per million years for the mitochondrial control region from a large-scale pedigree analysis of Adélie penguins is available (Millar *et al* 2008). This mutation rate was scaled by generation time (6.46 years, Ainley, 2002) and used to transform estimates of θ into long-term $N_{e(f)}$. We tested for correlations between colony estimates of θ and colony size in breeding pair numbers. Correlations were tested for significance using a 2-tailed t test.

2.4 Results

2.4.1 Summary Statistics and Population Structure

Haplotype diversities were consistently high among the colonies (> 0.9), while nucleotide diversity and mean number of pairwise differences varied somewhat, with the highest values pertaining to Cape Adare, the largest of the colonies. Most Tajima's *D* and all Fu's *Fs* values show significant negative departures from the assumed null hypothesis, consistent with population expansion.

The majority of variation resides within each of the colonies. This is common with species with high nucleotide diversity. Initially, the analysis was performed using groupings following the putative meta-populations suggested by Ainley (2002); however, no support was found in the structure analysis for these groupings so the analysis was carried out at the colony level (data not shown).

Table 2.1 Summary statistics for *HVRI* in fifteen Adélie penguin colonies

	PM	SI	DY	CA	CH	CW	CI	EP	AC	II	FI	BI	CB	CR	CC	Mean	s.d.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
n	29	29	18	90	29	9	25	10	19	21	17	22	130	29	29	33.73	32.56
bp	381	381	381	381	381	381	381	381	381	381	381	381	381	381	381	381	0
Pol. Loci	53	59	43	121	68	42	60	37	52	42	57	56	109	66	69	62.27	23.65
Ti	52	53	43	107	67	42	60	30	43	42	51	53	106	65	67	58.73	22.03
Tv	1	8	0	28	1	0	0	8	10	0	6	3	4	4	3	5.07	7.15
h	0.9975	0.9951	0.9804	0.9983	0.9975	1.0000	0.9900	1.0000	1.0000	0.9762	1.0000	0.9870	0.9979	0.9926	0.9877	0.9933	0.008
h s.d.	0.0099	0.0106	0.0243	0.002	0.0099	0.0524	0.0142	0.0447	0.0171	0.023	0.0202	0.0201	0.0013	0.0111	0.0133	0.0183	0.014
N. div.	0.0242	0.0374	0.0234	0.0713	0.0487	0.0470	0.0360	0.0590	0.0748	0.0319	0.0550	0.0436	0.0398	0.0417	0.0482	0.0455	0.015
N. div. s.d.	0.0128	0.0193	0.0127	0.0349	0.0247	0.0262	0.0187	0.0322	0.0386	0.0171	0.0287	0.0226	0.0198	0.0213	0.0245	0.0236	0.008
Θ_π	9.0049	13.9261	8.5229	26.5079	18.1478	17.500	13.383	21.956	21.006	8.9905	19.368	16.229	14.7234	15.387	17.835	16.166	5.066
Θ_π s.d.	4.7528	7.1630	4.6233	12.9904	9.2286	9.7569	6.9392	11.977	10.844	4.8149	10.098	8.3912	7.3425	7.8777	9.0756	8.3916	2.552
D	-1.25	-0.29	-1.30	0.37	0.17	0.67	-0.61	3.29	1.68	-0.91	0.62	0.22	-0.86	-0.32	0.04	0.10	1.19
Fs	-21.14	-12.68	-4.70	-23.98	-12.38	-1.60	-7.39	-1.62	-1.54	-5.29	-5.59	-5.62	-24.03	-9.48	-6.49	9.57	7.78

N = number of samples, *bp* = length of sequence in base pairs, Pol. Loci = polymorphic loci, *Ti* = transitions, *Tv* = transversions, *h* = haplotypic diversity, N. div. = nucleotide diversity, Θ_π = pairwise differences, *D* = Tajima's *D*, *Fs* = Fu's *Fs*, s.d. = standard deviation. Tajima's *D* and Fu's *Fs* are significant to $p<0.01$.

Significant structure was detected between the colonies (ϕ_{st} = 0.10515, p<0.01; exact test of differentiation, p<0.05) (Table 2.2). This structure is due to pairwise differentiation between some, but not all of the colonies. Pairwise values show patchy significant differentiation, most of which appears to be due to comparisons with colonies Port Martin, Sabrina Islet, Cape Adare, Cape Wheatstone and, to a lesser extent, Duke of York Island (Table 2.3). The southernmost colonies (from Edmonson Point south, see Fig. 2.1) are not significantly different from each other for the most part (Table 2.2).

Table 2.2: AMOVA results.

Scenario	Within-population variance	Among-population, within-group variance	Among-group variance	Exact test of differentiation
Population = colony	89.49%	10.51% ϕ_{st} = 0.10515, p<0.01	n/a	p<0.05

2.4.2 Estimates of genetic diversity (θ) and female effective population size (N_{ef})

Estimates of genetic diversity and 97.5% credibility intervals for these estimates for colonies were obtained using the Bayesian approach in MIGRATE (Table 2.2). A significant and positive linear correlation was recorded for genetic diversity and colony sizes (Fig. 2.2) (p<0.05). Colony size explains 40% of variation in colony diversity at the mitochondrial control region (r=0.63). We also estimated the parameter M (=m/μ), quantifying the amount of migration occurring between colonies grouped in putative meta-populations. Estimates of this parameter were not as robust as genetic diversity estimates but overall m=Mμ values were of the order of 0.001 proportion of migrants received into each colony in a meta-population per generation (data not shown).

Estimates of female effective population size (N_{ef}) were obtained from genetic diversity (θ) estimates using a known mutation (μ) rate: θ/2μ. These rates were rescaled for generation time (6.46 years in Adélie penguins). Genetic estimates were compared to

Table 2.3: Pairwise ϕ_{st} results.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PM	*	-	+	+	+	+	+	-	+	+	+	+	+	+	+
SI	-0.00612	*	+	+	+	+	+	+	+	+	+	+	+	+	+
DY	**0.03766**	**0.0485**	*	+	+	+	-	-	-	-	-	-	-	-	-
CA	**0.30788**	**0.31038**	**0.23713**	*	+	-	+	+	-	+	+	+	+	+	+
CH	**0.17676**	**0.17907**	**0.07672**	**0.05674**	*	-	-	-	-	-	-	-	-	+	-
CW	**0.51377**	**0.50301**	**0.39109**	-0.02102	0.07925	*	+	+	-	+	+	+	+	+	+
CI	**0.07343**	**0.06725**	0.00057	**0.18293**	0.0281	**0.2628**	*	-	-	-	-	-	-	-	-
EP	0.0298	**0.03829**	-0.00737	**0.26198**	0.10643	**0.39964**	0.01902	*	-	-	-	-	-	-	-
AC	**0.14714**	**0.14805**	0.0388	0.04492	-0.04972	0.08926	-0.00486	0.06965	*	-	-	-	-	-	-
II	**0.06451**	**0.06197**	-0.00236	**0.16123**	0.02268	**0.25195**	-0.01024	0.0257	-0.0216	*	-	-	-	-	-
FI	**0.07445**	**0.07769**	-0.00942	**0.15275**	0.00416	**0.23023**	-0.01853	0.01902	-0.03459	-0.02794	*	-	-	-	-
BI	**0.08759**	**0.08874**	-0.00216	**0.16007**	0.01232	**0.24122**	-0.01504	0.02672	-0.02811	-0.01682	-0.02154	*	-	-	-
CB	**0.06727**	**0.07101**	0.01506	**0.16606**	0.01779	**0.23363**	-0.00309	0.03435	-0.02511	-0.01093	-0.01729	-0.01113	*	-	-
CR	**0.04578**	**0.05145**	-0.01899	**0.22759**	**0.06638**	**0.34213**	-0.00342	-0.00967	0.03142	0.00277	-0.00791	-0.00803	0.01053	*	-
CC	**0.12399**	**0.13002**	0.0512	**0.09035**	-0.01025	**0.12688**	0.01815	0.07417	-0.04824	-0.00128	-0.00417	0.00424	0.00179	0.04411	*

Colonies are labeled as number (top row) or abbreviation of first letters (left column), e.g. 1 = PM = Port Martin. Lower diagonal indicates pairwise ϕ_{st} values, upper diagonal indicates significanc, $p<0.05$ = +, $p>0.05$ = -. Significant values are labelled in bold and shaded in grey.

Table 2.4: Estimates of population sizes from population counts together with genetic diversity in colonies of Adélie penguins.

Colonies are arranged by increasing number of breeding pairs.

Population		Parameters		
		θ (97.5 percentile)	N_b ± SD	N_{ef} (97.5 percentile)
8	Edmonson Point	0.04101 (0.00604-0.08298)	1793 ± 457	5768 (850 – 11671)
15	Cape Royds	0.12386 (0.05479-0.21095)	1892 ± 1000	17564 (7617 – 30138)
6	Cape Wheatstone	0.06593 (0.00749-0.17383)	2315 ± 514	9273 (1053 – 24449)
4	Duke of York	0.08225 (0.03892-0.11725)	3442 ± 1303	11568 (5474 – 16491)
2	Sabrina Islet	0.09948 (0.05892-0.14755)	7073	13992 (8287 – 20752)
9	Adélie Cove	0.05178 (0.01343-0.09012)	10123 ± 1509	7283 (1889 – 12675)
10	Inexpressible Island	0.05553 (0.02052-0.09754)	22897 ± 3420	7810 (2886 – 13719)
7	Coulman Island	0.07655 (0.03185-0.11131)	25453 ± 4140	10767 (4480 – 15655)
12	Beaufort Island	0.05887 (0.01986-0.11588)	36288 ± 8732	8028 (2779 – 15337)
14	Cape Bird	0.15196 (0.06508-0.27179)	36676 ± 9081	24485 (10625 – 43777)
5	Cape Hallett	0.08974 (0.04455-0.13041)	48729 ± 8356	12622 (6266 – 18342)
11	Franklin Island	0.07423 (0.02863-0.11983)	54877 ± 8960	10440 (4027 – 16854)
1	Port Martin	0.12982 (0.07011-0.18456)	55860	18259 (9861 – 25958)
13	Cape Crozier	0.11808 (0.05239-0.19519)	132928 ± 26167	17264 (7566 – 28353)
3	Cape Adare	0.18055 (0.09551-0.27892)	223291 ± 43564	25394 (13433 – 39229)

θ is the mitochondrial genetic diversity as estimated using the Bayesian coalescent approach in MIGRATE. N_b is the harmonic mean of the number of breeding pairs in each colony estimated by direct count over a number of years. N_{ef} is the effective population size of breeding females based on mitochondrial control region sequence variation for each colony.

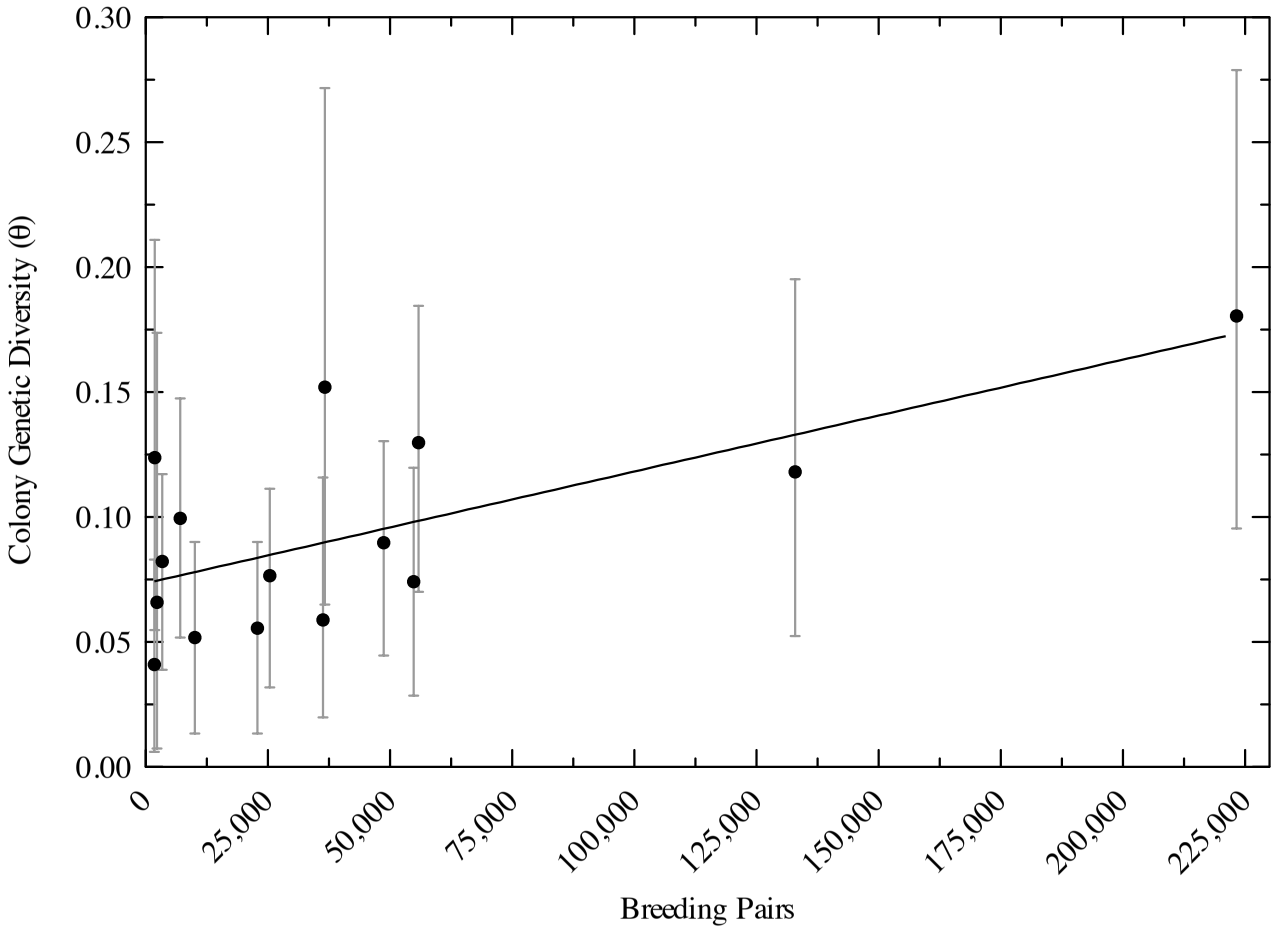


Figure 2.2: The relationship between colony size (harmonic mean of breeding pair counts; N_b on the X axis) and genetic diversity (as estimated in MIGRATE on the Y axis).

A positive correlation was found (r = 0.63) that was significant to $p<0.05$ (2-tailed t test).

demographic estimates obtained by calculating the harmonic mean of breeding pair numbers (Table 2).

Estimates of female effective population sizes (N_{ef}) of colonies and demographic harmonic means of breeding pairs per colony (N_b) were significantly and positively correlated ($r = 0.63$, $p<0.05$ 2-tailed).

2.5 Discussion

With this study, we set out to test Bazin *et al*'s (2006) assertion that mitochondrial DNA diversity does not increase with increasing population size, within a single species. In order to test this question adequately, several criteria need to be met that are often lacking in non-model organisms. Population sizes need to vary, there should be detectable structure between them, and a good measure of population size available. Ideally, the populations should be closed to gene flow from other populations and have stable sizes through time or a good understanding of the population history of the species.

These conditions are rarely met. Adélie penguins do not perfectly fit these criteria; there is evidence for low levels of episodic gene flow, and populations may have not been stable over time. Despite these drawbacks, Adélie penguins fulfil more of the conditions mentioned above much more closely than most other species of vertebrate. Colony sizes vary greatly, as shown through breeding pair counts (Table 2.2). The measure of colony size used here, namely the harmonic mean of the number of breeding pairs from direct counts since the 1960s, is arguably closer to a current effective size for these colonies than many estimates present in the literature, certainly those proceeding from indirect estimation methods. As for population structure, it has long been known that Adélie penguins normally exhibit a high degree of natal philopatry, returning to their colony of birth when the time comes to breed. Hence a proportionate degree of structure might be expected from genetic studies. Recent studies, however, point towards episodic movement between colonies at times of environmental stress (DUGGER *et al*. 2010; SHEPHERD *et al*. 2005). Also, a study

using nuclear microsatellite markers failed to recover any structure at the colony level (ROEDER *et al*. 2001).

Our AMOVA results, testing for population genetic structure among colonies, show the highest amount of variation exists within each of the colonies. This in and of itself is not surprising, and is common with populations presenting such high nucleotide diversity. Differentiation measures between the colonies shows a significant ϕ_{sc} value, 0.10515, indicating low to moderate differentiation (BALLOUX and LUGON-MOULIN 2002; WRIGHT 1978). This apparently low differentiation however, is consistent with markers with high variability having a low maximum ϕ_{st} (BALLOUX and LUGON-MOULIN 2002; HEDRICK 2005; MEIRMANS 2006). The effect of mutation-caused polymorphism deflates the highest expectations of F_{st} (BALLOUX and LUGON-MOULIN 2002). The exact test of differentiation also shows significant structure, to $p<0.05$. When variability within and among populations is high, as is the case with Adélie penguin colonies (RITCHIE *et al*. 2004), significance will be difficult to reach (Excoffier, 2008; comment in Genetic Software Forum). Hence, this further supports the likelihood of differentiation between the colonies. Therefore, overall we believe there to be significant, low to moderate differentiation between the colonies but no support for the meta-population grouping proposed by Ainley (2002), despite having sufficient power to detect significant structure at the group level (FITZPATRICK 2009) (data not shown).

However, pairwise estimates of differentiation indicate that a number of differentiated colonies are potentially driving the finding of significant differentiation between the colonies. In particular, the colonies distributed farther north in the Ross Sea presented greater differentiation when compared to other colonies, while the southernmost colonies appeared genetically indistinguishable to other colonies. Previous work carried out found that two mitochondrial lineages were distributed unevenly throughout the Ross Sea, forming a cline. It is possible the structure we detect has arisen from that cline. For the purposes of this chapter we follow the assumption that the observed differentiation is enough to warrant testing the relationship between population size and mitochondrial diversity, with the caveat that this test is invalidated if there is no structure.

Having established that Adélie penguins might serve as a good model to test the relationship between mitochondrial diversity and population size, our analyses yielded consistent and robust genetic diversity estimates for each colony. In addition, a significant correlation was recorded between genetic diversity within colonies and colony size as measured by the harmonic mean of number of breeding individuals (Fig. 2.2). Our estimates presented wide confidence intervals, however, since obtaining a statistically meaningful estimate of N_{ef} when dealing with large population sizes can be difficult (TURNER *et al*. 1999), resulting in wide credibility or confidence intervals (LEHMANN *et al*. 1998). The Bayesian inference method as implemented in MIGRATE also has a tendency to produce conservative confidence intervals (Beerli, 2006). Theoretically, genetic diversity and effective population size should present a slope of unity in this correlation. The relationship we have observed has a flatter slope, such that a ten-fold increase in population size is met with a ≈ three-fold increase in genetic diversity. This indicates our estimate of recent breeding pair counts is not a perfect approximation of coalescent effective population size (represented by genetic diversity); however, the relationship is still significant and positive indicating a signal is still present in the data.

The significant correlation between colony size and mitochondrial genetic diversity found suggests that mtDNA should be a good predictor of population history in the Adélie penguin, so we calculated the female effective population size (N_{ef}) of Adélie penguin colonies and compared these to the harmonic mean of breeding pair counts for the different colonies. Our estimates of N_{ef}, though correlating positively to current sizes, are generally much lower than these. This was found to be the case except for colonies smaller than ~22,000 breeding individuals; for these, N_{ef} estimates were higher than the harmonic mean of the number of breeding individuals. The exception was Cape Bird on Ross Island with ~36,000 breeding individuals, the estimated colony size from genetic data was not significantly different. For the smallest colonies, it is likely that direct counts of breeding pairs do not approximate the effective size as well as for larger colonies; the smaller colonies are more likely to donate immigrants during times of environmental instability, and receive them once conditions normalize (Dugger *et al*, 2010). We sought to investigate whether the

nearest large colony (five times larger or more) influenced this "excess" population size in the smallest colonies. A correlation was performed for distance to nearest large colony against excess population size (N_b - N_{ef}). No relationship was found (data not shown). The sample size (n=7 small-large colony pairs) for this analysis is too small to robustly test this possibility at present. Also, geographic distance between colonies may not be the best indicator of potentially increased gene flow, rather distance to shared feeding grounds may be a better indicator. It is also worth noting that our estimates of N_{ef} depend greatly on the quality of the mutation rate estimate and an accurate generation time. We have confidence in both of these estimates, however we cannot rule out some uncertainty. A slower rate would have resulted in smaller estimated sizes, conversely a faster rate would give us larger sizes. Generation time changes would also have a noticeable effect.

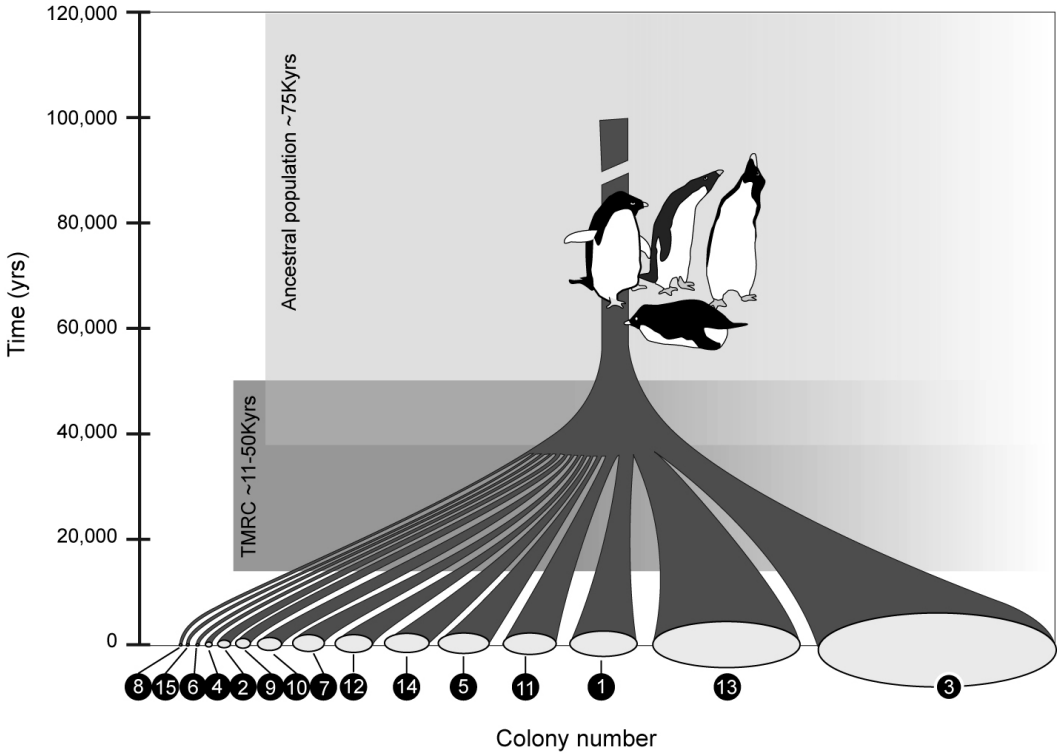


Figure 2.2: Schematic representation of the late Pleistocene migration of Adélie penguins into the Ross Sea of Antarctica, together with estimation times for the common ancestor of the current populations.

N_{ef} values actually estimate populations at points in the past. Current and long-term estimates of population size are similar in populations without much loss of genetic

diversity through inbreeding (WHITLOCK and BARTON 1997). It is apparent that for the Adélie penguin, population sizes have not been constant over time. In this species there is good evidence for a genetic bottleneck occurring during the last glacial maximum. Ritchie *et al* (2004) observed that Adélie penguins had the signature of an expanding population. Based on an evolutionary rate for the HVRI in Adélie penguins calculated from ancient DNA (0.96 s/s/Myr; Lambert *et al* 2002), these authors (Ritchie *et al*, 2004) suggested that current populations derived from an ancestral one ~75kya (95% HPD interval 37 – 122 kya) (Fig. 2.3). In a purely panmictic population the time to the most recent common ancestor (TMRCA) equals $2N_{ef}$ (Hudson 1990). Using N_{ef} (generated using the mutation rate from Millar *et al*, (2008) (0.55s/s/Myr) to estimate coalescence time ($2N_{ef}$ generations), the TMRCA is ~50 – 323kya. This is not significantly different from the age of the most recent common ancestor proposed by Ritchie *et al* (2004). Adélie penguins currently present a wide range of colony sizes (≈1,800 – 200,000 N_b). Mitochondrial evidence suggests that past colony sizes did not vary as much as current sizes (≈5,700 – 25,000 $N_{e(f)}$) and have, in most cases, increased in size since the Pleistocene, with rising temperatures and changes in coastal ice cover.

Bazin *et al* (2006) have suggested that mitochondrial genetic diversity is not a reliable indicator of population size. This work was based on a large-scale analysis of major groups, e.g. vertebrates versus invertebrates. Bazin *et al* (2006) suggested that this lack of a relationship was explained by reference to selective sweeps (see the commentary by (EYRE-WALKER 2006)). That is, the increase in frequency of advantageous variants in the mitochondrial genome incidentally also increases the frequency of alleles at linked loci. Without recombination, the process reduces genetic variation. Hence, since different modern populations of differing sizes will have been the subject of such selective sweeps at different points in the past, there will be no consistent relationship between population size and genetic diversity. Rather the population sizes will simply reflect the timing of these unknown events at different points in the past. We find a positive and significant relationship between colony size and mitochondrial control region diversity, suggesting that in Adélie penguin colonies this is not the case. Though we cannot rule out a selective sweep or hitchhiking for the larger colonies – or Hill-Robertson effect – lowering genetic

diversity, these interpretations of the data seem unlikely. In order to test definitively whether these mechanisms are involved or not, an independent estimate of past population sizes is needed, however past population dynamics are the most plausible explanation for the size differences. Some studies found that differences in mutation rates between species were a better predictor of differences in mitochondrial diversity than population size indicators (NABHOLZ *et al*. 2009; NABHOLZ *et al*. 2008). Within a species it is highly unlikely that different populations would show variations in mutation rate and therefore we can eliminate such a phenomenon in Adélie penguins. Several studies following on from Bazin *et al*'s (2006) paper have found that mitochondrial and nuclear diversity do correlate positively with increasing population size between mammal and bird species (Nabholz *et al* 2008, 2009), and between fish species (McCusker & Bentzen, 2010). (ATKINSON *et al*. 2008) found that estimates of relative population size through time (derived from mitochondrial diversity) showed concordance to broad modern human groups. To our knowledge, however, this is the first study to use populations of a modern species, differing by orders of magnitude, in order to test this important concept.

In the case of Adélie penguins, the well-documented population bottleneck in the late Pleistocene is consistent with the estimates of the effective population sizes based on current genetic diversity within colonies. Hence, rather than invoking selective sweeps, an understanding of the past population dynamics explains the results obtained. Our results illustrate the strength of combined methods, in which demographic data and genetic data are utilised. As (RUBINOFF and HOLLAND 2005) point out in their review of the use of mtDNA in phylogenetic studies, the solution to the observed drawbacks of mtDNA is not to discontinue its use altogether. Using multiple molecular data sources, or combining genetic data with demographic, ecological data, allows us to test apparent incongruencies, thus enhancing our understanding of evolutionary processes (RUBINOFF and HOLLAND 2005).

Finally, our data provide support for the hypothesis, based on the neutral theory of evolution, that mitochondrial diversity increases with increasing colony size in Adélie penguin colonies, and therefore that the use of mtDNA for interpreting population history in vertebrates is not misplaced. Our estimates of N_{ef} appear to be indicators of

the sizes of ancestral populations of Adélie penguins from the late Pleistocene that gave rise to the modern populations.

2.6 References

AINLEY, D. G., 2002 *The Adélie Penguin: Bellwether of Climate Change*. Columbia University Press.

ATKINSON, Q. D., R. D. GRAY and A. J. DRUMMOND, 2008 mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory. Molecular Biology and Evolution **25:** 468-474.

AVISE, J. C., 1992 Molecular Population Structure and the Biogeographic History of a Regional Fauna: A Case History with Lessons for Conservation Biology. Oikos **63:** 62-76.

BALLARD, J. W. O., and M. C. WHITLOCK, 2004 The incomplete natural history of mitochondria. Molecular Ecology **13:** 729-744.

BALLOUX, F., and N. LUGON-MOULIN, 2002 The estimation of population differentiation with microsatellite markers. Molecular Ecology **11:** 155-165.

BAZIN, E., S. GLÉMIN and N. GALTIER, 2006 Population size does not influence mitochondrial genetic diversity in animals. Science **312:** 570-571.

BEERLI, P., 2006 Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics **22:** 341-345.

BEERLI, P., and J. FELSENSTEIN, 1999 Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations Using a Coalescent Approach. Genetics **152:** 763-774.

BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. PNAS **98:** 4563-4568.

DUGGER, K. M., D. G. AINLEY, P. O. B. LYVER, K. BARTON and G. BALLARD, 2010 Survival differences and the effect of environmental instability on breeding dispersal in an Adélie penguin meta-population. Proceedings of the National Academy of Sciences **107:** 12375-12380.

EXCOFFIER, L., G. LAVAL and S. SCHNEIDER, 2005 Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics **1:** 47-50.

EXCOFFIER, L., and H. E. L. LISCHER, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10:** 564-567.

EYRE-WALKER, A., 2006 Size Does Not Matter for Mitochondrial DNA. Science **312:** 537-538.

FITZPATRICK, B. M., 2009 Power and sample size for nested analysis of molecular variance. Molecular Ecology **18:** 3961-3966.

FRANKHAM, R., 1996 Relationship of Genetic Variation to Population Size in Wildlife. Conservation Biology **10:** 1500-1508.

FRANKHAM, R., J. D. BALLOU and D. A. BRISCOE, 2005 *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge.

FU, Y.-X., 1997 Statistical tests of neutrality against population growth, hitchhiking and background selection. Genetics **147:** 915 - 925.

GERBER, A. S., R. LOGGINS, S. KUMAR and T. E. DOWLING, 2001 Does Nonneutral Evolution Shape Observed Patterns of DNA Variation in Animal Mitochondrial Genomes? Annu. Rev. Genet. **35:** 539-568.

GILLESPIE, J. H., 2000 Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. Genetics **155:** 909-919.

GILLESPIE, J. H., 2001 Is the Population Size of a Species Relevant to its Evolution? Evolution **55:** 2161-2169.

HARRISON, R. G., 1989 Animal Mitochondrial DNA as a Genetic Marker in Population and Evolutionary Biology. Trends in Ecology and Evolution **4:** 6-11.

HEDRICK, P. W., 2005 A Standardized Genetic Differentiation Measure. Evolution **59:** 1633-1638.

HURST, G. D. D., and F. M. JIGGINS, 2005 Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society Series B **272:** 1525-1534.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of Evolution in Ancient DNA from Adélie Penguins. Science **295:** 2270-2273.

LEHMANN, T., W. A. HAWLEY, H. GREBERT and F. H. COLLINS, 1998 The Effective Population Size of Anopheles gambiae in Kenya: Implications for Population Structure. Molecular Biology and Evolution **15:** 264-276.

MCCUSKER, M. R., and P. BENTZEN, 2010 Positive relationships between genetic diversity and abundance in fishes. Molecular Ecology **19:** 4852-4862.

MEIRMANS, P. G., 2006 Using the AMOVA Framework to Estimate a Standardized Genetic Differentiation Measure. Evolution **60:** 2399-2402.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al.*, 2008 Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

MULLIGAN, C. J., A. KITCHEN and M. M. MIYAMOTO, 2006 Comment on "Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals". Science **314:** 1390a.

NABHOLZ, B., S. GLEMIN and N. GALTIER, 2009 The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. BMC Evolutionary Biology **9:** 54.

NABHOLZ, B., S. GLÉMIN and N. GALTIER, 2008 Strong Variations of Mitochondrial Mutation Rate across Mammals - the Longevity Hypothesis. Molecular Biology and Evolution **25:** 120-130.

OHTA, T., 2003 Origin of the neutral and nearly neutral theories of evolution. Journal of Biosciences **28:** 371-377.

POSADA, D., 2008 jModelTest: Phylogenetic Model Averaging. Molecular Biology and Evolution **25:** 1253-1256.

RITCHIE, P. A., C. D. MILLAR, G. C. GIBB, C. BARONI and D. M. LAMBERT, 2004 Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. Molecular Biology and Evolution **21:** 240-248.

ROEDER, A. D., R. K. MARSHALL, A. J. MITCHELSON, T. VISAGATHILAGAR, P. A. RITCHIE *et al*., 2001 Gene flow on the ice: genetic differentiation among Adélie penguin colonies around Antarctica. Molecular Ecology **10:** 1645-1656.

RUBINOFF, D., and B. S. HOLLAND, 2005 Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. Syst. Biol. **54:** 952-961.

SAMBROOK, J., E. F. FRITSCH and T. MANIATUS, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York.

SEUTIN, G., B. N. WHITE and P. T. BOAG, 1991 Preservation of avian blood and tissue samples for DNA analyses. Canadian Journal of Zoology **69:** 82-90.

SHEPHERD, L. D., C. D. MILLAR, G. BALLARD, D. G. AINLEY, P. R. WILSON *et al*., 2005 Microevolution and mega-icebergs in the Antarctic. PNAS **102:** 16717-16722.

SOULÉ, M. E., 1976 Allozyme variation, its determinants in space and time. , pp. 60-77 in *Molecular Evolution*, edited by F. J. AYALA. Sinauer Associates, Sunderland, Massachusetts.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585 - 595.

TURNER, T. F., L. R. RICHARDSON and J. R. GOLD, 1999 Temporal genetic variation of mitochondrial DNA and the female effective population size of red drum (*Sciaenops ocellatus*) in the northern Gulf of Mexico. Molecular Ecology **8:** 1223-1229.

WHITLOCK, M. C., and N. H. BARTON, 1997 The Effective Size of a Subdivided Population. Genetics **146:** 427-441.

WILSON, P. R., D. G. AINLEY, N. NUR, S. S. JACOBS, K. J. BARTON *et al*., 2001 Adélie penguin population change in the pacific sector of Antarctica: relation to sea-ice extent and the Antarctic Circumpolar Current. Marine Ecology Progress Series **213:** 301-309.

WOEHLER, E. J., and M. J. RIDDLE, 1998 Spatial relationships of Adélie penguin colonies: implications for assessing population changes from remote imagery. Antarctic Science **10:** 449-454.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97-165.

WRIGHT, S., 1978 Evolution and the Genetics of Populations in *Variability within and among Natural populations*. University of Chicago Press, Chicago.

3 Chapter Three

EFFECTIVE POPULATION SIZE OF THE EXTINCT HUIA

3.1 Abstract

The extinct Huia (*Heteralocha acutirostris*) was one of New Zealand's most emblematic, endemic bird species. Maori and human settlement is thought to have brought about the species' extinction through habitat loss, introduced mammalian predator species, and hunting. Within the context of a study on sexual bill dimorphism of Huia (LAMBERT *et al*. 2009), no evidence for nuclear genetic structure among Huia populations was found. This contradicts the prediction that Huia social organization and limited powers of flight indicated population differentiation. The census population size of Huia was estimated using mitochondrial hypervariable region sequences and likely mutation rate and generation times. The estimated census population size of Huia was found to be moderately high (34,187 – 89,539 individuals). A signature of population expansion was also detected. These results support the theory that the decline of Huia occurred after human settlement of New Zealand.

3.2 Introduction

The New Zealand Huia (*Heteralocha acutirostris*) belonged to one of New Zealand's only endemic bird families, Callaeatidae. The closest relatives of the Huia are the North Isand Saddleback *(Philesturnus carunculatus*) and the North Island Kokako (*Callaeas cinereus*). Like the Huia, these species are also wattlebirds (Callaeatidae) and are characterised by a pair of colourful, fleshy wattles; strong feet; and short, rounded wings (HEATHER and ROBERTSON 1998). The wattlebirds arrived in New Zealand most probably through transoceanic dispersal (SHEPHERD and LAMBERT 2007). The wattlebirds are a monophyletic family of passerines, nested within the oscines but excluded from the Passerida and core Corvoidea (SHEPHERD and LAMBERT 2007). Relationships amongst the wattlebirds are not resolved, though data suggest either Huia or Kokako may have diverged first (SHEPHERD and LAMBERT 2007).

The pre-human distribution of Huia bones in caves, dunes, and middens indicates that they were once common throughout the North Island of New Zealand but were absent from the South Island (WORTHY and HOLDAWAY 2002). They mainly inhabited montane and lowland hardwood-podocarp forests with a dense understorey (PHILLIPS 1963). Huia were thought to be a specialist predator of the Huhu beetle (*Prionoplus reticularis*) larvae, though they also ate other invertebrates and fruit (BULLER 1888).

Huia were medium sized passerines; one recently killed bird weighed 406g (BULLER 1888). Their colouration was mostly blackish with a greenish or dark bluish gloss, and their feathers had concealed light-grey bases (HIGGINS and PETER 2002). Tail feathers were blackish with broad white tips (*c*. 25mm long) (HIGGINS and PETER 2002). Their legs and feet were described by (BULLER 1870) as bluish grey, their wattles flesh-white, and their bills ivory-white. Huia had the most extreme sex-linked bill dimorphism known in birds (BURTON 1974; SELANDER 1966; SELANDER 1972). Male Huia were thought to have short, stout bills, whereas females were characterised by long curved bills, about a third longer than those of males. Males and females had such distinctive bill morphologies that they were originally described as different species (Fig. 3.1) (GOULD 1837; LACK 1971). It was suggested that the different bill morphologies related to different feeding strategies. Male Huia, with strong, shorter

bills, would break up rotten wood by chiselling and gaping. Female Huia used their longer, more slender bills to probe holes and crevices (BULLER 1870; BURTON 1974; SELANDER 1966). Lambert *et al* (2009) found that female bill size varied, suggesting juvenile female Huia had similar bill morphologies to males. Observations by early naturalists suggested that the species was territorial, that juveniles lived with adults, and that family members cooperated in foraging (BULLER 1888).

Figure 3.1: Extreme reverse sexual bill dimorphism in Huia. Adapted from Lambert et al (2009).

Following Polynesian settlement, Huia, Saddleback and Kokako declined in numbers. All three species were adversely affected by habitat loss and fragmentation, introduced mammals, and to some extent, hunting (HIGGINS and PETER 2002). Saddleback and Kokako are currently globally threatened. In the case of Huia, now extinct, the species' range was initially contracted greatly after the arrival of Maori (HEATHER and ROBERTSON 1998). Maori considered Huia as sacred, and a variety of Huia remains (feathers, heads, bills, skins, etc) were highly valued and traded widely (PHILLIPS 1963). Further reduction ensued as hunting pressure increased, partly because Huia tail feathers became fashionable among Europeans, especially after the Duke of York (later King George V) wore one in his hatband. Huia bills were also commonly used as brooches and these became increasingly popular in the late nineteenth century. Increased hunting, clearance of lowland forest, and the introduction of predators finally led to the Huia's extinction (HEATHER and ROBERTSON 1998; PHILLIPS 1963). The last confirmed sighting was in 1907 (PHILLIPS 1963), however, evidence suggests that the species may have survived until the 1930s (PHILLIPS 1963; Lovis pers. comm.).

Little is known about the behaviour or social structure of Huia, apart from limited observations made by early naturalists. Buller (1888) observed that Huia inhabited thick forest and moved mainly on foot 'by a series of bounds or jumps'. Colenso

(1887) suggested that Huia were social birds, and Buller (1888) noted that they were almost always found in pairs and sometimes in groups of four or more. Huia were probably pair-bonded for life (COLENSO 1887). One mate of a captive pair became distressed and died within ten days of its mate being accidentally killed (BULLER 1888). Not many records exist that document breeding in Huia. It was thought that they raised one brood per season, and breeding occurred October to November (POTTS 1885). Clutch size has been described as between three to five eggs (DRUMMOND 1910). Potts (1885) observed Huia young accompanying what he assumed to be their parents for a considerable time after fledging. He gave an account of four juveniles, barely distinguishable from adults, still being fed by their parents. It was suggested (MOORHOUSE 1996) that Huia were highly territorial based on Buller's (1888) observation that pairs were attracted by imitations of their call, a trait also common to the Saddleback and Kokako (HEATHER and ROBERTSON 1998). Their social organisation and limited powers of flight suggests that Huia are likely to have exhibited a high level of population genetic structure.

Lambert *et al*. (2009) examined museum samples collected from both sexes and from a number of locations (Fig. 3.2) to understand the population genetic structure of Huia and the nature of their sexual dimorphism. They used several nuclear genotyping markers isolated from extant Saddleback to amplify ancient DNA from Huia. Using rigorous ancient DNA methodologies, they were able to determine the genotypes of a number of individuals unambiguously. To determine the relationship between bill morphology and sex in Huia, they used molecular sexing methods and correlated these results with morphometric measurements.

I utilized the data generated during the study to estimate the genetic diversity of the Huia as a whole from mitochondrial hypervariable control region sequences, which may serve as a proxy for the effective size of a population, to calculate the probable census size of the species prior to its decline.

Figure 3.2: Provenance of Huia samples used in this study. Precise sample locations are indicated by circles and approximate locations by stars. Other place names are for reference only. Reproduced with permission from Lambert *et al*. (2009).

3.3 Material and Methods

3.3.1 Samples

Twenty-four Huia footpad samples were provided by the Canterbury Museum, 13 from the Auckland Museum, seven from the Naturalis Museum (Leiden), 24 from the American Museum of Natural History, two from Macleay Museum, two from the Australian Museum, and one sample from a private collector (Table 3.1). Exact provenance data were not known for 46 of these samples; the collection locations of the eight samples used for microsatellite analysis, plus two labelled as 'possibly Pipiriki', are illustrated in Figure 3.2.

Table 3.1: Huia samples used in this study.

Museum Number	Location	Collector and Presentation Date	Morphological Sex	Molecular Sex (# times sexed)	Beak	
					L	D
AV1076	Wairarapa	Buller, 1891	Female	Female (2)	99.03	16.17
AV1078	Makuri	1892	Male	Male (1)	55	17
AV1079	Ngarara	Buller, 1891	Male	Male (1)	54.01	18.16
AV1081	-	-	Female	Female (1)	77.83	14.98
AV1082	Wairarapa	Buller, 1892	Male	Male (1)	50.09	18.61
AV1083	-	-	Female	Female (1)	51	11.9
AV1085	-	-	Male	Male (3)	50.48	17.3
AV1087	-	Moorhouse	Male	Male (3)	50.85	17.69
AV1126	-	-	Male	Male (2)	52.5	16.44
AV2244	-	Parker	Female	Female (1)	85.93	14.95
AV2245	-	Parker	Male	Male (1)	46.8	17.71
AV2283	-	Parker	Male	Male (1)	55.44	16.94
AV2727	-	Codmor	Female	Female (1)	67.35	11.9
AV2729	-	O'Connor	Male	Female (5)	56.16	15.27
AV2744	Wellington	-	Male	Male (3)	54.71	16.33
AV2745	Mangaroa Hill	Len Harris, 1885	Male	Male (1)	51.29	18.11
AV2746	Mangaroa Hill	Len Harris, 1885	Female	Female (3)	87.98	13
AV2747	Mangaroa Hill	Len Harris, 1885	Female (juvenile)	Female (2)	66.2	14.13
AV21283	-	P. Hall	Male	Male (1)	49.78	17.52
AV21289	-	-	Female	Female (2)	106.4	12.93
AV36838	-	F.Grimwood, 1870s Gifted by a North Island Maori Chief	Female	Female (3)	79.65	15.68
AV37493A	Possibly Pipiriki	Mrs F. Stewart	Male?	Female (2)	64	10.5
	Possibly Pipiriki	Mrs F. Stewart	Female	Female (4)	104	12.5
HBH	-	-	Female	Female (2)	66	14
AV1070	-	-	Female	Female (2)	52	11.7
LB4564	North Island	-	Male	Male (3)	62.7	16.3
LB4565	North Island	-	Male	Female (1)	68.0	15.5
LB4567	Ruahine Range	C.E. Clarke, 20 Aug 1931	Male	Male (1)	58.2	15.9
LB4568	Ruahine Range	C.E. Clarke, 20 Aug 1931	Female	Female (1)	99.2	16.5
LB4571	North Island	S.H. Mountford, 1941	Male	Male (1)	54.3	16.5
LB4572	North Island`	S.H. Mountford	Male	Male (1)	51.0	15.3
LB4573	North Island	S.H. Mountford	Female	Female (1)	64.2	15.8
LB4575	North Island	S.H. Mountford	Female	Female (1)	59.5	14.3
LB4576	North Island	C.A. Fleming	Female	Female (1)	80.3	11.7
LB4577	North Island	C.A. Fleming	Male	Male (1)	52.0	16.8

LB9213	-	J.A. Rentoul, 03 Dec 1969	Female	Female (1)	100.7	13.7
LB9215	-	-	Male	Male (1)	49.0	15.2
LB9217	-	-	Female	Female (3)	99.5	15.21
RMNH110.108	Rimutaka, Hills, Wellington	Travers, 1898	Male			
RMNH110.109	Rimutaka Hills, Wellington	Travers, 1898	Female			
AMNH669772	Wellington	1892	Male			
AMNH669774	Makuri Ranges		Male			
AMNH669775	Makuri Ranges		Male			

Museum numbers and presentation details if known and sex (morphological and molecular) are given. HBH = specimen obtained from Hastings Boys' High School; RMNH = Rijksmuseum van Natuurlijke Historie; AMNH = American Museum of Natural History, Beak length (L) and depth (D) are measured in millimetres. Reproduced with permission from Lambert *et al*. (2009).

3.3.2 Ancient DNA methods

Extraction of ancient Huia DNA was performed in a dedicated ancient DNA laboratory physically separated from where contemporary DNA and PCR products were handled. Decontamination was routinely carried out by UV-irradiation and sodium hypochlorite washes. Approximately two mm^2 of Huia footpad tissue was removed and cut into several pieces using a sterile razor blade. Huia DNA was extracted by incubating footpad fragments overnight at 50°C in 2.5 ml of extraction buffer (10mM Tris-HCl pH 8.0, 10mM NaCl, 10mM EDTA), 250 μl of 10% SDS, 15 μl of 200 mg/ml dithiothreitol (DTT), and 25 μl of 50 mg/ml Proteinase-K. Samples were then extracted with phenol followed by chloroform: isoamyl alcohol (24:1) and then concentrated by centrifugation through a Vivaspin-30 (Viva Science, U.K.) membrane. Negative extraction controls were included with every 6–12 sample extractions. All mitochondrial sequences were obtained by PCR as outlined below and sequenced in the forward and reverse direction from multiple independent amplifications. Huia *HVRI* sequences are deposited in the GenBank database with accession numbers GU176413-GU176433.

3.3.3 Huia mitochondrial hypervariable region sequences

A one hundred and ninety-nine bp of the Huia hypervariable region I (*HVRI*) region was amplified, purified, sequenced and edited from 21 individuals (Fig. 3.3) using primers huiaIF (5'-ATAAACCCAAGTGATCCTACCT) and huiaIIR (5'-TTGAGTAGCTCGGTTCTCGTGA).

3.3.4 Estimating genetic diversity and effective population size in Huia

The genetic diversity parameter (θ) was calculated using a Bayesian framework in LAMARC v2.1.2 (KUHNER 2006) that uses a coalescent approach to obtain a joint estimate of various population genetic parameters such as genetic diversity, growth, migration, and recombination rates. Prior analysis of genetic structure using microsatellite markers (LAMBERT *et al*. 2009) did not distinguish separate populations so we did not estimate migration rates. We also did not estimate recombination rates as the analysis was carried out on the presumably non-recombining mitochondrial control region.

We estimated θ from 21 individuals for a 199 bp fragment of the mitochondrial hypervariable region. To ensure that the Bayesian estimate of θ was robust, we performed a number of repeat analyses. Fourteen preliminary analyses were conducted with a range of starting parameters (e.g. sample size and sampling increment). Posterior probability distributions were compared between runs to assess whether we converged on an estimate of θ. We also assessed convergence by calculating the effective sample size (ESS) (using the program Tracer v1.4, (http://beast.bio.ed.ac.uk/Tracer). An ESS of 100–200 has been suggested to indicate convergence. Our estimates were well above this value, in the thousands or greater. After preliminary analyses were completed, a final estimate of θ was performed from 10 replicates, each with the following starting parameters: $\theta = 0.015$, and a linear prior of 0.0001 – 3. Two initial chains sampled 5000 trees with a sampling increment of 40, of which the first 7000 trees sampled were discarded, followed by 4 final chains to produce the estimate, in which, after a burn-in of 5000 trees, every 50^{th} tree of 5×10^6 trees was sampled. Adaptive chain heating was used (chain temperatures 1,

1.1, 1.2, and 1.3). We also performed separate LAMARC analyses to test for signatures of exponential growth or shrinkage using the same searching strategy.

3.4 Results

3.4.1 Estimating genetic diversity and population size of Huia

Eight mitochondrial haplotypes among the 21 Huia individuals examined were recorded (Fig. 3.3). The resulting dataset did not show extensive levels of artifactual mutations; samples 0.37386 and LB4568 only have two C>T singletons that could have arisen as a result of post mortem deamination of cytosine residues. If artifactual, these singletons would lead to a false overestimation of the extent of population expansion, and would consequently provide false population size estimates. However, these two samples were sequenced in the forward and reverse direction from independent amplifications and both sequence reads cover the singletons in question. Therefore, we are confident that they represent real variation.

Sample Number	Variable Site(s) 11 7788933 0167836	Haplotypes
AV1076	A------	1
AV1082	----C--	2
AV1083	--A----	3
AV1078	----C--	2
AV1126	----C--	2
AV1079	--A----	3
AV1085	-------	4
AV2744	-------	4
AV2745	-------	4
AV2746	--A----	3
AV2747	--A----	3
LB4567	---TC--	5
LB4568	-------	4
LB4577	-------	4
0.37386	-T--CA-	6
RMNH110.109	-------	4
RMNH110.108	A------	1
AMNH669775	A------	1
AMNH669774	----C-C	7
AMNH669772	----CA-	8
HBH	----C--	2
Consensus Sequence	GCGCTGT	

Figure 3.3: DNA nucleotide variation in 199bp of the mitochondrial hypervariable region (HVRI) among 21 Huia.

Eight haplotypes were found from 21 Huia. Modified from Lambert *et al*. (2009).

I used these data to obtain an estimate of genetic diversity (θ) of 0.011846 (95% credibility interval 0.005804–0.031121) from 199 bp of mitochondrial hypervariable region sequence. Figure 3.4 shows this estimate and the probability intervals. Mitochondrial data were used in this analysis because mutation rates for the hypervariable region have been more accurately estimated. Microsatellite data as reported in Lambert *et al*. (2009) were also used to obtain an independent estimate of genetic diversity. However, different loci gave highly divergent results (data not shown). These differences between loci are best expained by mutation rate differences. Without a better estimate of a per-locus mutation rate, diversity estimates are not accurately converted to an effective population size. Consequently, the analysis was carried out using only the mitochondrial data.

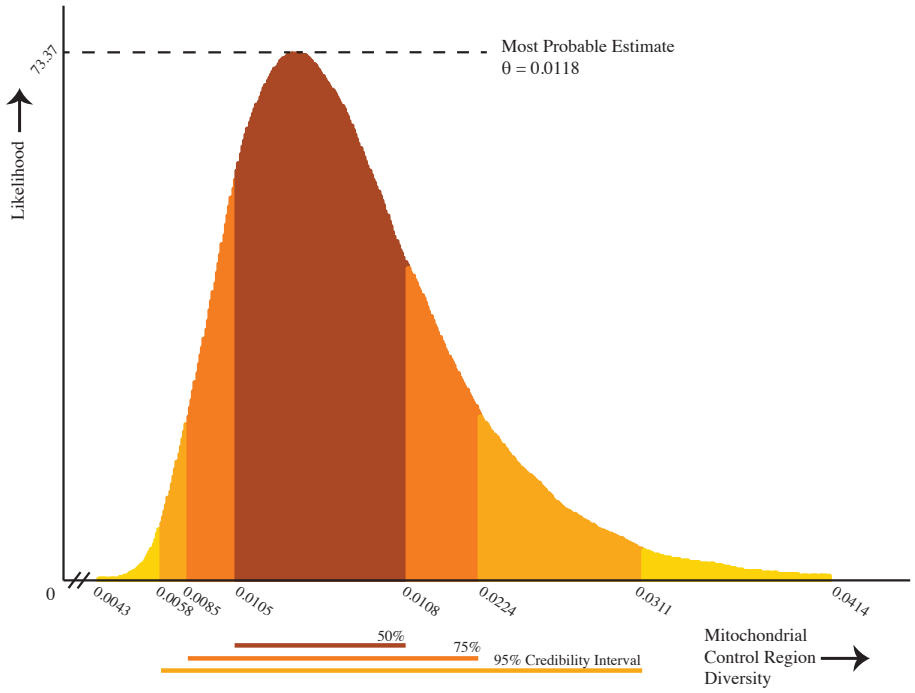


Figure 3.4: Most probable mitochondrial diversity (θ) estimate from LAMARC.
Confidence intervals for the estimate are shown.

The relationship between θ and the effective population size of breeding females is given by the expression $N_e(f) = \theta / 2\mu$, where μ is the mutation rate per base pair per

generation and generation time is defined as the average age at which a female reproduces. A generation time of 6.3 years was used in our analyses, and was based on the generation times of the related North Island Kokako (DOUBLE and MURPHY 2000) and Saddleback (ARMSTRONG *et al*. 2005). We used a mutation rate of 5.5 x 10^{-7} mutations per site per year for the hypervariable region, based on a pedigree study in Adélie penguins (MILLAR *et al*. 2008) as well as the phylogenetic estimate of 2.1 x 10^{-7} mutations per site per year. The former resulted in a mutation rate per generation of 3.47 x 10^{-6} for Huia and the latter 2.65 x 10^{-7}. Using the above values, $N_e(f)$ for Huia was estimated at 1709–4477. The range of estimates of the overall effective population size (N_e) of Huia using this generation time and these mutation rate variables was 3419–8954 breeding adults, assuming an equal sex ratio. The ratio of effective to census population size (N_c) in wild populations is often quite low and it has been suggested that a N_e: N_c ratio of 0.1 is appropriate (FRANKHAM 1995; FRANKHAM *et al*. 2005). This resulted in an expected census population size of 34,187 birds for the higher mutation rate. If we use the slower molecular rate of 2.1 x 10^{-7} (WENINK *et al*. 1994), the mean estimate of the census population size is 89,538. Table 3.2 gives details of the credibility intervals for all these estimates. This range of population size estimates is best described as 'moderately high'. Estimates of growth rate (g) were consistently positive and large (~1111), with confidence intervals excluding zero (419–2776), indicative of an expanding population (http://evolution.gs.washington.edu/lamarc/).

Table 3.2: Long-term population size estimates of Huia based on mitochondrial hypervariable region diversity. Reproduced with permission from Lambert *et al* (2009).

Mutation rate per base pair per generation (6.3 years)	Genetic diversity θ mean (95% credibility interval)	$N_e(f)$ ($\theta/2\mu$)	N_e ($N_e(f)$ x 2)	N census size (N_e:N_c ratio of 0.1)
3.47 x 10-6	**0.011846** (0.005804 – 0.031121)	**1709** (838 – 4491)	**3419** (1675 – 8982)	**34188** (16750 – 89815)

3.5 Discussion

Huia were characterised by a high level of genetic variation (LAMBERT *et al*. 2009). Two genetic studies of Kokako reported low levels of genetic structuring among modern populations for both mtDNA (DOUBLE and MURPHY 2000) and microsatellite

DNA loci (HUDSON *et al*. 2000). However the level of genetic variation in Kokako, prior to their reduction in numbers, remains unknown. Although Kokako adults remain in the same territory for many years, it has been suggested that juveniles disperse early to find mates and/or establish territories (INNES and FLUX 1999) and that this is responsible for the recorded low levels of genetic structure. In contrast, early observations of Huia behaviour, such as apparent site fidelity and the presence of family groups, suggest that Huia would exhibit a high level of population genetic structure. Contrary to this prediction, an examination of six polymorphic nuclear loci in Huia using a series of analytical methods to detect population differentiation suggested no evidence of structure among the Huia samples examined (LAMBERT *et al*. 2009). Huia therefore exhibit a lack of population genetic structure, despite known provenance samples coming from locations up to 300 km apart. However, the area from which these samples came was covered in continuous lowland forest prior to human settlement, which may have allowed extensive mixing of resident populations.

As a result the whole species can be considered a single, panmictic population and is amenable to a per-species estimate of genetic diversity and population size. The estimate of the population size of Huia is relevant to the period prior to the human settlement of New Zealand and, depending on likely mutation rates and generation times, indicates a 'moderate' historical population size of 34,187 to 89,539 individuals. It is likely, however, that the pre-human population size was higher than our estimate, as our analyses suggests a period of population expansion. The estimate could vary up to ≈ 200,000 individuals if the generation time of Huia were closer to that of Saddlebacks (2.5 years) instead of Kokako (6.3 years) (Fig. 3.5). In combination, these factors suggest moderate to high Huia population densities. As a consequence, gene flow within the population was likely to be significant and would have thereby promoted genetic homogeneity within the species. In addition, populations of moderate size typically exhibit a level of genetic inertia that may also have contributed to the overall genetic homogeneity of the species. This finding is in contrast to the original expectation that Huia populations would have exhibited high levels of population structure.

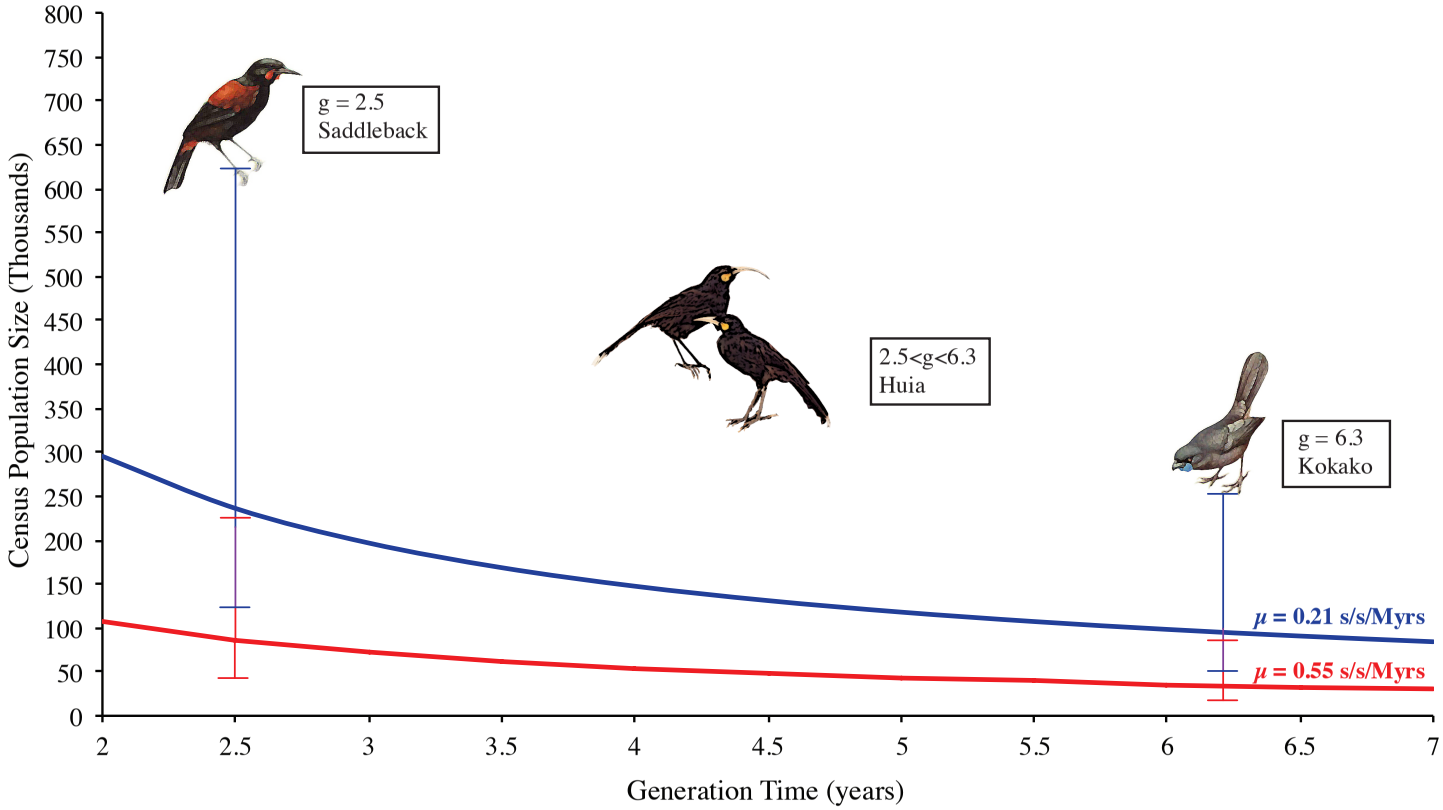


Figure 3.5: Effect of generation time and mutation rate on the population size estimate. For this study we employed a generation time from the Huia's confamilial, the Kokako, of 6.3 years, and two mutation rates. Shown also is the estimate we would have obtained if we had used the Saddleback's generation time of 2.5 years. 95% confidence intervals are shown for the census estimates for the two mutation rates and to generation times. Estimates of the census size are based on Frankham's (1995) observation of N_e:N_c ratios of approximately 0.1 in wild populations.

In conclusion, mitochondrial DNA variation in Huia allowed us to estimate the likely census population size for this species prior to its decline. Our estimate of pre-human settlement Huia population size was moderate to high and showed a signature of expansion. From this we may suggest that the species was not in decline prior to the arrival of humans. This corroborates the theory that human settlement brought about the extinction of the New Zealand Huia.

Kokako and Saddleback, both threatened, are emblematic species of conservation concern in New Zealand. No measures of genetic diversity prior to their decline exist at present. Using historical samples, pre-decline population structure and sizes could be compared to current ones, aiding species management. As many species of conservation concern become rarer in the wild, scientific programmes might benefit from using older specimens in museums. This would have less impact on modern populations of these species and would make use of a valuable museum resource. Hence, similar ancient DNA methods to those used here could be applied to scientific attempts to better understand the causes of species' decline.

3.6 References

ARMSTRONG, D., R. DAVIDSON, J. PERROTT, J. ROYGARD and L. BUCHANAN, 2005 Density-dependent population growth in a reintroduced population of North Island saddlebacks. Journal of Animal Ecology **74:** 160-170.

BULLER, W., 1870 Notes on the Ornithology of New Zealand. Trans. Proc. NZ Inst. **3:** 24-29.

BULLER, W., 1888 *A History of the Birds of New Zealand*. Van Vorst, London.

BURTON, P., 1974 Anatomy of head and neck in the huia (*Heteralocha acutirostris*) with comparative notes on other Callaeidae. Bulletin British Museum Natural History (Zoology) **27:** 1-48.

COLENSO, W., 1887 A description of the curiously deformed bill of a huia (*Heteralocha acutirostris,* Gould) an endemic New Zealand bird. Trans. Proc. NZ Inst. **19:** 140-145.

DOUBLE, M., and S. MURPHY, 2000 Genetic variation within and among populations of North Island kokako, pp. in *Science and Research Internal Report* Department of Conservation, Wellington.

DRUMMOND, J., 1910 Emu **10:** 60-62.

FRANKHAM, R., 1995 Effective population size / adult population size ratios in wildlife: a review. Genetical Research **66:** 95-107.

FRANKHAM, R., J. D. BALLOU and D. A. BRISCOE, 2005 *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge.

GOULD, J., 1837 *Synopsis of the birds of Australia and adjacent Islands*. Gould, London.

HEATHER, B., and H. ROBERTSON, 1998 *The Field Guide to the Birds of New Zealand*. Oxford University Press, Oxford.

HIGGINS, P., and J. PETER, 2002 Pardalotes to Shrike-thrushes, pp. 963-1017 in *Handbook of Australian, New Zealand and Antarctic Birds*. Oxford University Press, Melbourne.

HUDSON, Q., R. WILKINS, J. WAAS and I. HOGG, 2000 Low genetic variability in small populations of New Zealand kokako *Callaeas cinerea* wilsoni. Biological Conservation **96:** 105-112.

INNES, J., and I. FLUX, 1999 North Island Recovery Plan 1999-2009, pp. in *Threatened Species Recovery Plan 30*. Department of Conservation, Wellington.

KUHNER, M. K., 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics **22:** 768-770.

LACK, D., 1971 *Ecological Isolation in Birds*. Blackwell Scientific, Oxford.

LAMBERT, D. M., L. D. SHEPHERD, L. HUYNEN, G. BEANS-PICÓN, G. H. WALTER *et al.*, 2009 The Molecular Ecology of the Extinct New Zealand Huia. PLoS ONE **4:** e8019.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al.*, 2008 Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

MOORHOUSE, R., 1996 The extraordinary bill dimorphism of the huia (*Heteralocha acutirostris*): sexual selection or intersexual competition. Notornis **43:** 19-34.

PHILLIPS, W., 1963 *The Book of the Huia*. Whitcombe and Tombs, Christchurch.

POTTS, T., 1885 Oology of New Zealand. New zealand Journal of Science **2:** 373-484.

SELANDER, R., 1966 Sexual selection and dimorphism in birds. Condor **68:** 113-151.

SELANDER, R., 1972 Sexual selection and dimorphism in birds, pp. 180-230 in *Sexual Selection and the Descent of Man*, edited by C. B. Aldine, Chicago.

SHEPHERD, L. D., and D. M. LAMBERT, 2007 The relationships and origins of the New Zealand wattlebirds (Passeriformes, Callaeatidae) from DNA sequence analyses. Molecular Phylogenetics and Evolution **43:** 480-492.

WENINK, P., A. BAKER and M. TILANUS, 1994 Mitochondrial control-region sequences in two shorebird species, the turnstone and the dunlin, and their utility in population genetic studies. Molecular Biology and Evolution **11:** 22-31.

WORTHY, T., and R. HOLDAWAY, 2002 *The Lost World of the Moa*. Canterbury University Press, Christchurch.

4 Chapter Four

INTRON RECOVERY IN ADÉLIE PENGUINS

4.1 Abstract

Mitochondrial DNA markers have been of great use in Adélie penguins for the elucidation of population history. However there are uncertainties associated with relying entirely on one molecule to make inferences, and it is increasingly common to use multiple loci to gain a more complete picture of the ecology and evolution of a species or group of species. Nuclear intron markers, presumed to evolve in a nearly-neutral fashion, are now frequently used in phylogenetics and population genetics. We used PCR amplification and direct sequencing to assess 26 previously described nuclear intron markers that may be amenable for population and phylogenetic analysis in Adélie penguins. Eleven of 26 markers produced a single PCR product and six of the 26 produced high quality target sequence. Sequence data obtained from these loci suggest a moderately high SNP density of 1 SNP per 78 - 125bp of intron sequence in Adélie penguins. This work shows that cross-amplification and screening of existing nuclear marker resources is a time and cost-effective approach to marker selection, particularly when moderate numbers of markers are required.

4.2 Introduction

The fields of phylogenetics and population genetics have long relied on mitochondrial DNA markers, due to the unique biology of the mitochondrial genome. However, there are disadvantages when using mtDNA as a sole source of phylogenetic information (BALLOUX 2010; MEIKLEJOHN *et al*. 2007; RUBINOFF and HOLLAND 2005). This has contributed to a rise in multi-locus approaches, along with decreasing sequencing costs and improving technology. Non-coding introns are now routinely used in phylogenetics and population genetics (BRITO and EDWARDS 2009; HARE 2001; THOMSON *et al*. 2010; ZHANG and HEWITT 2003). Despite the steady increase in the number of studies using introns either independently or in conjunction with mtDNA, working with introns is not always straightforward. Unlike mitochondrial regions (with exceptions, e.g. *HVR1*), where nearly universal primers have been developed for vertebrates (SORENSON *et al*. 1999), widely tested nuclear primer sets with a high likelihood of working within a particular group of species are not available. The presence of length-variant heterozygotes, alignment, allele phasing, allelic dropout, gap treatment and recombination, can also influence the ease with which data can be obtained and used in empirical studies (CREER 2007; THOMSON *et al*. 2010; ZHANG and HEWITT 2003).

In the first instance, selecting an intron or series of intron markers for a project is not necessarily a routine procedure. Primer sequences are not normally available for most species and so need to be designed. If a fully annotated genome or EST sequences is available, primers can be designed within exonic regions flanking the variable intron (referred to as exon-primed, intron crossing primers: EPIC primers for short) (PALUMBI 1995; SLADE *et al*. 1993). Though there have been recent advances in non-model species genomics, largely brought on by progresses in second-generation sequencing technology (ALLENDORF *et al*. 2010; JOHANSSON 2009; OUBORG *et al*. 2010; SCIENTISTS 2009), most researchers do not have genomic resources for their study species available. In these cases primers can be designed from genomic data from a closely related species. A good example of this approach can be found in (BACKSTRÖM *et al*. 2008). These authors compared zebra finch (*Taenopygia guttata*) and chicken (*Gallus gallus*) genomes and

isolated 242 markers spread relatively evenly across the avian genome. Two hundred of these markers were subsequently sequenced in collared flycatchers (*Ficedula albicollis*) and PCR amplification was tested for 122 markers across a range of avian species. This approach relies on the assumption that single copy nuclear exon coding sequences are conserved in sequence and position across even distantly related species (FEDOROV *et al*. 2002; ROGOZIN *et al*. 2003). Another assumption is that no gene duplication has occurred in our study species that may result in inadvertent amplification of paralogous loci (CREER 2007). Reference marker sets have been developed using similar strategies for mammals (IGEA *et al*. 2010; LYONS *et al*. 1997), birds (Backström *et al*., 2008; Kimball *et al*., 2009; Primmer *et al*., 2002), fishes (HASSAN *et al*. 2002; LI *et al*. 2010), shrimp (BIERNE *et al*. 2000), and for non-vertebrate metazoa (CHENUIL *et al*. 2010), among others. Rather than develop many primers, one can instead first choose an array of potentially suitable primers from those that have been proven to work in species closely related to the study species. Another approach, infrequently used in population genetics, is to develop anonymous non-coding markers (ANMs) (JENNINGS and EDWARDS 2005; LEE *et al*. 2009). The use of these can be complicated by the frequent presence of repetitive elements, and their very anonymity makes comparisons with other taxa and studies more difficult (THOMSON *et al*. 2010).

As a result of the lack of readily accessible, 'tried and tested' nuclear marker sets for most study species, researchers generally aim to test an array of primers for different intron markers, and cost can often be an added issue to take into account. It may be more cost-effective to test a wide array of markers in search of those that work best in the study species, rather than to work with fewer markers and use more extensive optimization methods to get them working (CREER 2007; SLADE *et al*. 1993).

The objective of this study was to recover intron markers amenable for phylogenetic and population genetic analysis in Adélie penguins, as well as provide information about their variability within the species, without resorting to costly marker isolation.

Table 4.1 Details of primer pairs tested in Adélie penguin samples from the Ross Sea, Antarctica.

Primer information was taken from the original reference describing the primer and supplemented by looking up the loci on the chicken genome. Name* refers to primer name from the literature and working name I assigned to the primer for convenience. Closest Sequence = closest species to Adélie penguins sequenced to date for the marker; P = penguin species, C = species belonging to one of the closest orders to penguins, O = other species not belonging to closest orders (details in text).

Primer location			Primer		Length (bp)		Closest Sequence
Locus**	Intron	Chr; Exon	Name*	Sequence (5'-3')	Exon	Intron	
AK1[1]	5	?; 5-6	AK5b+ AK6c-	ATTGACGGCTACCCTCGCGAGGTG CACCCGCCCGCTGGTCTCTCC	99	350-500	P
Myo[2]	2	1; 2-3	Myo2 Myo3	GCCACCAAGCACAAGATCCC CGGAAGAGCTCCAGGGCCTT	10	700	P
LDH-B[3]	4	1; 3-4	/23F /23R	GGAAGACAAACTAAAAGGAGAAATGATGGA TTCCTCTGAAGCAGCTTGAGACGACTCTC		480	O
MPP[3]	4	4; 4-5	/22F /22R	TACATCTACTTTAACACCTGGACCACCTG TTGCAGATGGAGAGCAGGTTGGAGCC		390	O
ODC[3]	6-7	3: 6-8	/24F /24R	GACTCCAAAGCAGTTTGTCGTCTCAGTGT TCTTCAGAGCCAGGGAAGCCACCACCAAT	204	526	O
ODC1[4]	9	3: 9-10	/20F /20R	AGCGTGCAAAAGAACTTGACC CTGAGCTACCAATTTTAATGCATCTAA	115	339	O
ACTB[4]	2	A: 2-3	/21F /21R	AATGAGCTGAGAGTAGCCCCTG TGGCTACATACATGGCTGGG	50	494	O
ACTB[5]	3	A: 3-4	/13F /13R	CCTGATGGTCAGGTCATCA CAGCAATGCCAGGGTACAT		300	C
FGB[6]	5	A: 5-6	Fib-5 / 9F Fib-6 / 9R	CGCCATACAGAGTATACTGTGA GCCATCCTGGCGATTCTGAA			O
UCHL3[7]	5	1: 4-5	/16F /16R	GCTTGTGGGACAATTGGG TATTTGGCCCTCTCTTCAGG	104	357	O
HMG-2[7]	4	4: 3-4	/6F /6R	GAAATGTGGTCTGAACAGTC TTGCTCTTGGCACGATATGC	26	483	O
18142[7]		7; ?	/12F /12R	GTGTGGAGGCAGTTGATCC ACACTCTGAATGGGATCCAC	25	468	O
504[7]		23; ?	/1F /1R	AAAGCTGATGCTGGGAAG CCAGAAGTACACAGTTATC	54	810	O
CHDZ[4]	15	Z: 15-16	/2F /2R	TAGAGAGATTGAGAACTACAGT GACATCCTGGCAGAGTATCT	37	464	O
CHDZ[4]	18	Z: 18-19	/4F /4R	TACATACAGGCTCTACTCCT CCCCTTCAGGTTCTTTAAAA	69	239	O
CHDZ[4]	24	Z: 24-25	/11F /11R	CATTCACCTGCACTCCTGAG GGCCTTTAAAGGACARTTCA	76	384	O
CHD1W[8]	7	W: 7-8	/3F /3R	AGTATCAAGTAGTGGAAAGG AATAGTAATCTGGATAACCG		250	O
CHD1W[8]	11	W: 11-12	/10F /10R	CACAGACCAAGCGATTAAAATTTAAT GCTTCATCAATCCCTATAAACA		582	O
CHD1W[8]	25	W: 25-26	/7F /7R	AAAGGCCCAACATTCCGAATA CCAATCTATATCAAAATGAGCT		684	O
VLDLR[4]	7	Z: 7-8	/8F /8R	CAGAAGTGGAGAATGCATAG ACAGTCACATTCATAGCCAA	123	452	O
VLDLR[4]	9	Z: 9-10	/14F /14R	AAGTGTGAATGTAGCCGTGG TCGGTTGGTGAAAATCAGAC	33	376	O
OPSIN[8]	1	6: 1-2	/15F /15R	GAGGGCTTCATGGTCTCCTT CACAACGTAGCGCTCCAG		438	O
OPSIN[8]	3	6: 3-4	/17F /17R	TCAGCCTCATCCTCTTCTCC GGTGTCAGCTTCCTTCTGCT		204	O
OPSIN[8]	4	6: 4-5	/19F /19R	CACCGTGTACAATCCCATCA TGGTAACCACAGCACAGCAT		370	O

Primer location			Primer		Length (bp)		Closest Sequence
Locus**	Intron	Chr; Exon	Name*	Sequence (5'-3')	Exon	Intron	
RHO[8]	1	26: 1-2	/5F /5R	TCTTTGGAGTAACAGGGTGC CTTACAGACCACCACATATC		818	O
RHO[8]	4	26: 4-5	/18F /18R	GCATTCTTTGCCAAGAGCTC GGTTCTTGCCGCAGCAGAGG		356	O

References for each primer pair are as follows: Adenylate Kinase (*AK1*) intron 5[1] (SHAPIRO and DUMBACHER 2001), Myoglobin (*Myo*) intron 2[2] (JOHANSSON and ERICSON 2004), Lactate dehydrogenase (*LDH-B*) intron 3, Myelin Proteolipid Protein (*MPP*) intron 4, Ornithine Decarboxylase (*ODC*) introns 6-7[3] (FRIESEN *et al*. 1999), *ODC1* intron 9, Beta-Actin (*ACTB*) intron 2, Chromo-helicase DNA binding on Z (*CHDZ*) introns 15, 18, 24, Very low density lipoprotein/Vitellogenin receptor (*VLDLR*) introns 7, 9[4] (BORGE *et al*. 2005), *ACTB* intron 3[5] (WALTARI and EDWARDS 2002), Beta-Fibrinogen (*FGB*) intron 5[6] (LOVETTE and RUBINSTEIN 2007), Ubiquitin carboxyl-terminal esterase L3 (*UCHL3* intron 5, High mobility group protein B2 (*HMG-2*) intron 4, Proteasome 26S subunit, non-ATPase, 14 (*18142*), Non histone chromosomal protein *HMG-17* (*504*)[7] (BACKSTRÖM *et al*. 2008), Chromo-helicase DNA binding 1 on W (*CHD1W*) introns 7, 11, 25, Opsin (*OPSIN*) introns 1, 3, 4, Rhodopsin visual pigment (*RHO*) introns 1, 4[8] (AXELSSON *et al*. 2004).

4.3 Methods

4.3.1 Choosing intron markers for Adélie penguins

Initially I performed a literature search for studies using intron markers for phylogenetic or population genetic studies in Adélie penguins. The search was successively broadened to encompass all penguins (Sphenisciformes), closely related avian orders (Procellariformes, Ciconiiformes, Pelecaniiformes, Gaviiformes) (HACKETT *et al*. 2008), and finally birds in general.

Markers for testing were chosen based on their having amplified in penguin species or other closely related species (see above) (FAIN *et al*. 2007; JOHANSSON and ERICSON 2004; SHAPIRO and DUMBACHER 2001; SUNDSTRÖM *et al*. 2003; WALTARI and EDWARDS 2002), and also having a size range between 300-1000bp for ease of downstream work. As not many markers fit these criteria, other markers isolated from chicken (ARMSTRONG *et al*. 2001; AXELSSON *et al*. 2004; FRIDOLFSSON and ELLEGREN 1999; FRIESEN *et al*. 1999), and PCR tested across an array of avian species (eg marbled murrelet, *Brachyramphus marmoratus*) (FRIESEN *et al*. 1999), various Galliforms (ARMSTRONG *et al*. 2001), pied and collared flycatchers (*Ficedula hypoleuca* and *Ficedula albicollis*)

(BORGE *et al*. 2005; PRIMMER *et al*. 2002), Sturnidae (LOVETTE and RUBINSTEIN 2007)) were also considered. If possible, markers were also given preference if mention was made of their usefulness in population studies. A total of twenty-six primer pairs were thus selected to screen in modern Adélie penguin samples (details and references in Table 4.1).

4.3.2 DNA Extractions

Genomic DNA was extracted from 40-100 μl of blood preserved in buffer (SEUTIN *et al*. 1991), following the standard phenol-chloroform protocol (SAMBROOK *et al*. 1989), and diluted to 50 ng/μl in water for downstream work. The blood samples used originated from three colonies of Adélie penguins from the Ross Sea area of Antarctica, collected by D. Lambert and colleagues (Fig. 4.1).

4.3.3 Screening intron markers in modern Adélie penguins

Given that the majority of the markers selected for testing in Adélie penguins had never been amplified in the species before, it was likely each marker might require extensive optimization to produce a clean amplification product and sequence. The strategy implemented here was to screen more markers than might be needed, follow a few optimization steps and select the markers that worked best.

Two loci that were considered highly likely to amplify in Adélie penguins were the subject of more extensive optimization, including a wide range of annealing temperatures and using a minimum of five DNA extracts. These loci were adenylate kinase intron 5 (*AK1i5*) (SHAPIRO and DUMBACHER 2001) and myoglobin intron 2 (*myo2*) (JOHANSSON and ERICSON 2004). For these two loci, separate optimization PCRs were carried out (details in Table 4.2 and 4.3). The published PCR protocol of each was used as the

starting conditions for optimization. Based on results from the first gradient PCRs, conditions were more finely adjusted to improve amplifications.

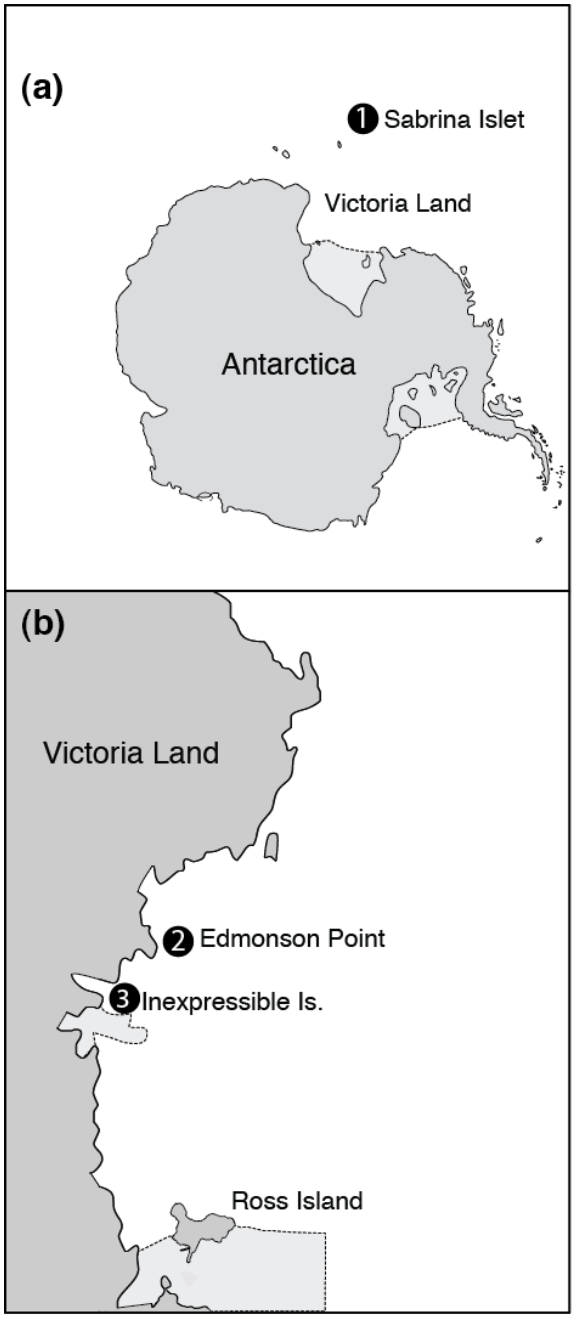


Figure 4.1: Distribution of Adélie penguin samples used in this study

In order to quickly screen the twenty-four remaining primer pairs, while minimizing optimization time, the primers were first grouped according to melting temperatures (see

Table 4.2 and 4.3 for testing conditions, Table 4.4 for primer groupings). Those that had the closest melting temperatures were screened together in groups of four using temperature gradient PCR covering three different annealing temperatures and two different concentrations of $MgCl_2$ (1.5 and 2mM), as these two variables were the most likely to have an effect on PCR success. Standard PCR conditions were the following: dNTPs 0.2mM, primers 0.4-0.5μM, platinum Taq 0.5units, 1x buffer (all Invitrogen), 1-2μl of 50ng/μl DNA sample, and water to 25μl.

Table 4.2 PCR conditions and annealing temperature ranges tested during optimization runs for 26 primer pairs in modern samples.

Testing Group	T_{anneal} (°C)	$[MgCl_2]$ (mM)	PCR program	Other
1	48.5, 49.2, 50	1.5, 2	G1	
2	51, 52, 53.3	1.5, 2	G1	
3	53.5, 54.2, 55	1.8, 2.2	G2	
4	55.5, 56, 57	1.8, 2.2	G2	
5	57.5, 58, 59	1.5, 2	G1	
6	59.5, 60, 61	1.5, 2	G1	
AK1i5	53, 53.8, 55.1, 56.9, 59.4, 61.3, 62.5, 63	1.5, 2	G3, T1	+/- BSA 1mg/ml
Myo2	56, 56.7, 57.7, 59.1, 61.2, 62.7, 63.6, 64	1.5, 2	G4, T2	+/- BSA 1mg/ml

'Other' refers to other components of the PCR varied during optimization.

Table 4.3 PCR program details for modern samples.

Program	Description
G1	5' 95°C; 30 x [45" 94°C, 30" (annealing gradient), 1' 72°C]; 5' 72°C
G2	5' 95°C; 35 x [45" 94°C, 45" (annealing gradient), 1' 72°C]; 5' 72°C
G3	5' 94°C; 40 x [45" 94°C, 1' (annealing gradient), 1' 72°C]; 5' 72°C
G4	5' 94°C; 40 x [40" 94°C, 40" (annealing gradient), 1' 72°C]; 5' 72°C
T1	5' 94°C; 9 x [40" 94°C, 40" 60°C, 1' 72°C]; 31 x [40" 94°C, 40"58°C, 1' 72°C]; 5' 72°C
T2	5' 94°C; 9 x [40" 94°C, 40" 62°C, 1' 72°C]; 31 x [40" 94°C, 40"58°C, 1' 72°C]; 5' 72°C
C1	5' 94°C; 40 x [30" 94°C, 1' 55°C, 1' 72°C]; 10' 72°C

Programs labelled 'G' are temperature gradient programs, 'T' are two-step or touchdown programs.
' : minute, '' : second.

For each of the primer pairs amplifying a product in modern Adélie penguin samples, PCR conditions were identified that yielded a strong, single band of the correct size. If a locus consistently presented multiple amplification products (including one of the correct size), it was not followed up on for further testing. Promising primer pairs were used to amplify intron markers across a wider range of samples using PCR conditions shown to work best. In some cases further slight modifications of annealing temperatures or touchdown programs and $MgCl_2$ concentrations were necessary for primer pairs that initially worked well but then showed mispriming. PCR products were purified using Agencourt AMPure beads following the manufacturer's protocol, and were quantified on 2% agarose gels using Invitrogen Low Mass Ladder as a standard, as Agencourt AMPure magnetic beads interfere with Nanodrop readings (pers. obs.).

Purified products were sent for direct Sanger sequencing to the Auckland University sequencing service (http://www.biosciencesresearch.co.nz/services/dna-sequencing/), in two directions using the forward and reverse PCR primers as sequencing primers. Electropherograms were visualised using 4peaks (http://mekentosj.com/science/4peaks/) and sequences were edited in SEQUENCHER 4.6 (Gene Codes Corporation). Forward and reverse sequences were aligned in SEQUENCHER 4.6 in order to resolve sequence ambiguities arising either from error or the presence of two alleles (single-nucleotide or length-variant heterozygotes). Length-variant heterozygotes resulting from insertion-deletion polymorphisms (PALUMBI 1995) in particular present problems as electropherograms will appear superimposed and sequence-editing software cannot separate the overlapped reads (Fig. 4.2). As a first step in the analysis of these reads, the programs INDELLIGENT (DMITRIEV and RAKITOV 2008) and CHAMPURU (FLOT 2007) were used. The program INDELLIGENT can resolve mixed traces arising from two allelic traces superimposed onto each other, in most cases, provided each of the double peaks on the raw electropherogram can be called accurately following IUPAC nomenclature (eg as Y, R, W…). CHAMPURU works in a similar fashion, requiring both the forward and the reverse ambiguous sequences. Both programs were used in order to reach better consensus sequences

Table 4.4 Results of PCR condition testing intron screen in modern Adélie penguins.

Test Conditions	Primer Pair ID	Locus	Expected Size (bp)	DNA sample	PCR			Sequencing	
					Band	T_{anneal} (°C)	$[MgCl_2]$ (mM)	Success	Ex-Int length (bp)
GROUP 1	1	504	864	T164[2]	FDB	48.5	2		
	2	CHD1Z15	501	T168[2]	FDB	48.5	2		
	3	CHD1W7	250	T169[2]	F				
	4	CHD1Z18	308	T173[2]	B	49	2	FAIL	
GROUP 2	5	RHO1	818	T164[2]	F				
	6	HMG2	509	T168[2]	B	53.3	2	GOOD	30 – **470** - 18
	7	CHD1W25	684	T169[2]	F				
	8	VLDLR7	575	T173[2]	SB	52	2	FAIL	
GROUP 3	9	Fib5	545	T347[1]	F	53.5-55	2.2		
	10	CHD1W11	582	T347[1]	F				
	11	CHD1Z24	460	T346[1]	SB	54	2.2	BAD	
	12	18142	493	T346[1]	FDB	54	2.2		
GROUP 4	13	ACTB3	300	T345[1]	F				
	14	VLDLR9	399	T345[1]	SB	55.5-57	2.2	BAD	
	15	OPSIN1	438	T344[1]	F				
	16	UCHL3	461	T344[1]	SB	55.5	2.2	GOOD*	5 – **358** - 50
GROUP 5	17	OPSIN3	204	T332[1]	SB	58-59	2	FAIL	
	18	RHO4	356	T332[1]	DB	57.5-59	2		
	19	OPSIN4	370	T336[1]	F				
	20	ODC9	454	T336[1]	F				
GROUP 6	21	ACTB2	544	T337[1]	B	59.5-61	2	GOOD	68 – **531** - 23
	22	MPP	390	T337[1]	SB	59.5-61	2	GOOD	71 – **217** - 8
	23	LDH-B	480	T340[1]	F				
	24	ODC6	730	T340[1]	SB	59.5-60	2	GOOD	65 – **157** – 82 – **334** - 7

Table 4.4 continued

Test Conditions	Primer Pair ID	Locus	Expected Size (bp)	DNA sample	PCR			Sequencing	
					Band	T_{anneal} (°C)	[$MgCl_2$] (mM)	Success	Ex-Int length (bp)
AK1i5	n/a	AK1i5	350-500	T01-T05[3]	SB	60	1.5	GOOD*	**474** - 47
Myo2	n/a	Myo2	700	T01-T05[3]	FB	63	1.5		

Test conditions detailed in text, Table 4.2 and Table 4.3 Primer pair ID refers to Table 4.1. PCR amplification band descriptions: SB=strong band, B=band, DB=double band, FDB=faint double band, F=failed. T_{anneal} (°C) and [$MgCl_2$] (mM) refers to best annealing temperatures and $MgCl_2$ concentrations found for each intron marker. GOOD* sequencing result=length variant heterozygotes found. Grey boxes indicated markers dropped due to bad PCR/Sequencing results. Length of primer trimmed sequences obtained in final column – exon sequence and intron sequence length (intron in bold). Colonies from which DNA samples originate are marked in superscript (eg T340[1]), where 1=Sabrina Islet, 2=Edmonson Point, 3=Inexpressible Island (Fig. 4.1).

.

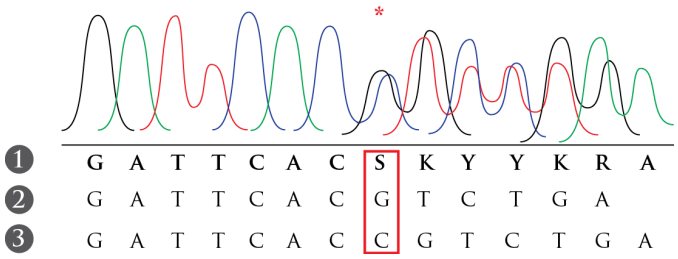


Figure 4.2 Hypothetical length variant heterozygote electropherogram read showing mixed peaks as a result of a 1bp indel (marked in red).

Sequence 1 is the mixed read from the direct sequencing, with ambiguous peaks called correctly using IUPAC notation (key in Fig. 4.6). Sequences 2 and 3 show the phased alleles after detection of a 1bp shift.

For sequences with more than one heterozygote position present, or for which superimposed reads could not be resolved, purified PCR products were cloned in order to adequately phase the alleles. An Invitrogen TOPO-TA cloning kit was used mainly following the manufacturer's protocol, with the exception that each vial of One-Shot competent cells was used for five cloning reactions. Ten white colonies per PCR product were picked from LB ampicillin and X-gal plates. Following this the colonies were incubated 10 minutes at 95°C to lyse the cells, centrifuged briefly to separate the cellular debris, and cloning PCR was performed with 3μl of the supernatant. PCR conditions were as follows: M13 primers (0.25μM), $MgCl_2$ (2.5mM), dNTPs (0.31mM), buffer (1x), Taq Polymerase (0.5u) (Invitrogen) (Program details: program C1, Table 4.3). Cloning PCR products were purified using AMPure magnetic beads and submitted to direct Sanger sequencing at the Auckland University Sequencing Service using M13 primers as sequencing primers.

4.4 Results

4.4.1 Literature and NCBI search

Seven intron markers have been sequenced for five penguin species previously (Friday, August 27th 2010 search on NCBI). In *Pygoscelis adeliae*, two studies report

introns from the glyceraldehyde-3-phosphate dehydrogenase (VAN TUINEN *et al*. 2001), and *CHD1Z/W* intron C (SUNDSTRÖM *et al*. 2003). Other markers targeted in penguins are beta-fibrinogen intron 7 (*Eudyptula minor*, *Spheniscus humboldti*) (ERICSON *et al*. 2006; FAIN and HOUDE 2004), myoglobin intron 2 (*Spheniscus humboldti*, *Eudyptula minor*) (ERICSON *et al*. 2006; HACKETT *et al*. 2008), interferon regulatory factor 2 intron 2 (*Eudyptula minor*) (HACKETT *et al*. 2008), ornithine decarboxylase intron 6-7 (*Spheniscus humboldti*) (ERICSON *et al*. 2006), and adenylate kinase intron 5 (*Spheniscus mendiculus*) (SHAPIRO and DUMBACHER 2001).

In total, primer sequences for 92 nuclear markers were collected from the literature (data not shown). According to the selection criteria detailed in the methods section the primer pairs were reduced to twenty-six potentially useful in Adélie penguins (Table 4.1).

4.4.2 Primer Screen and Initial Intron Sequencing Results

At each stage of the PCR screen, primer pairs were omitted from further optimization efforts, so that those that did not amplify a single band (B or SB in Table 4.4) were not sent for sequencing. Those primer pairs that produced sequence that could not be identified either directly or using indel detection software were also excluded from downstream analyses. Of the twenty-six primer pairs, seventeen amplified a product of similar size to the expected (65.4% success rate). Of these, eleven produced a PCR product with no non-specific bands, and of these nine produced sequence with mixed results. Six of the primer pairs produced high quality sequence (Fig. 4.3), thereby giving a 23% overall success rate (Table 4.4). PCR conditions that worked best for tested primer pairs are shown in Table 4.4.

4.4.2.1 Adenylate kinase intron 5 and Myoglobin intron 2

Of the two markers that were the subject of more extensive optimization, *AK1i5* was selected for further work. *Myo2* proved difficult to amplify in Adélie penguins despite

substantial optimization efforts. A band of the correct size was regularly obtained on agarose gels after PCR, but non-specific bands were almost always present. In order to improve amplification it would be necessary to use gel purification of the correct band, sequencing, then design nested primers more specific to Adélie penguins. The presence or absence of BSA at a concentration of 1 mg/ml was tested for both *Myo2* and *AK1i5*. No beneficial effect was detected by the inclusion of this compound, if anything a slight weakening was noted for *AK1i5* and was therefore not added to further PCRs for this marker.

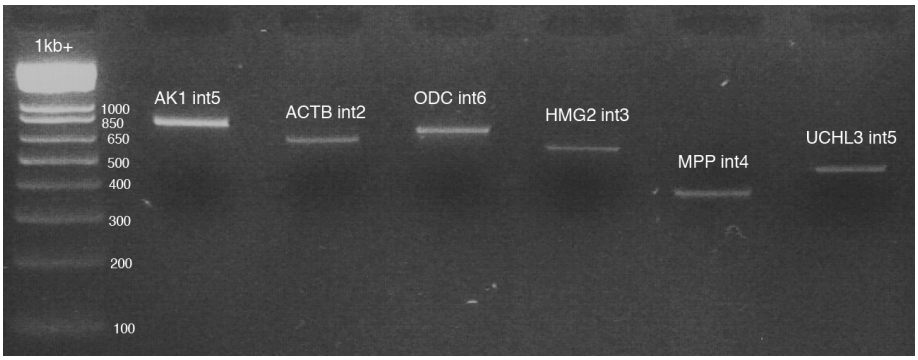


Figure 4.3 2% agarose gel showing the six intron markers that produced high quality sequence well in Adélie penguins.

On the left an Invitrogen 1kb+ size marker is shown with band size indicated in base pairs. Intron markers are labeled with locus ID names and intron number. In Adélie penguin samples sequenced, amplification product lengths excluding primers are: AK1i5 = 521 bp*, ACTB2 = 622 bp, ODC6 = 645 bp, HMG2 = 518 bp, MPP = 296 bp, UCHL3 = 413 bp. *Note: in the gel photo AK1i5 appears larger than 521bp due to amplification from a clone using M13 primers.

PCR amplification of *AK1i5* was successful. Direct sequencing of purified amplification products produced sequences 521 bp in length (excluding primers), of which 47 bp were from exon 6. A number of the sequences appeared unreadable and closer examination revealed this to be the consequence of two superimposed sequencing reads, either the result of mis-priming or an insertion-deletion polymorphism (indel). Cloning of five samples showing mixed traces revealed the presence of a 5bp indel near the 5' end of the intron (Fig. 4.4). Further mixed reads were resolved using Indelligent and Champuru. However, not all sequences could be resolved using these programs as it was not always possible to call ambiguous peaks

correctly. Twenty-eight sequences were obtained (including seven length variant heterozygote reads that were able to be phased into two alleles) and aside from the indel, eight variable sites were confirmed (Fig 4.6). A consensus sequence was compared to published sequences on the NCBI repository using BLAST and the closest hit was compared to the consensus (Fig 4.6), which was *AK1i5* for Galapagos penguin (*Spheniscus mendiculus*) (SHAPIRO and DUMBACHER 2001) (AF307894.1). The 5bp string of bases that was repeated either two or three times in Adélie penguins (TGGCA) was present twice in the Galapagos penguin sample compared. As no other samples of the same species have been sequenced, it is possible that this indel polymorphism occurs within that species as well or could be fixed.

As a result of the presence of the 5bp indel and due to some mispriming observed with the published primers, four primers were designed for use in modern samples (Fig. 5.5). AklongF (CCAAGGGTTTCCTCATTGAC) sits 5' to the published AK5b+ primer (SHAPIRO and DUMBACHER 2001), overlapping in 13bp. AklongR (GCTTCTTGATGGTCTCCTCG) is located 3' to the published AK6c- (SHAPIRO and DUMBACHER 2001) and does not overlap with it. Both of these primers were designed from the chicken gene. The internal primers Akint5aF (ATCCCAGCACAGATCCCATC) and Akint5aR (TGGGACGGTGCCCGAGTTTAG) are placed 3' to the 5bp indel to amplify the intron in two fragments and aid in phasing length-variant heterozygotes.

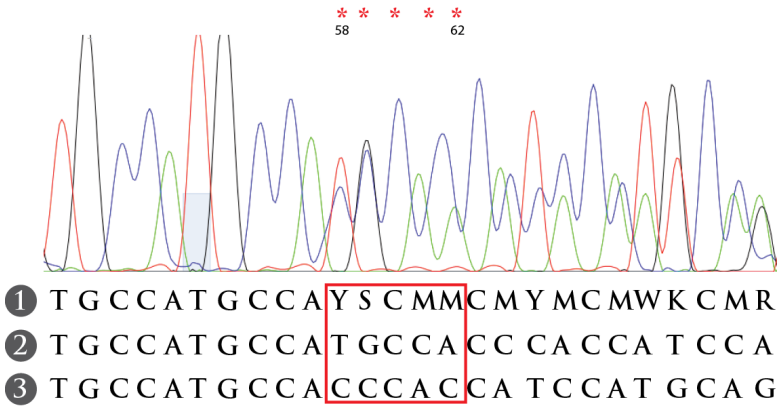


Figure 4.4 Adenylate kinase intron 5 length variant heterozygote.

A portion of a mixed trace for sample T176 sequenced with primer Ak5b+, *AK1i5*, showing the position of the 5bp indel, the mixed base calling (sequence 1) and the phasing of the two alleles using

Indelligent (DMITRIEV and RAKITOV 2008) (sequences 2 and 3). Mixed bases are called according to IUPAC nomenclature (Fig. 4.6).

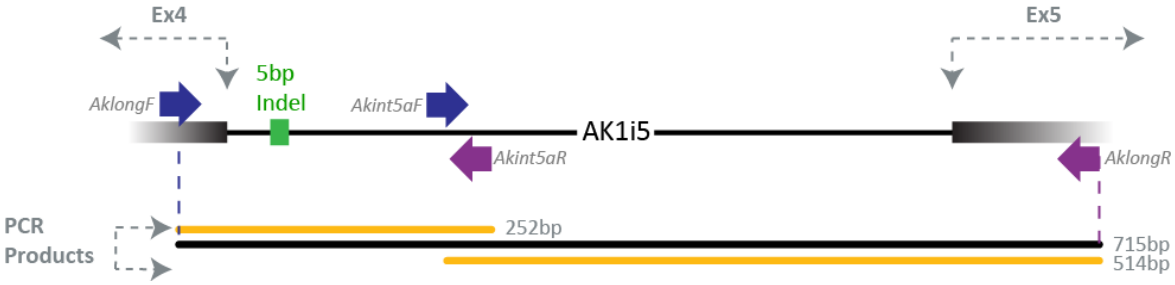


Figure 4.5 Primer positions and amplicon sizes for Adenylate kinase intron 5 external and internal primers (not to scale).

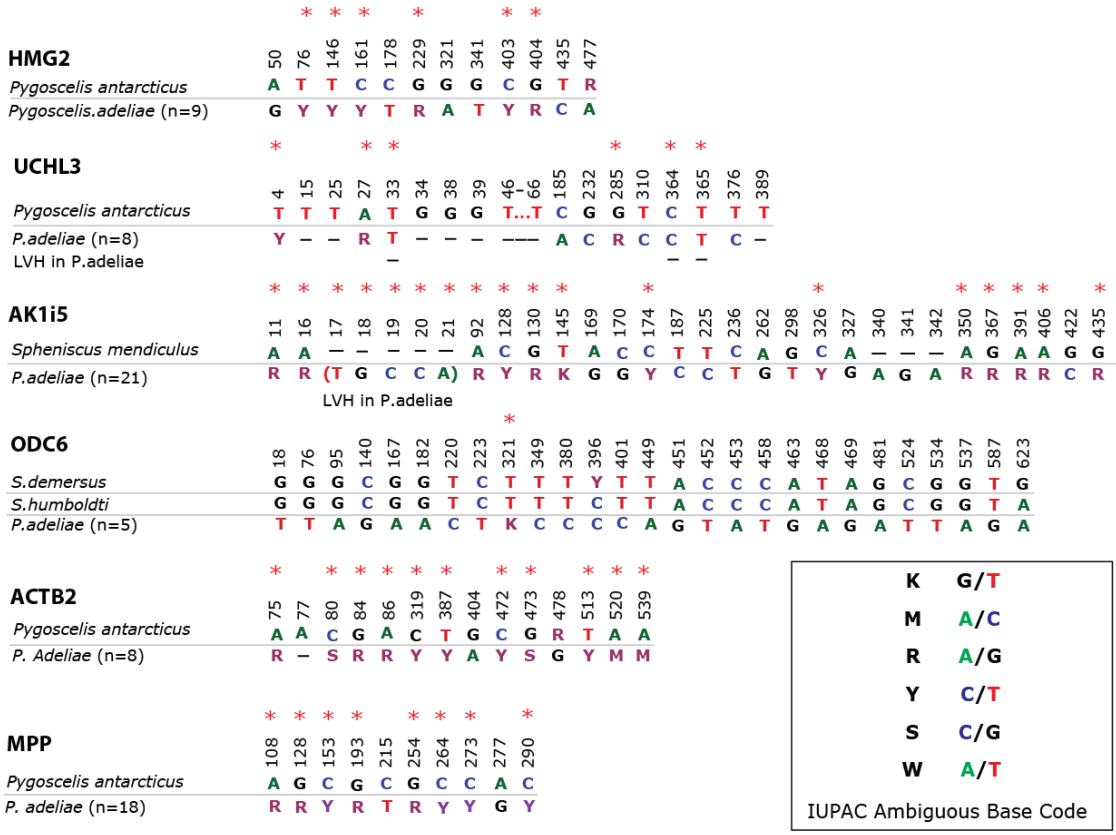


Figure 4.6 Variable sites of introns sequenced in the Adélie penguin as compared to available sequences from other penguin species.

Variable sites as compared to closest available sequence from other penguin species are shown. Chinstrap penguin sequences (*Pygoscelis antarcticus*) were obtained for this study (described in Chapter 5). Red asterisks mark sites that were found to be polymorphic in Adélie penguins sequenced

(sample numbers given above). Deletions are marked with a line. Accession numbers: *S. mendiculus* = AF307894.1, *S. demersus* = EF552742.1, *S. humboldti* = DQ881781.1.

4.4.2.2 Sequencing results for other primer pairs tested

Sequences were also obtained for markers *MPP4*, *UCHL3*, *HMG2*, *ODC6* and *ACTB2* (Fig. 4.3). Marker *VLDLR9* produced some sequence but the quality was very low. For each of these markers, a variable number of sequences were obtained from modern populations of Adélie penguins in order to gauge marker variability. Consensus sequences were compared to published penguin sequences using BLAST (Fig. 4.6) or, if no alternate penguin sequence was available, a sequence was obtained for Chinstrap penguin (*Pygoscelis antarcticus*, sample ID T361). Locus *UCHL3* contained several sites with insertion-deletion polymorphisms that were for the most part able to be resolved with the bioinformatic approach described in the methods section. For markers *ODC6*, *ACTB2*, *HMG2*, and *UCHL3* sequence sample size was low (sequence n = 5, 8, 9, 8 respectively).

4.5 Discussion

Nuclear intron markers are increasingly used in phylogenetics and population genetics as the per-base cost of sequencing decreases and technology to resolve complications arising from diploidy, recombination, etc improves. Nevertheless marker choice is still a non-trivial component of any population study, as is optimization of these markers for routine amplification in a species of interest. There are a number of approaches one may take to obtain enough informative nuclear sequence data for a particular study. No one approach is perfect, and all the options should be considered on a case-by-case basis. In the present case, the objective of identifying nuclear markers in Adélie penguins is primarily to address population genetic questions and to recover nuclear DNA from ancient Adélie populations (Chapters Five and Six). In this study, more time and effort was expended for the optimization of PCR amplification conditions for adenylate kinase intron 5 and myoglobin intron 2 than for the remaining 24 primer pairs tested together. However, the initial gradient PCR of

AK1i5 and *Myo2* gave good results for the first and inadequate results for the second. All further testing served to confirm this result and thus lend support to the assertion that devoting more resources to optimization of particular loci over others is not a cost-effective way to recover intron markers (CREER 2007).

The PCR screen of the 26 primer pairs yielded a 65.4% amplification success rate (percentage of markers for which a product of the correct size was amplified). This result is within reported success rates from other studies screening nuclear markers, e.g. (SPINKS *et al*. 2010) who reported a 59% success rate for avian, squamate and mammalian nuclear markers tested in turtles. Two other studies reported a 58% (IGEA *et al*. 2010) and 73% (BACKSTRÖM *et al*. 2008) amplification success rate.

Markers for which a single product of the correct size was amplified with no non-specific amplification (eleven of the twenty-six loci) in Adélie penguins were sequenced, and six of these produced high quality target sequence (23% overall success rate). A similar overall success rate (30.7%) was reported for a screen of intron markers in the kelp gull, *Larus dominicanus* (DE MENDONÇA DANTAS *et al*. 2009). This is a successful result, especially considering the large phylogenetic distance between chicken or passerine species (from which the markers were developed) and penguins. This success rate could easily be improved if necessary. In order to save time and resources, primer pairs were grouped according to melting temperatures and subjected to gradient PCRs testing three annealing temperatures and two concentrations of $MgCl_2$. Testing a slightly higher [$MgCl_2$] or a wider range of annealing temperatures would in all likelihood add a number of markers to the successful list. Simply by using a larger number of DNA extracts to test each marker further positive results could be obtained. Due to the large number of markers and conditions tested one DNA extract was not used for all and the possibility exists that some of the failures were caused by a degraded extract (possibly the case for primer pairs 3 and 7, Table 4.4)

Five out of the eleven markers that performed the best in PCR screens did not produce high quality sequence. Further efforts with these markers would very likely give better results. They should be considered as potential markers to work on if the set of

six markers needed to be expanded, in particular markers *CHDZ* intron 24 and *VLDLR* intron 9, both located on the Z chromosome (Table 4.4).

Two out of the three markers tested that had previously been used in penguin species (*AK1i5*, *Spheniscus mendiculus* (SHAPIRO and DUMBACHER 2001); *ODC6*, *Spheniscus humboldti* and *S. demersus (ERICSON et al. 2006)*) both amplified and sequenced well. It was surprising that myoglobin intron 2 did not prove successful in the Adélie penguin, as it was previously amplified from two penguin species, *Spheniscus humboldti* and *Eudyptula minor* (ERICSON *et al.* 2006). These species, though they do not belong to the same genus as the Adélie penguin (*Pygoscelis*), belong to different Genera themselves. However, this illustrates the appropriateness of the approach used in this study. Testing an array of markers, including a majority not previously used for Adélie penguins, produced more successful results than if I had focused solely on those markers that had been used previously in penguins.

The majority of the markers that proved successful in this study have been used for phylogenetic and population genetic studies in different avian species. Adenylate kinase intron 5 has been widely used in phylogenetics, for example in studies of New World jays (Corvidae) (BONACCORSO and TOWNSEND PETERSON 2007), aquiline eagles (Accipitriformes) (HELBIG *et al.* 2005), and *Pitohui* species (Pachycephalidae) (DUMBACHER *et al.* 2008). Ornithine decarboxylase introns 6-7 have been used in phylogenetics of New World orioles (*Icterus*) (ALLEN and OMLAND 2003), and Passerida (JOHANSSON *et al.* 2008) among others. Primers for markers Myelin proteolipid protein intron 4 (*MPP4*) and Beta-actin intron 2 (*ACTB2*) were designed for vertebrates in general from chicken, human and mouse resources (among others) and were originally tested in marbled murrelets (*Brachyramphus marmoratus*) and other, non-avian, vertebrates (SHAPIRO and DUMBACHER 2001). *MPP4* has been used as a marker for elucidating population genetic structure in the seabird *Larus dominicanus*, along with markers *AK1i5* and *ODC6* among others, (DE MENDONÇA DANTAS *et al.* 2009) and for phylogenetics of the genus *Oxyura* (Anatidae) (MCCRACKEN and SORENSON 2005). *ACTB2* has been applied to population genetics of the greater prairie chicken (*Tympanuchus cupido*) (ROSS *et al.* 2006). This marker has also been utilized for phylogenetics of jungle fowl *Gallus g. gallus* and *Gallus*

varius (SAWAI *et al*. 2010). Primers for markers *HMG2* and *UCHL3* were designed for avian taxa by comparing chicken and zebra finch genomes, and were tested across five bird species (BACKSTRÖM *et al*. 2008), and have not yet been applied to avian phylogenetics or population genetics.

In total, 3015 bp of nuclear sequence (primers trimmed) was generated for a variable number of Adélie penguin individuals, and 42 SNPs were detected – roughly 1 SNP/72bp. This result is high compared to other studies. For example, (BACKSTRÖM *et al*. 2008) found one SNP per 130bp in collared flycatcher (*Ficedula albicollis*), and (AITKEN *et al*. 2004) found one per 400bp in chimpanzees. However, within the kelp gull *Larus dominicanus* 1 SNP/78bp was reported (N=12 introns) (DE MENDONÇA DANTAS *et al*. 2009). This species is closer phylogenetically to penguins than the taxa from the other studies. Two loci also contained indel polymorphisms, which complicated direct sequence electropherogram reading due to length variation of heterozygotes. Within adenylate kinase intron 5 there is a 5bp repeat that is present either twice or three times in the Adélie penguin, and is present twice in *S. mendiculus*. *AK1i5* presented 13 SNPs ($N_{samples}$ = 21) in this study, fewer than those found in the kelp gull for the same marker (*Larus dominicanus*) (18 SNPs, N_{seq} = 24), though in the kelp gull 890bp were sequenced compared to 521bp in this study (DE MENDONÇA DANTAS *et al*. 2009). As a result SNP density is similar: 1 SNP per 40 bp in Adélie penguins, and 1/50 in kelp gulls. Some short repeat regions within the intron complicated the resolution of the intron alleles further, necessitating cloning. For marker *UCHL3* (Ubiquitin carboxyl-terminal esterase L3), two separate indels were found (1 bp and 2 bp). Sequence reads were able to be resolved in all cases using the Indelligent online software (DMITRIEV and RAKITOV 2008). The other four markers showed no evidence of indels, though a larger sample size for some of the markers may reveal more polymorphisms in general. *ODC6* is likely to be more variable; 11 SNPs (N_{seq}=38) have been reported in the kelp gull, *L. dominicanus*. In the same study *MPP4* presented 11 SNPs (N_{seq}=48) (DE MENDONÇA DANTAS *et al*. 2009), compared to 8 SNPs ($N_{samples}$=18) in Adélie penguins found during this study. This lower diversity may be a result of sample size. Full characterization of variability of these markers was not the main objective of this study; further sequencing for several of these markers is carried out in Chapter Five. The screening

carried out in this study indicates markers *MPP4* (1 SNP/37 bp) and *AK1i5* (1 SNP/40 bp) are the most variable and potentially the most amenable to population genetic studies. Markers *ACTB2* (1 SNP/56 bp), *HMG2* (1 SNP/86 bp) and *UCHL3* (1 SNP/137 bp) are slightly less variable. Further sequencing of *ODC6* is required to determine its variability.

Choosing to screen a larger number of intron primer pairs rather than concentrating solely on those that had previously been amplified in penguins proved to be an effective and relatively rapid way of recovering nuclear loci for Adélie penguins. Not all the primer pairs amplifying introns in penguins enabled amplification in Adélie penguins (myoglobin intron 2), and a number that had not been amplified from penguin species were found to work. Hence this approach is a relatively cost-effective one for identifying single copy nuclear loci for population genetic work. This technique successfully identified six markers for use in Adélie penguins, potentially variable enough for population genetics. Another five or more markers of those screened in this study could be recovered from Adélie penguins with a little further optimization effort.

The rise of second-generation sequencing technologies is quickly making it more feasible to obtain genomic data for any species. Second-generation sequencing has now been used for microsatellite marker development (ABDELKRIM *et al*. 2009; ALLENTOFT *et al*. 2009) and for identification of anonymous nuclear markers (THOMSON *et al*. 2010). It is rapidly becoming easier to develop one's own suite of nuclear markers. Still, the cost of marker development will not be within the reach of most research groups for some time, and alternative, less costly approaches will continue to be important for marker selection.

4.6 References

ABDELKRIM, J., B. C. ROBERTSON, J.-A. L. STANTON and N. J. GEMMELL, 2009 Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques **46:** 185-192.

AITKEN, N., S. SMITH, C. SCHWARZ and P. A. MORIN, 2004 Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. Molecular Ecology **13:** 1423-1431.

ALLEN, E. S., and K. E. OMLAND, 2003 Novel Intron Phylogeny Supports Plumage Convergence in Orioles (*Icterus*). The Auk **4:** 961-969.

ALLENDORF, F. W., P. A. HOHENLOHE and G. LUIKART, 2010 Genomics and the future of conservation genetics. Nat Rev Genet **11:** 697-709.

ALLENTOFT, M. E., S. C. SCHUSTER, R. N. HOLDAWAY, M. L. HALE, E. MCLAY *et al*., 2009 Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. BioTechniques **46:** 195-200.

ARMSTRONG, M. H., E. L. BRAUN and R. T. KIMBALL, 2001 Phylogenetic Utility of Avian Ovomucoid Intron G: A Comparison of Nuclear and Mitochondrial Phylogenies in Galliformes. The Auk **118:** 799-804.

AXELSSON, E., N. G. C. SMITH, H. SUNDSTRÖM, S. BERLIN and H. ELLEGREN, 2004 Male-biased Mutation Rate and Divergence in Autosomal, Z-Linked and W-Linked Introns of Chicken and Turkey. Molecular Biology and Evolution **21:** 1538-1547.

BACKSTRÖM, N., S. FAGERBERG and H. ELLEGREN, 2008 Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Molecular Ecology **7:** 964-980.

BALLOUX, F., 2010 The worm in the fruit of the mitochondrial DNA tree. Heredity **104:** 419-420.

BIERNE, N., S. A. LEHNERT, E. BÉDIER, F. BONHOMME and S. S. MOORE, 2000 Screening for intron-length polymorphisms in penaeid shrimps using exon-primed intron-crossing (EPIC)-PCR. Molecular Ecology **9:** 233-235.

BONACCORSO, E., and A. TOWNSEND PETERSON, 2007 A multilocus phylogeny of New World jay genera. Molecular Phylogenetics and Evolution **42:** 467-476.

BORGE, T., M. T. WEBSTER, G. ANDERSSON and G.-P. SAETRE, 2005 Contrasting Patterns of Polymorphism and Divergence on the Z Chromosome and Autosomes in Two Ficedula Flycatcher Species. Genetics **171:** 1861-1873.

BRITO, P., and S. EDWARDS, 2009 Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica **135:** 439-455.

CHENUIL, A., T. HOAREAU, E. EGEA, G. PENANT, C. ROCHER *et al*., 2010 An efficient method to find potentially universal population genetic markers, applied to metazoans. BMC Evolutionary Biology **10:** 276.

CREER, S., 2007 Choosing and using introns in molecular phylogenetics. Evolutionary Bioinformatics **3:** 99-108.

DE MENDONÇA DANTAS, G. P., R. GODINHO, J. S. MORGANTE and N. FERRAND, 2009 Development of new nuclear markers and characterization of single nucleotide polymorphisms in kelp gull (*Larus dominicanus*). Molecular Ecology Resources **9:** 1159-1161.

DMITRIEV, D. A., and R. A. RAKITOV, 2008 Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels. PLoS Computational Biology **4:** e1000113.

DUMBACHER, J. P., K. DEINER, L. THOMPSON and R. C. FLEISCHER, 2008 Phylogeny of the avian genus *Pitohui* and the evolution of toxicity in birds. Molecular Phylogenetics and Evolution **49:** 774-781.

ERICSON, P. G. P., C. L. ANDERSON, T. BRITTON, A. ELZANOWSKI, U. S. JOHANSSON *et al*., 2006 Diversification of Neoaves: integration of molecular sequence data and fossils. Biology Letters **2:** 543-547.

FAIN, M. G., and P. HOUDE, 2004 Parallel Radiations in the Primary Clades of Birds. Evolution **58:** 2558-2573.

FAIN, M. G., C. KRAJEWSKI and P. HOUDE, 2007 Phylogeny of "core Gruiformes" (Aves: Grues) and resolution of the Limpkin-Sungrebe problem. Molecular Phylogenetics and Evolution **43:** 15.

FEDOROV, A., A. F. MERICAN and W. GILBERT, 2002 Large-scale comparison of intron positions among animal, plant, and fungal genes. PNAS **99:** 16128-16133.

FLOT, J.-F., 2007 CHAMPURU 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Molecular Ecology Notes **7:** 974-977.

FRIDOLFSSON, A.-K., and H. ELLEGREN, 1999 A simple and universal method for molecular sexing of non-ratite birds. Journal of Avian Biology **30:** 116-121.

FRIESEN, V. L., B. C. CONGDON, M. G. KIDD and T. P. BIRT, 1999 Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. Molecular Ecology **8:** 2147-2149.

HACKETT, S. J., R. T. KIMBALL, S. REDDY, R. C. K. BOWIE, E. L. BRAUN *et al.*, 2008 A Phylogenomic Study of Birds Reveals Their Evolutionary History. Science **320:** 1763-1768.

HARE, M. P., 2001 Prospects for nuclear gene phylogeography. Trends in Ecology and Evolution **15:** 700-706.

HASSAN, M., C. LEMAIRE, C. FAUVELOT and F. BONHOMME, 2002 Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. Molecular Ecology Notes **2:** 334-340.

HELBIG, A. J., A. KOCUM, I. SEIBOLD and M. J. BRAUN, 2005 A multi-gene phylogeny of aquiline eagles (Aves: Accipitriformes) reveals extensive paraphyly at the genus level. Molecular Phylogenetics and Evolution **35:** 147-164.

IGEA, J., J. JUSTE and J. CASTRESANA, 2010 Novel intron markers to study the phylogeny of closely related mammalian species. BMC Evolutionary Biology **10:** 369.

JENNINGS, W. B., and S. V. EDWARDS, 2005 Speciational History of Australian Grass Finches (*Poephila*) inferred from Thirty Gene Trees. Evolution **59:** 2033-2047.

JOHANSSON, M. L., 2009 Next Generation Sequencing in Nonmodel Organisms: Has the Future Arrived? J Hered **100:** 807-.

JOHANSSON, U. S., and P. G. P. ERICSON, 2004 A re-evaluation of basal phylogenetic relationships within trogons (Aves: Trogonidae) based on nuclear DNA sequences. JZS **43:** 166-173.

JOHANSSON, U. S., J. FJELDSÂ and R. C. K. BOWIE, 2008 Phylogenetic relationships within Passerida (Aves: Passeriformes): A review and a new molecular phylogeny based on three nuclear intron markers. Molecular Phylogenetics and Evolution **48:** 858-876.

LEE, J. Y., S. V. EDWARDS and M. WEBSTER, 2009 Divergence Across Australia's Carpentarian Barrier: Statistical Phylogeography of the Red-Backed Fairy Wren (Malurus melanocephalus). Evolution **62:** 3117-3134.

LI, C., J.-J. RIETHOVEN and L. MA, 2010 Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. BMC Evolutionary Biology **10:** 90.

LOVETTE, I. J., and D. R. RUBINSTEIN, 2007 A comprehensive molecular phylogeny of the starlings (Aves: Sturnidae) and mockingbirds (Aves: Mimidae):

Congruent mtDNA and nuclear trees for a cosmopolitan avian radiation. Molecular Phylogenetics and Evolution **44:** 1031-1056.

LYONS, L. A., T. F. LAUGHLIN, N. G. COPELAND, N. A. JENKINS, J. E. WOMACK *et al.*, 1997 Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nature Genetics **15:** 47-56.

MCCRACKEN, K. G., and M. D. SORENSON, 2005 Is Homoplasy or Lineage Sorting the Source of Incongruent mtDNA and Nuclear Gene Trees in the Stiff-Tailed Ducks (Nomonyx-Oxyura)? Systematic Biology **54:** 35-55.

MEIKLEJOHN, C. D., K. L. MONTOOTH and D. M. RAND, 2007 Positive and negative selection on the mitochondrial genome. Trends in Genetics **23:** 5.

OUBORG, N. J., C. PERTOLDI, V. LOESCHCKE, R. BIJLSMA and P. W. HEDRICK, 2010 Conservation genetics in transition to conservation genomics. Trends in Genetics **26:** 177-187.

PALUMBI, S. R., 1995 Nucleic acids II: the polymerase chain reaction. , pp. 205–247 in *Molecular Systematics, 2nd edn*, edited by M. C. HILLIS D, Sinauer, Sunderland, MA.

PRIMMER, C. R., T. BORGE, J. LINDELL and G.-P. SAETRE, 2002 Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. Molecular Ecology **11:** 603-612.

ROGOZIN, I. B., Y. I. WOLF, A. V. SOROKIN, B. G. MIRKIN and E. V. KOONIN, 2003 Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Current Biology **13:** 1512-1517.

ROSS, J., A. ARNDT, R. SMITH, J. JOHNSON and J. BOUZAT, 2006 Re-examination of the historical range of the greater prairie chicken using provenance data and DNA analysis of museum collections. Conservation Genetics **7:** 735-751-751.

RUBINOFF, D., and B. S. HOLLAND, 2005 Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. Syst. Biol. **54:** 952-961.

SAMBROOK, J., E. F. FRITSCH and T. MANIATUS, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York.

SAWAI, H., H. L. KIM, K. KUNO, S. SUZUKI, H. GOTOH *et al.*, 2010 The Origin and Genetic Variation of Domestic Chickens with Special Reference to Junglefowls *Gallus g. gallus* and *G. varius*. PLoS ONE **5:** e10639.

SCIENTISTS, G. K. C. O., 2009 Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,Äâ000 Vertebrate Species. Journal of Heredity **100:** 659-674.

SEUTIN, G., B. N. WHITE and P. T. BOAG, 1991 Preservation of avian blood and tissue samples for DNA analyses. Canadian Journal of Zoology **69:** 82-90.

SHAPIRO, L. H., and J. P. DUMBACHER, 2001 Adenylate Kinase Intron 5: A New Nuclear Locus for Avian Systematics. The Auk **118:** 248-255.

SLADE, R. W., C. MORITZ, A. HEIDEMAN and P. T. HALE, 1993 Rapid assessment of single-copy nuclear DNA variation in diverse species. Molecular Ecology **2:** 359-373.

SORENSON, M. D., J. C. AST, D. E. DIMCHEFF, T. YURI and D. P. MINDELL, 1999 Primers for a PCR-Based Approach to Mitochondrial Genome Sequencing in Birds and Other Vertebrates. Molecular Phylogenetics and Evolution **12:** 105-114.

SPINKS, P., R. THOMSON, A. BARLEY, C. NEWMAN and H. BRADLEY SHAFFER, 2010 Testing avian, squamate, and mammalian nuclear markers for cross amplification in turtles. Conservation Genetics Resources **2:** 127-129.

SUNDSTRÖM, H., M. T. WEBSTER and H. ELLEGREN, 2003 Is the Rate of Insertion and Deletion Mutation Male Biased?: Molecular Evolutionary ANalysis of Avian and Primate Sex Chromosome Sequences. Genetics **164:** 259-268.

THOMSON, R. C., I. J. WANG and J. R. JOHNSON, 2010 Genome-enabled development of DNA markers for ecology, evolution and conservation. Molecular Ecology **19:** 2184-2195.

VAN TUINEN, M., D. B. BUTVILL, J. A. W. KIRSCH and S. B. HEDGES, 2001 Convergence and divergence in the evolution of aquatic birds. Proceedings of the Royal Society of London Series B **268:** 1345-1350.

WALTARI, E., and S. V. EDWARDS, 2002 Evolutionary Dynamics of Intron Size, Genome Size, and Physiological Correlates in Archosaurs. The American Naturalist **160:** 539-552.

ZHANG, D.-X., and G. M. HEWITT, 2003 Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Molecular Ecology **12:** 563-584.

5 Chapter Five

USING INTRONS TO ELUCIDATE ADÉLIE PENGUIN POPULATION HISTORY AND PENGUIN PHYLOGENY

5.1 Abstract

Adélie penguin (*Pygoscelis adeliae*) demographic history has been well-studied using mitochondrial markers. Population inference based on multiple loci allows independent estimates of demographic history. Nuclear intron markers have the potential to contribute to our understanding of Adélie penguins, and so far they have not been used in this context. Mitochondrial *HVRI* analyses of Adélie penguins previously revealed the presence of two distinct monophyletic lineages, called Antarctic (A) and Ross Sea (RS). Adélie penguin individuals, previously identified as belonging to either the A or RS mitochondrial lineages, were sequenced for six nuclear introns (myelin proteolipid protein intron 4, ubiquitin carboxyl-terminase esterase L3 intron 5, adenylate kinase intron 5, high mobility group intron 2, ornithine decarboxylase introns 6-8). Statistical parsimony haplotype networks were constructed, and A and RS samples were distributed randomly throughout the networks, indicating the lineages have not been maintained in the nuclear genome, and that breeding has occurred freely between the two. Neutrality indices and mismatch distribution analyses revealed a population expansion signature for 4/5 loci, consistent with previous findings from mtDNA analyses. The utility of four of the introns (*UCHL3*, *MPP4*, *AK1i5*, *ODC6*) in resolving penguin phylogenetic signals was also ascertained. Individually, none of the markers contained enough phylogenetic signal to resolve relationships among penguin species; a concatenated dataset (1926 bp incl. gaps), however, recovered the majority of splits between penguin species with significant support.

5.2 Introduction

Population and phylogenetic studies of non-model organisms are increasingly moving past the point of relying solely on mitochondrial DNA markers. Additional unlinked loci are now being incorporated to provide independent or combined insights into the evolutionary relationships between populations and species. Nuclear markers offer distinct, independent genealogies which allow us to fully address the population history of organisms (BALLARD and WHITLOCK 2004) at different time scales due to differences in mode of inheritance, effective population size and mutation rate for nuclear and mitochondrial DNA markers. Multi-locus approaches are increasingly used due to two principal characteristics of single-locus studies. Firstly, mutation and genetic drift are stochastic, and can create variable signatures in DNA even if different loci have experienced identical population histories (EDWARDS and BEERLI 2000; KNOWLES and MADDISON 2002; ROSENBERG and NORDBORG 2002). Single-locus studies, secondly, do not address the possibility that selection, rather than population history has generated the observed patterns in DNA (BAZIN *et al*. 2006; BENSCH *et al*. 2006). A signature of a population expansion, for example, is similar to that of a selective sweep, whereby variation is reduced at loci closely linked to a locus undergoing selection. Mutation, drift and selection affect unlinked loci independently, so using multiple loci can strengthen inferences about population history and help distinguish between these effects.

Nuclear introns have not, as yet, been widely used for any penguin species. Previous work carried out as part of this thesis (chapter four) identified a number of intron markers that could prove suitable for intra- and inter-specific studies of Adélie penguins, and potentially other penguin species. The utility of these introns for addressing Adélie penguin population genetics and to resolve the phylogenetic relationships among penguin species is assessed in this chapter.

Mitochondrial *HVRI* analyses of Adélie penguin populations have indicated that the A lineage has been recorded from colonies of Adélie penguins around the continent, while the RS lineage is found only in the Ross Sea area. It has been hypothesized that these two lineages originated due to isolation of Adélie penguins in separate refugia

during the last glacial maximum (75 kya, 32-122 kya) (RITCHIE *et al*. 2004). The presence of these lineages has not yet been assessed using nuclear intron data. The presence or absence of these lineages within nuclear introns will contribute to our understanding of the duration of the divergence within Adélie penguins, thanks to the longer coalescence time of nuclear markers compared to mitochondrial. In the present study nuclear intron sequences for five introns were sequenced for samples belonging to ten colonies distributed throughout the Ross Sea area of Antarctica. These were previously sequenced for the *HVRI* region of the mitochondrial genome and scored as belonging to the A or RS lineages (Chapter 2, Lambert *et al* (2002), Ritchie *et al* (2004)).

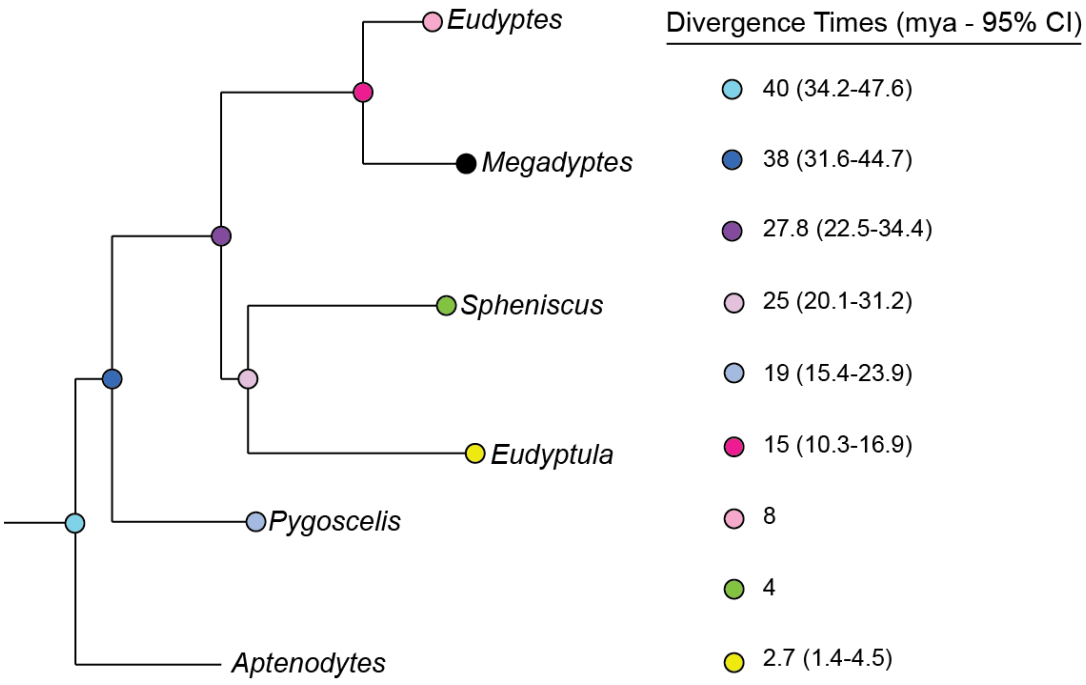


Figure 5.1 Phylogenetic representation of extant penguin genera and divergence times.

Divergence times are illustrated for each node (divergence times of major splits) and for most tips (divergence times within genera). All times are taken from (BAKER *et al*. 2006). A molecular and fossil phylogeny places the time to the most recent common ancestor of extant penguins (Spheniscidae) at 15Mya rather than 40Mya (CLARKE *et al*. 2007).

A number of recent studies have shown that, when compared to the more commonly-used mitochondrial genes and nuclear protein-coding genes, nuclear non-coding introns can provide equally useful phylogenetic characters, due to their lack of

functional constraints, high substitution rate and typically lower levels of homoplasy (CREER *et al*. 2006; FRIESEN 2000; FRIESEN *et al*. 1997). A well-supported phylogeny of extant penguins (Sphenisciformes: Spheniscidae) has been constructed using 5851bp of mitochondrial and nuclear DNA and most probable divergence dates for extant penguins estimated (Fig. 5.1) (BAKER *et al*. 2006) (but see (CLARKE *et al*. 2007) for fossil-supported divergence times). A nuclear exon of the *RAG-1* gene was used (2802bp); however no other nuclear regions were included. In the present study, four presumably unlinked intron loci were used to offer a complementary view of the penguin phylogeny.

5.3 Methods

5.3.1 DNA extractions

Genomic DNA was extracted from 40-100 μl of blood preserved in buffer (SEUTIN *et al*. 1991), following the standard phenol-chloroform protocol (SAMBROOK *et al*. 1989), and diluted to 50 ng/μl in water for downstream work.

The Adélie penguin blood samples used originated from ten colonies from the Ross Sea area of Antarctica, collected by D. M. Lambert and colleagues (Table 5.1). The samples selected had been used previously for mtDNA *HVRI* Adélie penguin population analyses and thus could be identified as belonging to either the "A" or "RS" mitochondrial DNA lineages. Twenty-three individuals were extracted, twelve of which belonged to the A lineage, and 11 to the RS.

Blood samples or DNA extracts originating from eleven of the seventeen other penguin species were provided by collaborators C. D. Millar (University of Auckland) or A. J. Baker (Royal Ontario Museum) (Table 5.2).

Table 5.1 Adélie penguin sample provenance, together with mtDNA lineage and intron haplotypes

COLONY	SAMPLE ID	MtDNA lineage	*MPP4*	*UCHL3*	*ODC6*	*HMG2*	*AK1i5*[*]
Cape Hallett	CM14	A	1 & 11	1	1	3 & 1	8 & 9[+]
Inexpressible Island	T18	A	1	1 & 6[+]	1	1 & 9[+]	1 & 5
Port Martin	T311	A	1 & 2	1 & 8	1	1 & 8	25 & 26[+]
Cape Wheatstone	CM39	A	1 & 9	1 & 6[+]	1 & 5[++]	1 & 6	3 & 7
Port Martin	T315	A	1 & 3	2 & 9[++]	1	1	13
Edmonson Point	T162	A	1 & 8	-	1	1	2 & 23
Adélie Cove	T67	A	1	1 & 13[++]	1	5	1
Sabrina Islet	T347	A	-	-	-	2	5 & 29[+]
Adélie Cove	T69	A	4 & 7	-	1 & 2	7 & 10	16
Duke of York Island	RT02	A	1 & 2	1 & 6[+]	1 & 3	1 & 4	-
Sabrina Islet	T340	A	1 & 3	1 & 11[++]	1	3	27 & 28[+]
Edmonson Point	T163	A	1 & 8	1 & 8	-	-	4 & 12
Port Martin	T302	RS	1	3 & 8	1 & 6[++]	1 & 10	3 & 24[+]
Inexpressible Island	T02	RS	-	1 & 10[++]	-	1	11
Cape Wheatstone	CM42	RS	1 & 2	4 & 6[+]	1	1 & 4	2 & 10
Franklin Island	CM60	RS	1 & 3	6	1 & 11[+]	1	19 & 20[+]
Franklin Island	CM56	RS	1 & 5	1 & 5	8 & 9	1 & 2	17 & 18[+]
Inexpressible Island	T15	RS	1 & 3	-	1	1 & 7	21 & 22
Cape Adare	T431	RS	1 & 2	1	-	1 & 2	-
Adélie Cove	T64	RS	-	1 & 6[+]	8 & 10[++]	1	14 & 15
Sabrina Islet	T339	RS	1 & 6	1	1	1	1 & 4[+]
Cape Wheatstone	CM15	RS	1 & 10	7 & 12[++]	4 & 7	1	6

Intron haplotype numbers refer to those detailed in Figures 5.2 and 5.3. [*]*AK1i5* samples marked in red and underlined are those heterozygotes that were not unambiguously resolved in PHASE. For CM39, CM42 and T163, $p<0.85$. For CM14 and T64, one position could not be resolved. [+]Sample was a length-variant heterozygote (LVH). [++]Sample was a LVH presenting an indel at a different position.

Table 5.2 Penguin species sequenced for four introns

Penguin Species	Common Name	SAMPLE ID	*MPP4*	*UCHL3*	*ODC6*	*AK1i5*
Eudyptula minor	Little Blue	Tax 96-41[+]	P	H	P, LVH	P[1]
Eudyptes pachrhynchus	Fiordland Crested	Tax 96-18[+]	P	P[1]	P	H
Eudyptes sclateri	Erect Crested	LB 12718[+]	P	H	P, LVH	LVH[2]
Eudyptes chrysocome	Rockhopper	MV2258[+]	H	H	H	H
Eudyptes schlegeli	Royal	R13F[*]	H	P, LVH	P	P[1]
Megadyptes antipodes	Yellow-eyed	YEP-1[+]	H	H	P, LVH	H
Aptenodytes patagonicus	King	AP26[+]	P	LVH[2]	P[1]	P
Aptenodytes forsteri	Emperor	K33[+]	P[1]	P, LVH	P	P, LVH
Spheniscus demersus	Black-footed	JAP9[*]	H	P, LVH	H	P
Spheniscus demersus	Black-footed	EF552742.1[★]	-	-	P[1]	-
Spheniscus humboldti	Humboldt's	DQ881738.1[★]	-	-	H	-
Spheniscus magellanicus	Magellanic	MF1[*]	H	P	H	P
Spheniscus mendiculus	Galapagos	AF307894.1[★]	-	-	-	H
Pygoscelis antarcticus	Chinstrap	T361[*]	P	H	P, LVH	P
Pygoscelis adeliae	Adélie	T67[❖]	H1	H1, H13, LVH	H1	H1

[+]Samples obtained from C. D. Millar. [*]Samples obtained from A. J. Baker. [★]Obtained from NCBI, accession number given. T67[❖] = sample sequenced and used in the present intra-specific Adélie penguin study, haplotype number indicated for each intron. H = homozygote. P = heterozygote, unambiguously phased. P[1] = heterozygote, ambiguously phased. LVH = length variant heterozygote. LVH[2] = length variant heterozygote, unresolved. H1 = Adélie penguin intron haplotype 1. H13 = Adélie penguin haplotype 13.

5.3.2 PCR and direct sequencing

All DNA extracts were quantified using Nanodrop and 50 ng/μl working dilutions prepared. All samples were PCR-amplified for introns of four genes, adenylate kinase intron 5 (*AK1i5*), myelin proteolipid protein intron 4 (*MPP4*), ornithine decarboxylase introns 6 and 7 (*ODC6*), and ubiquitin carboxyl-terminase esterase L3 intron 5 (*UCHL3*). Primers used for PCR and sequencing of these introns were published previously for *MPP4*, *ODC6* (FRIESEN *et al*. 1999), and *UCHL3* (BACKSTRÖM *et al*. 2008). Primers AklongF and AklongR (designed and described in Chapter four of this thesis) were used for *AK1i5*. Two other introns were also PCR-amplified for Adélie penguins exclusively: high mobility group protein B2 intron 4 (*HMG-2*) and beta-actin intron 2 (*ACTB2*). Primers for these two introns were

obtained from the literature, *ACTB2* (BORGE *et al*. 2005) and *HMG-2* (BACKSTRÖM *et al*. 2008). Marker *ACTB2* was not sequenced due to resource constraints. Previous work recovering introns in Adélie penguins (Chapter Four) identified the best conditions for PCR amplification of each intron marker. [$MgCl_2$] (Invitrogen) for primers Aklong and *UCHL3* was 2.2mM; for primers *MPP* and *ODC6* it was 2mM. Standard PCR conditions were the following: dNTPs 0.2mM, primers 0.4μM, platinum Taq 0.5units, 1x buffer (all Invitrogen), approximately 150ng of DNA, and water to 25μl. To expedite PCR, Adélie penguin introns were amplified using a two-step touchdown program (5' 94°C; 20 x [40" 94°C, 40" 60°C, 1' 72°C]; 20 x [40" 94°C, 40"56°C, 1' 72°C]; 5' 72°C).

Introns were amplified for other penguin species using the same touchdown program. For sample/intron combinations that did not amplify during a first attempt, the second annealing temperature was dropped to 55ºC, the primer concentration was increased to 0.5μM, and the DNA concentration was increased from 150ng to 200ng per reaction. The $MgCl_2$ concentration was also slightly increased, from 2 to 2.1mM or 2.2 to 2.3 mM depending on the intron (see above).

PCR products were size fractioned on a 2% agarose gel to confirm amplification. They were purified using Agencourt AMPure XP magnetic beads following the manufacturer's protocol, and were quantified on 2% agarose gels using Invitrogen Low Mass Ladder as a standard. All products were sequenced in two directions, using the forward and reverse primers employed for PCR amplification, at the Griffith University Sequencing Service.

5.3.3 Sequence analysis and phasing of introns

Sequences were edited in GENEIOUS PRO 5.0.4 (DRUMMOND *et al*. 2010), and forward and reverse sequences were aligned in order to resolve sequence ambiguities arising either from error or from the presence of two alleles (single-nucleotide or length-variant heterozygotes). Length-variant heterozygotes resulting from insertion-deletion polymorphisms were phased using the program INDELLIGENT, which disentangles superimposed allelic sequences and identifies the insertion or deletion (DMITRIEV and

RAKITOV 2008). Heterozygotes containing one ambiguous position were unambiguously phased manually into haplotypes. Sequences with more than one heterozygote position present were subjected to analysis to statistically separate haplotypes based on probability, using the program PHASE (STEPHENS *et al*. 2001). Input files for PHASE were prepared using SeqPHASE (FLOT 2010). The most probable haplotypes for each DNA sequence were selected for further analysis.

5.3.4 Adélie population genetic analysis

Best-fit models of sequence evolution were selected using the modeltest function implemented in MEGA 5.04 (TAMURA *et al*. 2011) for each intron alignment, based on the Akaike information criterion (AIC) (AKAIKE 1974), as well as transition – transversion ratios. For four introns, the Tamura-3-parameter substitution model (T92) was the best fit (TAMURA 1992); for AK1i5 the Tamura-Nei model (T93) provided the best fit (TAMURA and NEI 1993), although a subset of unambiguously phased AK1i5 sequences was characterized by a best fit to T92 (Table 5.3).

A number of statistics were determined using the program ARLEQUIN 3.5.1.2 (default parameters, except for the substitution models and transition – transversion ratios found in MEGA 5.04) (EXCOFFIER and LISCHER 2010). PGDSpider 2.0.0.2 was used to generate the input files for each intron alignment. Genetic diversity measures were estimated (nucleotide diversity and mean number of pairwise differences among sequences) and their standard deviations. Different factors, such as population growth, selective sweeps, and background selection could affect patterns of DNA polymorphism. Two neutrality tests (Tajima's *D* statistic (TAJIMA 1989b) and Fu's *Fs* (FU 1997)) were used to test whether the observed polymorphism fit with neutral model expectations. Tajima's *D* describes the scaled difference between the estimate of θ ($4N_e\mu$; N_e = effective population size, μ = mutation rate per generation) based on the average number of pairwise differences (π) and the number of segregating sites (S) in a DNA sequence. Departures from the null model can be caused by a number of factors. Population size changes, selection (either direct or through linkage to a locus under selection) can lead to an excess of low frequency variants and significant

negative values of *D*. Population subdivision, balancing selection or recent bottlenecks can cause an excess of intermediate frequency variants, leading to positive *D* values (FAY and WU 1999; NIELSEN and WAKELEY 2001; SIMONSEN *et al*. 1995; TAJIMA 1989a; TAJIMA 1993). Fu's *Fs*, based on the haplotype distribution, was also estimated. Negative *Fs* values differing significantly from zero are indicative of population demographic expansions. Fu and Li's F^* and D^* (FU and LI 1993) were also calculated in DnaSP 5.10.01 (LIBRADO and ROZAS 2009). A range expansion is suggested when Tajima's *D* and Fu's *Fs* are significantly negative, while F^* and D^* are not (FU 1997).

To test for historical population expansion events in Adélie penguins, the observed frequency distribution of pairwise nucleotide differences among individuals (mismatch distribution (ROGERS and HARPENDING 1992)) was compared to expected distributions from a sudden population expansion model using ARLEQUIN 3.5.1.2 (EXCOFFIER and LISCHER 2010). Populations at demographic equilibrium or in decline should in theory produce a multimodal distribution of pairwise differences. Populations that have experienced a sudden demographic expansion, on the other hand, should be characterized by a unimodal distribution of values (ROGERS and HARPENDING 1992). The method assumes a sudden population expansion from N_0 to N_1, t generations ago, followed by demographic equilibrium. The parameters θ_0=4 $N_0\mu$, θ_1=4$N_1\mu$, and τ=2μt are estimated (μ=locus mutation rate). Estimates of time since demographic expansion were obtained using the parameter τ and a mutation rate μ obtained for four of the five introns combined (*AK1i5*, *MPP4*, *UCHL3*, *ODC6*). Methods for obtaining a mutation rate are described below. One thousand parametric bootstrap replicates were used to generate expected distributions using the model of sudden demographic expansion (EXCOFFIER and LISCHER 2010) for each of the five intron alignments. *P* values were calculated as the proportion of simulations producing a larger sum-of-squared deviation (SSD) than the observed SSD. Harpending's raggedness index was also calculated and its significance determined. The raggedness index quantifies the smoothness of the observed mismatch distribution; small values are typical of expanding populations while higher values are observed among stationary or bottlenecked populations (HARPENDING 1994; HARPENDING *et al*. 1993).

The population parameter θ ($4N_e\mu$; N_e=effective population size, μ=mutation rate per generation) was estimated, simultaneously with the growth parameter g ($\theta_t=\theta_{modern}^{-gt}$) and the recombination parameter r (C/μ; C= rate of recombination per inter-site link per generation) for each locus, using the Bayesian method in the program LAMARC 2.1.6 (KUHNER 2006). An r value of 1 means recombination and mutation are equally likely to occur at a site, which indicates a relatively high recombination rate. The parameter g, if negative, indicates a population has been shrinking, while a positive value indicates population growth. Starting parameters for the first run used empirical base frequencies, empirical transition – transversion ratios, and default starting parameters (θ=0.01, g=1, r=0.01). Prior bounds for each parameter were logarithmic for θ (1 x 10^{-5} to 10) and r (1 x 10^{-5} to 10), and linear for g (-500 to 1000). 100,000 trees were sampled every 50 genealogies, and the first 200,000 genealogies were discarded as burn-in. Three chains were run simultaneously (heating; temperatures 1, 1.1 and 1.2). Convergence was assessed using the program TRACER 1.5 (http://beast.bio.ed.ac.uk/Tracer) to calculate the effective sample size (ESS). ESS values of 100-200 or greater indicate convergence has been achieved. A second and third run with an upper g prior of 2000, 1,00,000 trees sampled and heating temperatures of 1, 1.3 and 1.5 were carried out for several markers (*ODC6*, *MPP4* and *HMG2*). Markers *AK1i5* and *UCHL3* did not converge well and runs were prohibitively time-consuming. All estimates reported here are the median estimate for each parameter, combined over one to three runs, as well as 95% bayesian credibility intervals.

To visualize phylogenetic relationships among intron haplotypes and alleles, statistical parsimony haplotype networks were generated for each intron using the program TCS 1.2.1 (CLEMENT *et al*. 2000). For all loci, gaps were considered a 5th state. Pairwise Φ_{st} for each intron between "A" and "RS" lineage individuals were estimated and significance assessed. Reticulate networks were generated in SPLITSTREE 4 (HUSON and BRYANT 2006) and a Phi test for recombination was also run. Additionally, several recombination detection methods, incorporated into RDP3 (MARTIN *et al*. 2010), were implemented (default parameters, with different

substitution models and transition – transversion ratios as determined previously for each intron alignment).

5.3.5 Penguin intron phylogenetics

Intron sequences for the different penguin species included in this study (*AK1i5*, *UCHL3*, *MPP4* and *ODC6*) were edited as described above and aligned in GENEIOUS PRO 5.0.4 (DRUMMOND *et al*. 2010). Penguin sequences available from GenBank were included in the alignments. Alignments were trimmed to the length of the primer-trimmed Adélie penguin sequence used for each separate intron alignment (sample ID T67), and including some flanking coding sequence. Best-fit models of sequence evolution were selected using the modeltest function implemented in MEGA 5.04 (TAMURA *et al*. 2011) for each intron alignment, based on the Akaike information criterion (AIC) (AKAIKE 1974) as well as transition – transversion ratios. A number of sequences obtained were heterozygotes. Certain heterozygotes were phased unambiguously; others were phased arbitrarily. For the concatenated dataset (detailed below), only unambiguously phased heterozygotes were separated into haplotypes (including length-variant heterozygotes); heterozygotes with more than one variable position presented ambiguous bases at variable sites. Length-variant heterozygotes were phased as described above for the Adélie-only dataset. For all introns, the Tamura-3-parameter substitution model (T92) gave the best fit (TAMURA 1992).

A mutation rate for the four introns together was estimated, based on a concatenated dataset which included only non-coding sites, with gaps removed (1485 bp). An estimate was obtained using two methods. First, in MEGA, a UPGMA tree was generated (Tamura-Nei model with uniform rates and complete deletion) and the root to tip distance was estimated and divided by the divergence time to the most recent common penguin ancestor (Fig. 5.1) (BAKER *et al*. 2006), i.e. a t_{mrca} of 40Mya. These divergence times are disputed by a study based on fossil evidence, which suggests a much more recent divergence time for the Spheniscidae crown group of 15Mya minimum (CLARKE *et al*. 2007). Second, a mutation rate was also generated in BEAST

(DRUMMOND and RAMBAUT 2007); genealogies were calibrated for a t_{mrca} of 40Mya (35 – 45 Mya prior), or of 15 Mya (10 – 35 Mya uniform prior). The HKY + gamma + five rate categories model was used, as well as an uncorrelated log normal molecular clock and a Yule birth rate for the coalescent model. Convergence of the estimate was verified in TRACER as described for LAMARC estimates above. Separating the introns resulted in too much stochasticity and rate estimates were unreliable. Hence, only the combined result is reported.

Phylogenetic analyses were performed for each intron alignment (including coding regions) separately as well as in three combined datasets. Outgroup sequences were included in each dataset from the phylogenetically nearest sequence available on NCBI. The combined datasets were: all available intron sequences (no outgroup), *AK1i5* and *MPP4* (outgroup *Larus dominicanus*), and *AK1i5* and *ODC6* (two Passeriform outgroups). Phylogenetic analysis was performed using the maximum likelihood method implemented in PhyML and the Bayesian method implemented in MrBayes 3.1.2 (HUELSENBECK *et al*. 2001; RONQUIST and HUELSENBECK 2003). Maximum likelihood node support was evaluated by bootstrap analysis with 500 replicates. Bayesian analyses were conducted using uniform priors, random starting trees, and four simultaneous Markov chains run for 1,000,000 generations, with trees sampled every 100 generations and the first 2,500 generations were discarded as burn-in. Rate variation across sites was estimated using the "invgamma" option, in which a proportion of sites is invariable and the rate for the remaining sites is drawn from a gamma distribution. The substitution model used was the HKY model (HASEGAWA *et al*. 1985), as being the best fit of the available models for the introns. Bayesian runs were replicated to ensure result convergence, which was assessed based on several diagnostics suggested in the MrBayes manual. The average standard deviation of split frequencies was under 0.01 at the end of a run, the plot of generation versus log-likelihood scores for the replicated runs together resulted in a "white-noise" plot, and all PSRF (potential scale reduction factor) scores are equal to or close to 1.000 (RONQUIST *et al*. 2005).

5.4 Results

5.4.1 Adélie penguin intron analyses

Haplotype diversity, nucleotide diversity and mean number of pairwise differences were similar for four of the intron markers and generally high (e.g. haplotype diversities ranged from 0.6738 (*ODC6*) to 0.8065 (*UCHL3*) (Table 5.3 & Figs. 5.2 – 5.3). Adenylate kinase intron 5 diversity indices were higher than the other four (e.g. haplotype diversity was 0.9882), showing a larger number of rare and intermediate frequency alleles. *AK1i5* had a high proportion of singleton haplotypes (n=1 occurrence) (66%) compared to the other four introns (12-35%). For most of the statistics calculated, *AK1i5* presented different results from the other four introns. Due to the high diversity of this marker, phase calling of haplotypes was ambiguous for seven of the samples. In order to ensure that results based on this locus were not driven by erroneously called haplotypes, a "b" dataset including only unambiguously phased haplotypes was used for further analyses. This reduced dataset presented similarly high diversity indices (Table 5.3). Markers *UCHL3*, *ODC6* and *AK1i5* all presented insertion-deletion polymorphisms.

LAMARC estimates of θ for the five loci converged well (ESS > 400 for all loci), and were between 0.01 and 0.05. Confidence intervals for all five loci overlapped (Table 5.4). These estimates were converted to effective population size estimates using the two mutation rates obtained (based on 40Mya time to most recent common extant penguin ancestor, μ^1, or 15Mya, μ^2) (Table 5.5). Averaging over the five loci, N_e (μ^1) = 2,390,000 (438,000 – 8,690,000) or N_e (μ^2) = 757,000 (139,000 – 2,760,000).

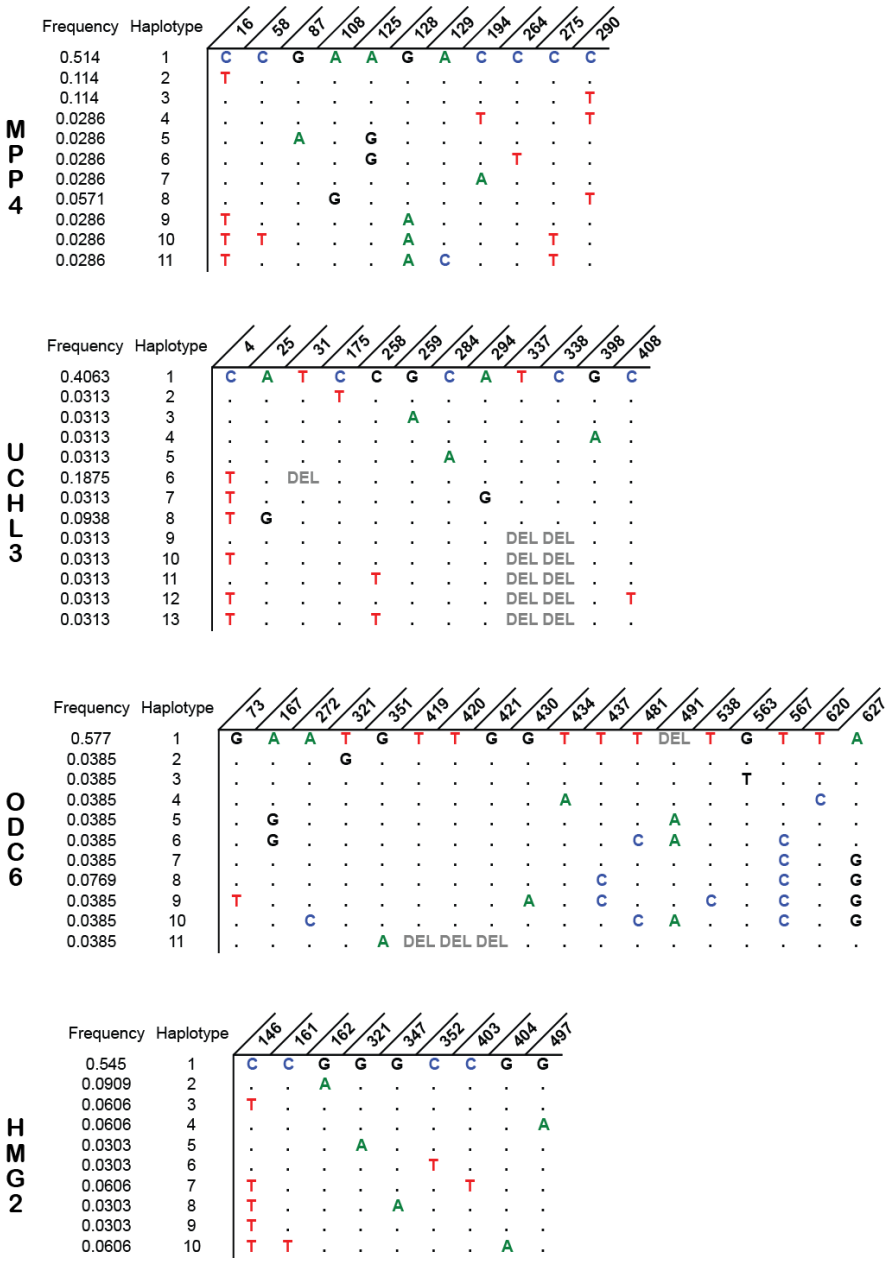


Figure 5.2 Haplotypes found for four introns in Adélie penguins.

Frequency of each haplotype is given, as well as the position of the variable sites defining each haplotype.

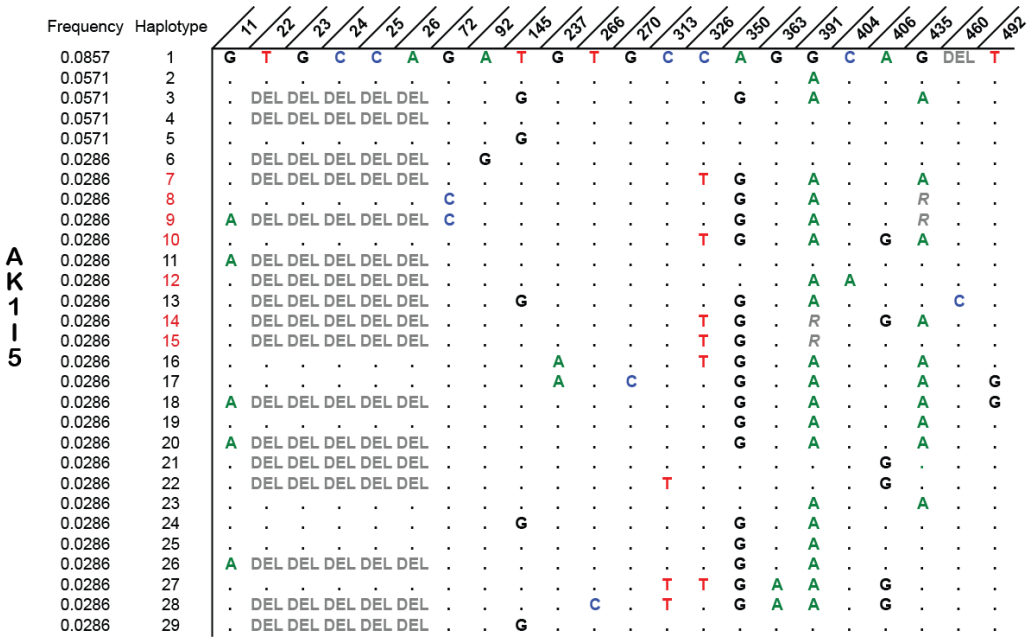


Figure 5.3 Haplotypes found for Ak1i5 in Adélie penguins.

Frequency of each haplotype is given, as well as the position of the variable sites defining each haplotype. *R* = ambiguous base, A/G. Five samples were phased with probabilities under 85%. The seven unique haplotypes derived from these ambiguously phased sequences are indicated in red.

Tajima's *D* values showed significant, negative departures from the assumed null hypothesis of neutral evolution for four of the introns, ranging from -0.775 to -1.388 ($p<0.01$). Adenylate kinase intron 5, however, showed a significant and large positive departure from zero, $D = 6.894$ ($p<0.05$). Tajima's *D* for the reduced, unambiguous dataset (*AK1i5b*), was still positive, but non-significant (1.3419). Fu's *Fs* values were significantly negative as well for all five introns, four of which ranged from -3.385 to -5.228 ($p<0.01$). Adenylate kinase intron 5 presented a value of *Fs* of -10.287 (or -12.281 for *AK1i5b*). Overall these significantly negative values could be consistent with either a population expansion or selective sweep. Fu and Li's F^* and D^*, however, were not significant. These statistics are more powerful than *Fs* and *D* for detecting background selection. The non-significance of F^* and D^* lends support to a potential population expansion.

Table 5.3 Summary statistics for five intron loci in Adélie penguins.

Locus	I	*bp (ex)*	Sub. Model	Ti:Tv	N_{sam}	N_{seq}	N_H	Indels	Pol. Sites	*h*	Nucleotide Diversity	Θ (π)	*D*	*Fs*	F^*	D^*
MPP4	4	296 (79)	T92	4.7	19	35	11	0	11	0.7210 +/- 0.0765	0.005097 +/- 0.003494	1.508697 +/- 1.034082	-0.881**	-3.974**	-1.473	-1.756
UCHL3	5	415 (57)	T92	9	18	32	13	2	12	0.8065 +/- 0.0600	0.002838 +/- 0.002083	1.177817 +/- 0.864312	-0.775**	-4.988**	-2.216	-2.325
ODC6	6-7	649 (158)	T92 + G (0.05)	1.8	18	26	11	2	18	0.6738 +/- 0.1044	0.002951 +/- 0.001935	1.906155 +/- 1.249786	-1.388*	-3.385**	-2.057	-2.265
AK1i5[1]	5	522 (48)	TN93 + I (0.95)	13.1	20	35	29	2	21	0.9882 +/- 0.1010	0.006537 +/- 0.003808	3.41238 +/- 1.987646	6.894*	-10.319**	-	-
AK1i5[2]	5	522 (48)	T92 + G (0.05) + I (0.49)	3.73	15	25	22	2	20	0.9867+/- 0.0167	0.006461 +/- 0.003966	3.529307 +/- 1.857979	1.3419	-12.281**	0.337	0.195
HMG2	4	518	T92	13.1	21	33	10	0	10	0.6970 +/- 0.0857	0.002435 +/- 0.001752	1.206145 +/- 0.878400	-1.395**	-5.228**	-0.426	-0.850

I = intron. Length (*bp*), (*ex*) = length of coding sequence included. Sub. Model = substitution model. Ti:Tv = transition/transversion ratio under the chosen substitution model. N_{sam} = number of Adélie penguin samples sequenced per locus. N_{seq} = number of sequences generated from the samples (includes phased heterozygotes). N_H = number of haplotypes. Indels = number of insertion-deletion polymorphisms (one indel can be ≥ 1bp in length. Pol. Sites = Polymorphic Sites. Genetic/haplotypic diversity = *h*. Θ (π) = theta based on pairwise differences. *D* = Tajima's *D*. D^{**} = p<0.01. D^* = p<0.05. *Fs* = Fu's *Fs*. Fs^{**} = p<0.01. F* and D* = Fu and Li's F* and D*. AK1i5[1] = summary statistics for Adenylate Kinase intron 5, including ambiguously phased haplotypes. AK1i5[2] = summary statistics for Adenylate Kinase intron 5, including only unambiguously phased haplotypes.

Mismatch distribution analyses for the five loci independently indicated that the sudden expansion model could not be rejected. Distributions were unimodal (Fig. 5.4) and exhibited low, non-significant raggedness indices (Table 5.4), though it is worth nothing that, for marker *UCHL3*, raggedness was higher than for the other loci by one order of magnitude and SSD approached significance (p=0.0836). *Tau* (τ) values ranged from 0.553 to 8. These values were converted to time since expansion by dividing τ by the two mutation rates generated from phylogenetic analyses of penguins (based on a 40Mya (μ^1) and a 15Mya (μ^2) hypothesized time to the most recent common ancestor of extant penguins). These mutation rates were scaled by locus length (see Table 5.4) and generation time (6.46 years). Expansion times ranged from 1.67 Mya to 13.7 Mya, based on μ^1, or 530 kya – 4.35 Mya (based on μ^2). The precision of this estimate was quite low, however, as confidence intervals for τ were large (Table 5.4).

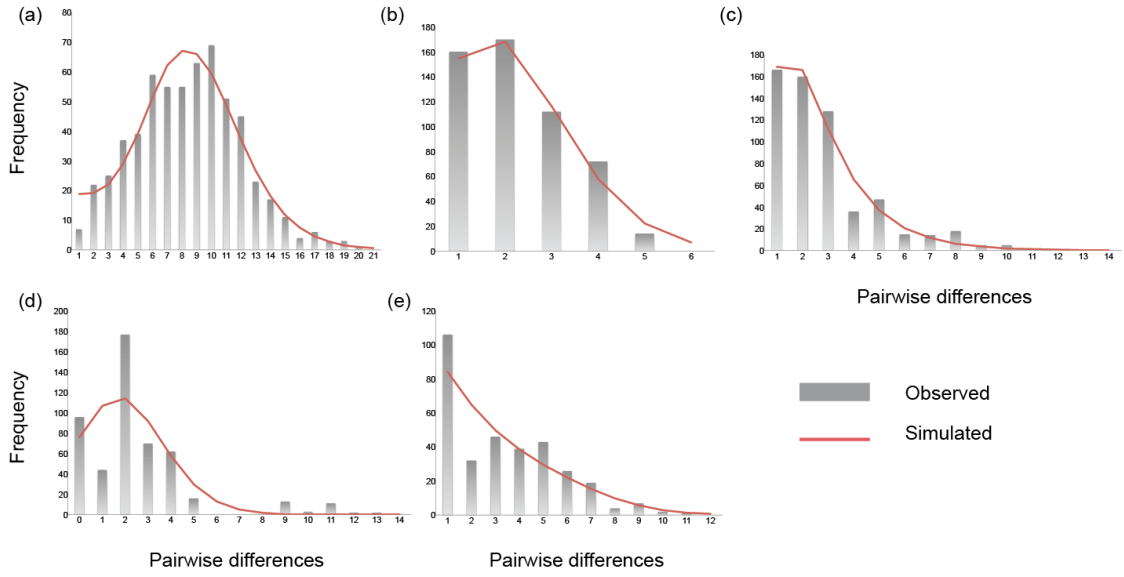


Figure 5.4 Mismatch distributions for five intron loci in Adélie penguins.
(a) *AK1i5b*, (b) *HMG2*, (c) *MPP4*, (d) *UCHL3*, (e) *ODC6*.

LAMARC estimates of the growth rate (*g*) were large and positive for four loci, indicating population expansion (*g* = 364 to 12221). Convergence for this parameter, however, was difficult to achieve and confidence intervals for the estimates of *g* were

large and in two cases included negative values (*UCHL3* and *ODC6*). LAMARC values of *g* are biased upwards and if these values include zero, there could be little or no growth. If the intervals exclude zero, the finding is generally reliable (LAMARC manual). Markers *HMG2* and *MPP4* gave *g* estimates supporting population growth, while *g* confidence intervals of *UCHL3* and *ODC6* included zero, implying either population growth is not supported or the genealogy sampling did not converge. Marker *AK1i5* gave no indication of population growth (*g* = 6.5 (-44,5-154 95% CI)).

Nuclear intron haplotype networks did not provide support for genetic structure based on the mtDNA *HVRI* A and RS lineages (Figs. 5.5 – 5.6). No evidence for any other structuring was found. Networks for most introns are star-like; however, the adenylate kinase intron 5 network is complex, showing a high degree of reticulation (Fig. 5.6). This reticulation was somewhat reduced when only unambiguously phased sequences were included in the analysis, however, the network still shows greater complexity when compared to the other intron markers. Pairwise Φ_{st} estimates between A and RS individuals for each intron yielded no significant values (data not shown), except for locus *ODC6* (Φ_{st}=0.018, $p<0.05$).

Phi tests in SPLITSTREE showed no evidence of recombination for any of the markers. RDP3 recombination tests were also negative for all markers. LAMARC runs, however, showed evidence of recombination in several of the markers (*AK1i5*, *UCHL3* and *ODC6*, *r* = 6.7163, 6.8204 and 2.8084, respectively). While these parameters did not converge well for all the markers (ESS below 100 for *UCHL3* and *AK1i5*), they were relatively higher than estimates for the presumably non-recombining *MPP4* and *HMG2* (*r* = 0.028 and 0.089, respectively).

Table 5.4 Results of the mismatch analysis for five intron loci in Adélie penguins and θ estimates.

	Mismatch Analysis						
Locus	**τ (95% CI)**	**SSD P**	***r***	**μ (s/l/g)[1]**	**μ (s/l/g)[2]**	***t*[1]**	***t*[2]**
MPP4	0.553 (0 - 4.572)	0.7458	0.0307	9.1×10^{-7}	2.9×10^{-6}	1.67×10^{6}	5.30×10^{5}
UCHL3	2.629 (0.594 - 4.756)	0.0836	0.1410	1.3×10^{-6}	4.0×10^{-6}	5.67×10^{6}	1.80×10^{6}
ODC6	5.223 (0.281 - 91.223)	0.6995	0.0600	2.0×10^{-6}	6.3×10^{-6}	7.20×10^{6}	2.28×10^{6}
HMG2	1.439 (0 - 3.584)	0.9019	0.0309	1.6×10^{-6}	5.0×10^{-6}	2.48×10^{6}	7.88×10^{6}
AK1i5b	8 (3.904 - 12.676)	0.9027	0.0053	1.6×10^{-6}	5.1×10^{-6}	1.37×10^{7}	4.35×10^{6}

For marker AK1i5, estimates were obtained for the reduced, unambiguously phased dataset (b). Raggedness indices (r) were non-significant in all cases. μ (s/l/g)[1,2] = substitutions per locus per generation, based on a generation time of 6.46 years and a divergence time of 40Mya[1] and 15Mya[2]. $t^{1,2}$ = years to expansion event, calculated with a mutation rate based on a divergence time of extant penguins of 40Mya[1] and 15Mya[2]. Θ estimates = median.

Table 5.5 Effective population size estimates for five intron loci in Adélie penguins

	Lamarc Analysis		
Locus	**θ (95% CI)**	N_e^1	N_e^2
MPP4	0.066 (0.0064 - 0.29)	5.35×10^{6} (5.19×10^{5} – 2.36×10^{7})	1.70×10^{6} (1.65×10^{5} – 7.49×10^{6})
UCHL3	0.023 (0.0047 - 0.065)	1.88×10^{6} (3.82×10^{5} – 5.28×10^{6})	5.96×10^{5} (1.21×10^{5} – 1.67×10^{6})
ODC6	0.027 (0.0041 - 0.11)	2.18×10^{6} (3.31×10^{5} – 8.85×10^{7})	6.93×10^{5} (1.05×10^{5} – 2.81×10^{6})
HMG2	0.013 (0.0022 - 0.038)	1.06×10^{6} (1.79×10^{5} – 3.12×10^{6})	3.35×10^{5} (5.69×10^{4} – 9.89×10^{5})
AK1i5b	0.018 (0.0096 – 0.032)	1.46×10^{6} (7.79×10^{5} – 2.61×10^{6})	4.63×10^{5} (2.47×10^{5} – 8.27×10^{5})

Θ estimates = median. $N_e^{1,2} = \theta/4\mu$, where μ is the mutation rate per site per generation (generation time = 6.46 years). $\mu^1 = 3.07 \times 10^{-9}$ s/s/gen (based on 40Mya divergence time between extant penguins). $\mu^2 = 9.69 \times 10^{-9}$ s/s/gen (based on 15Mya divergence time between extant penguins).

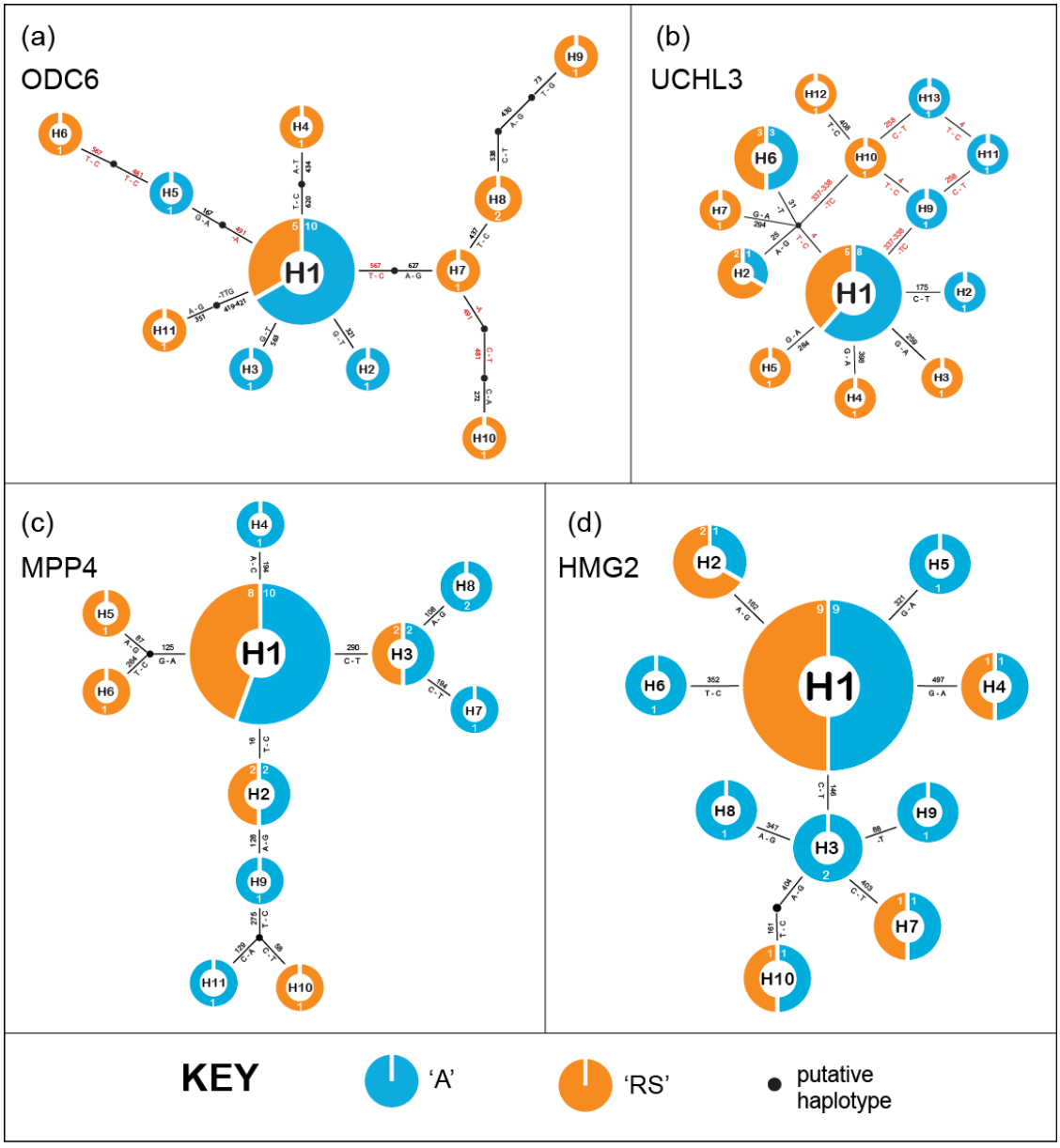


Figure 5.5 Haplotype networks constructed using sequence data from four introns of Adélie penguins.

Each line represents one mutation step – multiple base pair indels are treated as one step. Samples are colour-coded depending on their belonging to the A or RS mitochondrial lineages. Haplotype pie charts are proportional to the number of sequences, which are indicated on each pie. Substitutions shown in red are those represented more than once in the network.

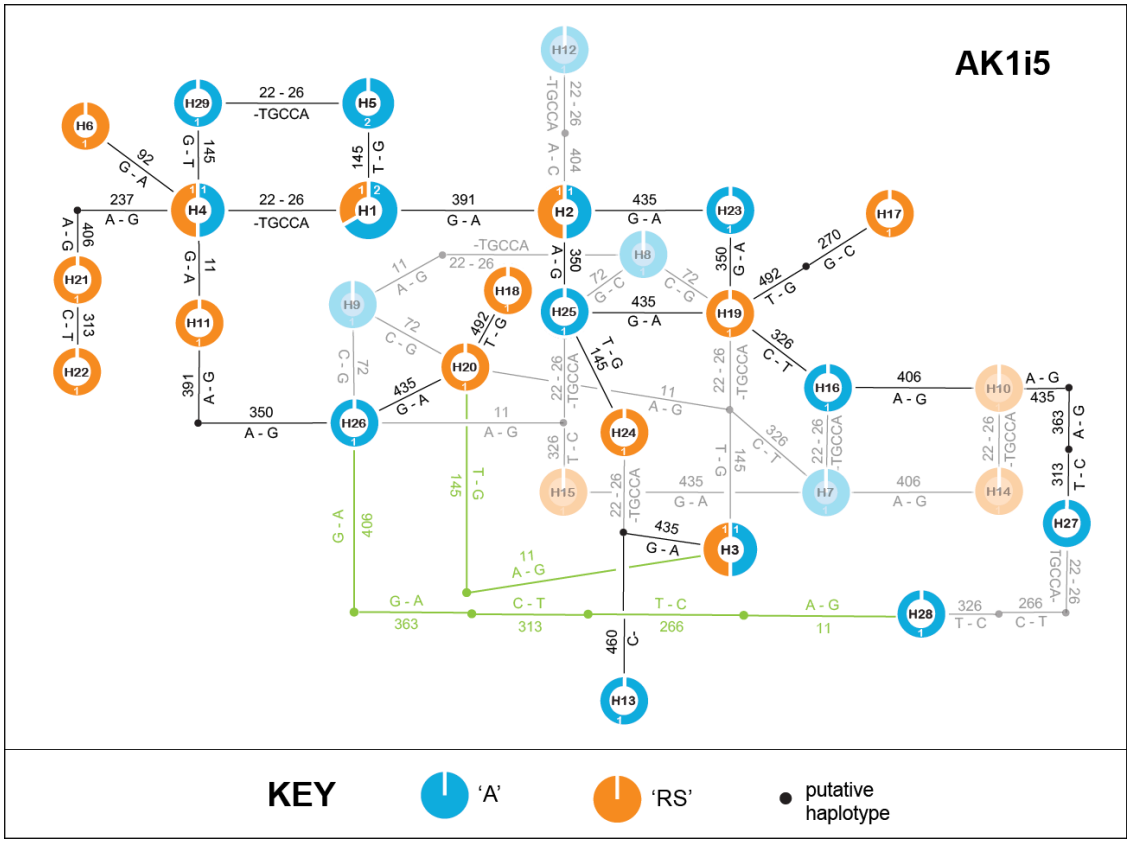


Figure 5.6 Haplotype network constructed using sequence data from *AK1i5* of Adélie penguins.

Each line represents one mutation step – multiple base pair indels are treated as one step. Samples are colour coded depending on their belonging to the A or RS mitochondrial lineages. Haplotype pie charts are proportional to the number of sequences, which are indicated on each pie. Two haplotype networks were combined. Connections and haplotypes that have been faded were present using the full dataset (ambiguously phased haplotypes included). Green connections are those present only in the partial dataset analysis (unambiguously phased haplotypes only). Black connections and dark haplotypes are those present in both analyses.

5.4.2 Penguin phylogenetic analyses

All four intron loci presented heterozygotic sequences for different penguin species, of which *UCHL3*, *ODC6* and *AK1i5* presented length variant heterozygotes. No *MPP4* sequences showed evidence of insertion-deletion polymorphisms. The four loci contained different amounts of phylogenetic signal. The amount of phylogenetic signal could be related to intron length and/or differences in the rate of accumulation among loci.

A root-to-tip distance for Adélie penguins of 0.017123 was estimated from a UPGMA tree (see Appendix Two). Divided by a divergence time of 40Mya (BAKER *et al*. 2006), a substitution rate of 4.28 x 10^{-10} s/s/yr was estimated. The substitution rate obtained using BEAST was very similar, 4.76 x 10^{-10} s/s/yr. If a more recent divergence time of 15Mya (CLARKE *et al*. 2007) was used instead, then the rate is an order of magnitude faster (1.14 x 10^{-9} s/s/yr from MEGA and 1.5 x 10^{-9} s/s/yr from BEAST). The rates obtained from BEAST were used to estimate population parameters.

No individual intron analysis recovered a robust phylogenetic tree (see Appendix Two, Fig. II.5); the combined dataset however provided robust support for all genera splits except *Eudyptula* (Fig. 5.7). No strongly supported nodes (over 90%) from individual analyses contradicted the concatenated phylogenetic tree. For the combined dataset, the position of *Eudyptula minor* as basal to the genus *Spheniscus* showed only 60% support (Bayesian posterior probability and maximum likelihood bootstrap scores). This split was strongly supported by the *AK1i5* trees (including the combined *AK1i5*/*ODC6* analysis, Fig. 5.8), however for *UCHL3*, *Eudyptula* was placed basal to *Pygoscelis* (63-78%), which lowered the support for the combined dataset. Species-level splits within *Pygoscelis* and *Spheniscus* were well supported. For the combined dataset, *Aptenodytes patagonicus* was not included as no reliable sequence was obtained from marker *UCHL3* for this species; however, this split was well supported by *AK1i5* and *ODC6* (both individually and combined, Fig. 5.8, 100-99.6%). The splits among *Eudyptes* species were generally not well supported. The split between *Eudyptes chrysocome* and *Eudyptes pachrhynchus* was the most strongly supported (90% Bayesian posterior probability, but only 46% maximum likelihood bootstrap support). The position of *Eudyptes sclateri* and *Eudyptes schlegeli* was basal to the other two *Eudyptes* species (no Bayesian support, 84% maximum likelihood bootstrap score), however which of the species is most basal is not resolved.

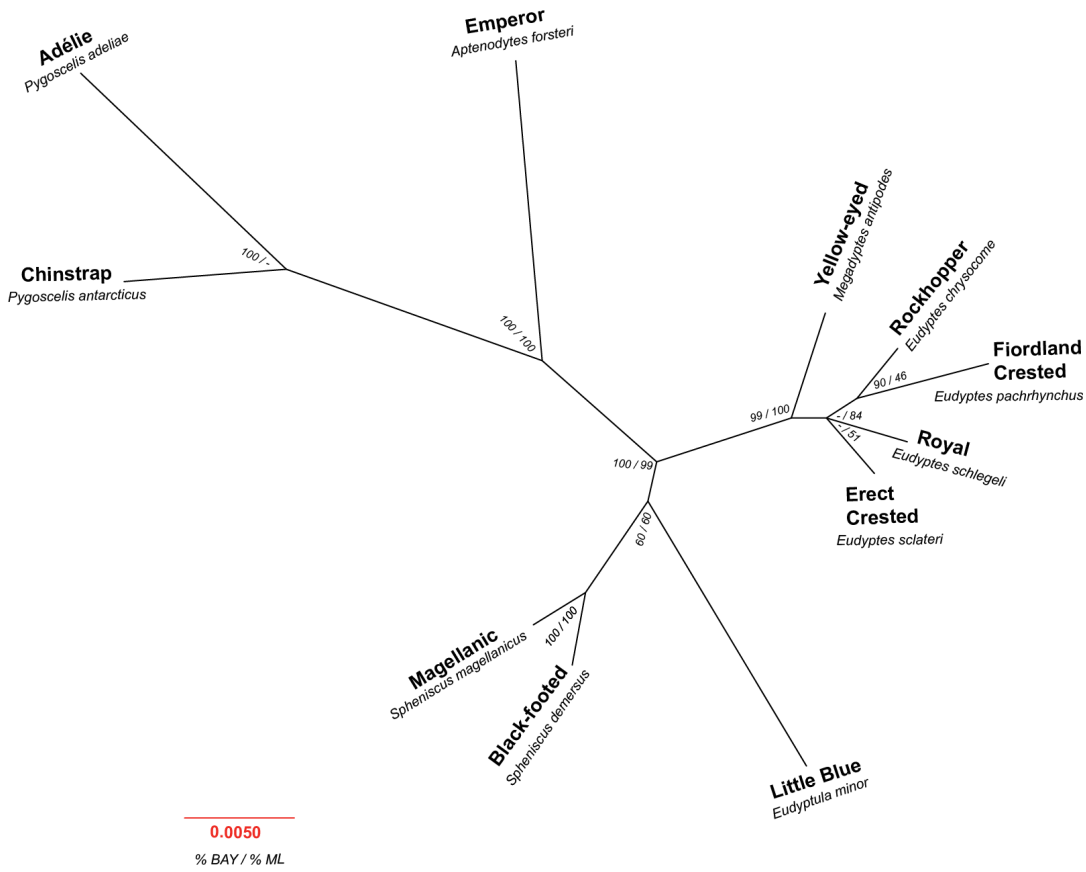


Figure 5.7 Unrooted Bayesian modern penguin phylogenetic consensus tree for the concatenated four intron dataset (1926 bp incl. gaps).

Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. Only species for which all four introns (*UCHL3*, *AK1i5*, *MPP4*, *ODC6*) were obtained were included in the analysis. The scale indicates a branch length equivalent to 0.005 substitutions per site.

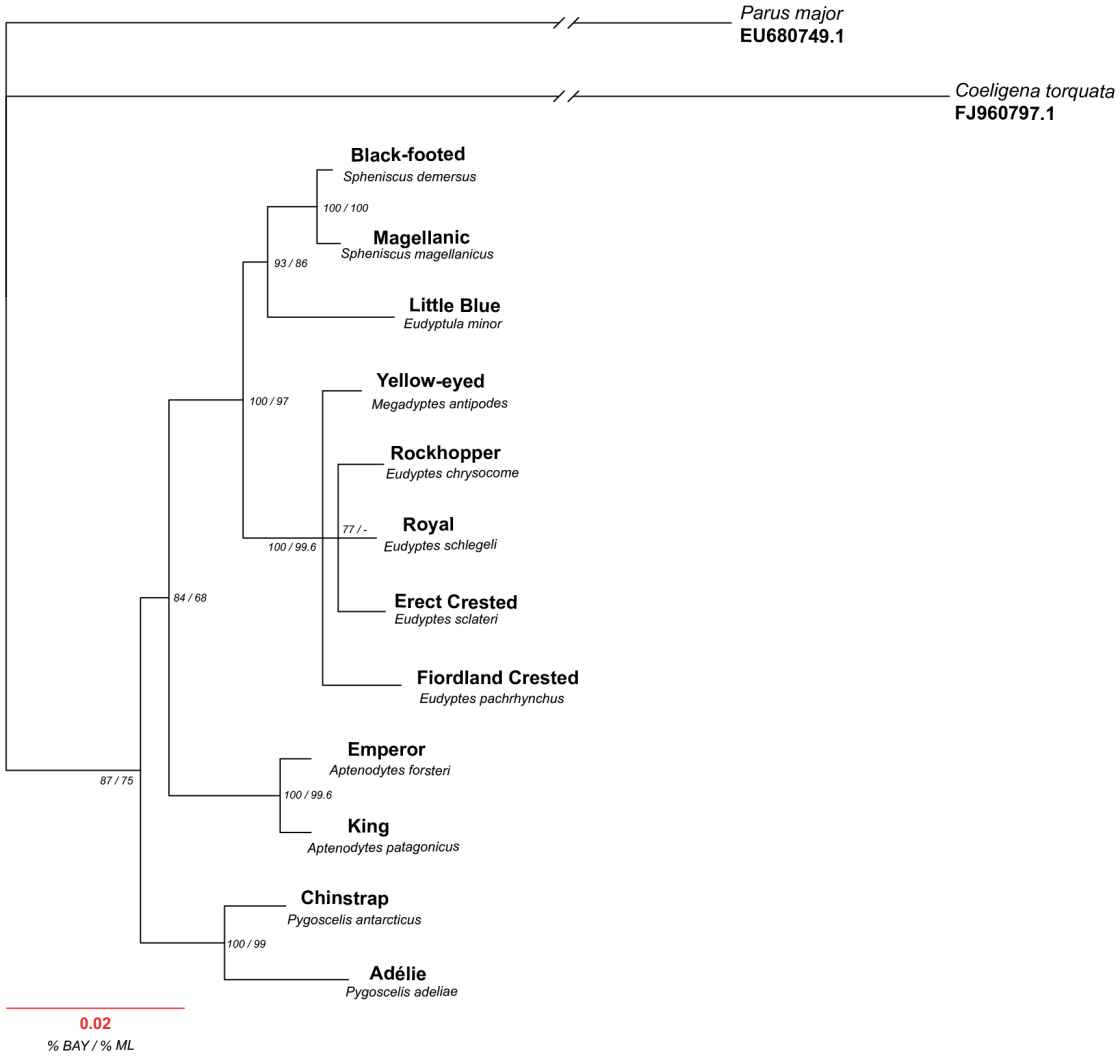


Figure 5.8 Rooted Bayesian modern penguin phylogenetic consensus tree for the concatenated *AK1i5*/*ODC6* dataset (1216 bp incl. gaps).

Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. Outgroup NCBI accession numbers are shown. The scale indicates a branch length equivalent to 0.02 substitutions per site.

No suitable outgroup was available for the four-intron dataset (Fig. 5.7). The combined dataset of *AK1i5* and *ODC6* with two passerine outgroup species showed a topologically similar phylogeny; however, the relationship between *Megadyptes* and *Eudyptes* genera was not resolved. When compared to the phylogenetic reconstruction based on *RAG-1* and four mitochondrial genes (BAKER *et al*. 2006),

the position of *Aptenodytes* and *Pygoscelis* were reversed (*Pygoscelis* appears most basal in the present study). Otherwise, there were no topological discrepancies between the present four-intron phylogeny and the published nuclear and mitochondrial coding phylogeny.

5.5 Discussion

In this study, the usefulness of nuclear intron markers for the elucidation of Adélie penguin population inference and for the recovery of penguin phylogenetic relationships was assessed. Previous work using mainly mitochondrial DNA markers has provided a good understanding of population-level relationships within Adélie penguins (LAMBERT *et al*. 2002; MILLAR *et al*. 2008; RITCHIE and LAMBERT 2000; RITCHIE *et al*. 2004; SHEPHERD *et al*. 2005) and a well-resolved phylogeny for extant penguin species (BAKER *et al*. 2006). Overall, the nuclear intron markers used provided an independent view of a history of population expansion in Adélie penguins, strengthening mitochondrial marker-based hypotheses. Despite the difficulties inherent to using allelic, recombining, more slowly evolving markers, when compared to more traditionally used mitochondrial markers, in population studies, it is important to obtain independent assessment of population processes. More loci rather than more mitochondrial sites helps increase our confidence in signals detected.

Also, four introns combined to provide sufficient phylogenetic signal to resolve relationships among penguin genera and most species. In the future, increasing the number of intron markers should extend this result to also provide tip resolution. These and other intron markers can easily be applied to more detailed interspecific or intraspecific penguin studies, as the primers used cross-amplify across all penguin species tested using one standard touchdown PCR program.

5.5.1 Adélie penguin intron population genetics

One of the most striking results found from the analyses of nuclear introns across ten Adélie penguin populations was the different patterns of variation observed for the marker *AK1i5* compared to the other four sequenced. Due to the methodology used for phasing the haplotypes of the heterozygotic introns obtained, which relied on Bayesian probabilities implemented in PHASE, we cannot exclude the possibility that some haplotypes were incorrectly called. *AK1i5* overall presented the most polymorphic sites (1 per 24.9 bp of sequence, compared to 1/26.9 (*MPP4*), 1/34.6 (*UCHL3*), 1/36.1 (*ODC6*), and 1/51.8 (*HMG2*)). PHASE utilizes the information from confirmed haplotypes to aid in assigning probabilities to putative haplotypes. For *AK1i5*, confirmed haplotypes showed high diversity. As a result, a number of potential phased haplotypes were almost equally probable (60%-40%). In order to address this uncertainty, cloning of PCR products from unphased heterozygotes and sequencing of ten clones or more per sample would be required, and/or increased sampling from further Adélie penguin individuals. At present this is beyond the scope of the current project, however, a smaller dataset for *AK1i5* containing only unambiguously phased and homozygotic sequences showed similarly high diversities and a complicated network. The high reticulation and diversity at this marker may indicate a higher rate of evolution and potentially recombination. None of the analyses carried out could conclusively detect recombination. LAMARC estimates, despite not converging enough to provide a reliable estimate for all markers, did indicate varying amounts of recombination per marker. Estimates of recombination were highest for markers *UCHL3*, *AK1i5* and *ODC6*, and lowest for *HMG2* and *MPP4*. This is consistent with the statistical parsimony haplotype networks generated for these markers. Both *ODC6* and *UCHL3* presented several probable connections involving repeated substitutions (marked in red; Fig. 5.5). *AK1i5* presented a much higher degree of reticulation still (Fig. 5.6). Recombination at these loci could provide one explanation for these patterns; however, a larger sample size for each intron would be necessary to fully address this possibility.

In the present study I sought to investigate whether variation in nuclear introns followed a similar pattern as that found in mitochondrial DNA, providing support for

the hypothetically refugial mitochondrial lineages, A and RS. No support was found; haplotype networks constructed and coded according to mitochondrial lineage showed a random distribution of haplotypes with no clustering based on lineage. Coalescence time for nuclear markers is twice that of the mitochondrial genome however, and previous research hypothesized that the time to the most recent common ancestor of the A and RS lineages was 75 kyr BP (37-122 kyr), during the last glacial maximum. Also, nuclear introns, presenting mutation rates much lower than mitochondrial DNA in general and the control region in particular, accumulate changes more slowly and therefore are more prone to incomplete lineage sorting. The influence of recombination also aids to remove any evidence of genetic structuring when barriers to gene flow have disappeared. The estimate of the t_{mrca} of the mitochondrial A and RS lineages was based on a mitochondrial HVRI substitution rate calculated using radiocarbon-dated ancient Adélie subfossil bones (LAMBERT *et al*. 2002). Ritchie *et al* (2004) justify their use of this faster rate (0.96 s/s/Myr) for their estimates of lineage divergence times, rather than the more frequently used phylogenetic rate (0.208 s/s/Myr) due to low likelihood that the A and RS lineages would have remained isolated through multiple glacial cycles. The lack of any A/RS split among the five introns sequenced for this study lends support to their use of this faster rate. If A and RS lineages had remained separated from 120 kya until the Last Glacial Maximum leading to the Holocene, nuclear regions would have presumably accumulated enough substitutions to affect haplotypic representations.

No other structuring was observed in these networks. Networks were not coded according to sample provenance, as a fine-scale population analysis was not the objective of the present study. To further investigate whether any nuclear intron structure or differentiation exists between Adélie penguin colonies, sampling sizes of introns for each colony need to be increased by at least ten individuals, and further loci should be added. However, considering evidence for sporadic gene flow between Adélie penguin colonies (DUGGER *et al*. 2010; SHEPHERD *et al*. 2005), it is unlikely that any genetic structuring at the nuclear level would be evident. This finding supports prior research using microsatellite markers that also failed to find any evidence of structure among Ross Sea colonies (ROEDER *et al*. 2001).

Analysis carried out in this study indicated a probable historical demographic expansion. Haplotype networks were mostly star-shaped, which is thought to be characteristic of expanding populations (SLATKIN and HUDSON 1991). In support of this hypothesis, negative and significant Tajima's *D* and Fu's *Fs* were found for four of the introns, while Fu and Li's F^* and D^* were non-significant. The mismatch distributions failed to reject the sudden expansion model, and growth estimates from LAMARC were large and different from zero for two of the markers (for two others, confidence intervals included zero, while marker AK1i5 did not support growth). Interestingly, the three markers showing evidence of recombination were those that least supported population expansion. Further sampling, identification of recombinants, and re-analysis with recombinant-free alignments could aid in determining whether the effect of recombination is masking the signature of demographic growth. While individually these tests may not be enough to distinguish between selection or demographic effects, and further LAMARC runs could be carried out to increase confidence in the estimates obtained, taken together they indicate that Adélie penguins have the signature of an expanding population at nuclear loci as well as at the mitochondrial genome. Due to the difference in coalescence times for nuclear and mitochondrial regions, it is likely that they are identifying separate expansion events, since each glacial Antarctic cycle probably resulted in repeated population bottlenecks due to the loss of ice-free breeding areas (RITCHIE *et al*. 2004). Depending on the mutation rate used, the estimated times since expansion varied drastically. Averaged over the five loci, these values ranged from 1.95 Mya to 6.14 Mya. This range is wider still if one takes the confidence intervals into account (6.23 kya – 38.3 Mya). As a result, this reported time to expansion cannot be accurately determined at present. However, even the most recent estimate of time to expansion is well outside the confidence intervals for the divergence between the A and RS mitochondrial lineages (upper limit, 122 kya), which would have predated Adélie penguin expansion after the last glacial maximum. Almost certainly the signature of expansion reported here for nuclear introns originates from earlier population contractions and expansions. This illustrates the importance of using molecular markers that reflect different time periods within population history of a species.

5.5.2 Penguin phylogeny

The phylogenetic analyses carried out in this chapter served two main purposes. Firstly, to ascertain whether the published phylogeny could be recovered with four intron markers, and to use the calibrated divergence times published previously by Baker *et al* (2006) and Clarke *et al* (2007) to estimate an evolutionary rate for the intron markers used. None of the nuclear intron markers used in this study could fully resolve relationships among closely related species when used individually. The phylogeny based solely on *ODC6* was probably the best of the four used, consistent with previous findings that this marker performs better than many other reported introns for species-level phylogenetics (ALLEN and OMLAND 2003). Concatenating the intron sequences, however, provided much better resolution when compared to Baker *et al*'s (2006) phylogeny. This concatenated phylogeny, while containing a small number of differences, presents a topology generally very consistent with Baker *et al*'s (2006) phylogeny, especially considering the alignment used to generate the trees consisted of 1926 sites (gaps included) compared to 5691 bp of sequenced used previously. The few discrepancies observed would most likely be resolved if the number of loci included were increased. Tip resolution for more recently diverged species was lacking, for example the most recently diverged *Eudyptes* species pairs diverged within the last 2Myr. The largest discrepancy occurs at the base of the phylogeny. In this present phylogeny, *Pygoscelis* appears basal to all other penguin genera, while in Baker *et al* (2006), *Aptenodytes* is basal to all other penguins. Divergence of *Aptenodytes* and *Pygoscelis* occurred 40-38 Mya, and considering the four intron loci together could not provide tip resolution for speciation events up to 8 Mya, again it is likely increasing the number of loci used would provide a better estimate of this split. Adding a slower-evolving nuclear gene such as *RAG-1* would aid in resolving basal relationships, while identifying nuclear markers that are more quickly evolving, such as ANMs, might help with tip resolution. These divergence times would be even more recent if one accepts Clarke *et al*'s (2007) estimates. Also, in the present analyses, sites containing gaps were excluded, which reduces the amount of phylogenetic information available in each alignment. Further analyses, with two to four more intron loci, substitution models allowing for gaps, and common, suitable outgroups could provide better resolution.

In this study, a substitution rate for four loci combined (*ODC6*, *UCHL3*, *AK1i5* and *MPP*) was generated. The reliability of this estimate, however, is determined by the confidence in the time to the most recent common ancestor of Spheniscidae, or the extant crown penguin group, used to calibrate the rate. Baker *et al* (2006) constructed a robust molecular phylogeny of extant penguin species, and, based on this, estimated a t_{mrca} of 40 million years. Clarke *et al* (2007), however, countered this estimate by using fossil evidence and could not find support for a t_{mrca} beyond 15 million years. These two estimates, when converted to substitution rates, translate to an order of magnitude difference between them. Using 40Mya as a t_{mrca} gives a rate for the four loci together of 4.76×10^{-10} s/s/yr, compared to 1.5×10^{-9} s/s/yr from the 15Mya t_{mrca}. The question of which of these divergence times is closer to the true time to the most recent common ancestor cannot be ascertained at present; however, previously reported substitution rates for *ODC6* within birds are of the order of 10^{-9} s/s/yr, for example 1.2×10^{-9} s/s/yr in Gadwall (*Anas* strepera) (PETERS *et al*. 2008). Also, assuming a slower rate derived from the 40Mya divergence time resulted in very high estimates of both average effective population size of Adélie penguins (N_e = 2,390,000) and time to expansion (6.14 Mya). Adélie penguins, despite having very large population sizes around the Antarctic today, are likely to have undergone range and species contractions over successive warming and cooling periods, which would have reduced effective population sizes. This large size is therefore less likely biologically likely than that obtained from the faster rate (N_e = 757,000). Time to expansion using this faster rate also may be more likely (1.95 Mya) than that estimated using the faster rate (6.14Mya). These observations indicate that the t_{mrca} of Spheniscidae is, in all likelihood, younger than 40Mya. Current work by S. Subramanian also shows that a divergence time of 10Mya is more likely, estimated by calibrating the mitochondrial tree of all extant penguins with an evolutionary rate obtained from ancient and modern Adélie mitochondrial genomes (SUBRAMANIAN *et al*. 2009) (S. Subramanian, *personal communication*). This high level of uncertainty is often typical of substitution rates obtained through phylogenetic calibration methods. Alternate methods, for example using ancient DNA techniques, could offer a better understanding of evolutionary rates of non-coding regions in penguins.

5.5.3 Conclusions

Intron markers have great potential to provide complementary data for population genetics and phylogenetic relationships within penguins, as demonstrated through this study. One of the five markers had not been used previously for either population genetics or phylogenetics (*UCHL3*). Within Adélie penguins, the five intron markers used showed no evidence for the presence of the two mitochondrial lineages (A and RS), most likely due to lower mutation rates and larger effective population sizes for the intron markers relative to mitochondrial DNA, as well as gene flow between these lineages. The markers also indicated a population expansion event 1.95 to 6.14 million years ago. These introns are probably not applicable to finer-scale population genetics analyses than those carried out here; wider sampling within Adélie penguins is needed to assess their utility in this context fully. These markers will very likely prove useful within other penguin species, as this study also showed the cross-amplification of the primers and amplification conditions across all extant penguin genera.

Four intron markers individually did not provide enough resolution for phylogenetic analysis among penguins, due to the lack of topological resolution and low statistical support. The concatenated phylogeny, however, successfully recovered the majority of splits as reported in a robust phylogeny based on mitochondrial DNA and one nuclear exon (BAKER *et al*. 2006). While introns are still not as widely used in phylogenetics as mitochondrial DNA markers or even certain nuclear coding genes (eg *RAG-1*), a number of recent studies have used introns to resolve difficult phylogenies (CREER *et al*. 2006; DALEBOUT *et al*. 2008; JACOBSEN *et al*. 2010; YU *et al*. 2011) successfully. While individual introns frequently do not provide enough resolution individually, concatenation of different intron markers provides more power (JACOBSEN *et al*. 2010; YU *et al*. 2011). Working with introns is, however, not straightforward, due to difficulty in acquiring their sequences, phasing heterozygotes, gap-filled alignments, recombination, among others. In the past these difficulties were perhaps reason enough to avoid using introns, however new sequencing and bioinformatics developments are making using introns more feasible.

5.6 References

AKAIKE, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19:** 716-722.

ALLEN, E. S., and K. E. OMLAND, 2003 Novel Intron Phylogeny Supports Plumage Convergence in Orioles (*Icterus*). The Auk **4:** 961-969.

BACKSTRÖM, N., N. KARAISKOU, E. H. LEDER, L. GUSTAFSSON, C. R. PRIMMER *et al.*, 2008 A Gene-Based Genetic Linkage Map of the Collared Flycatcher (*Ficedula albicollis*) Reveals Extensive Synteny and Gene-Order Conservation During 100 Million Years of Avian Evolution. Genetics **179:** 1479-1495.

BAKER, A. J., S. L. PEREIRA, O. P. HADDRATH and K.-A. EDGE, 2006 Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. Proceedings of the Royal Society Series B **273:** 11-17.

BALLARD, J. W. O., and M. C. WHITLOCK, 2004 The incomplete natural history of mitochondria. Molecular Ecology **13:** 729-744.

BAZIN, E., S. GLEMIN and N. GALTIER, 2006 Population size does not influence mitochondrial genetic diversity in animals. Science **312:** 570 - 571.

BENSCH, S., D. E. IRWIN, J. H. IRWIN, L. KVIST and S. ÅKESSON, 2006 Conflicting patterns of mitochondrial and nuclear DNA diversity in *Phylloscopus* warblers. Molecular Ecology **15:** 161-171.

BORGE, T., M. T. WEBSTER, G. ANDERSSON and G.-P. SAETRE, 2005 Contrasting Patterns of Polymorphism and Divergence on the Z Chromosome and Autosomes in Two *Ficedula* Flycatcher Species. Genetics **171:** 1861-1873.

CLARKE, J. A., D. T. KSEPKA, M. STUCCHI, M. URBINA, N. GIANNINI *et al.*, 2007 Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. PNAS **104:** 11545-11550.

CLEMENT, M., D. POSADA and K. CRANDALL, 2000 TCS: a computer program to estimate gene genealogies. Mol Ecol **9:** 1657 - 1659.

CREER, S., C. E. POOK, A. MALHOTRA and R. S. THORPE, 2006 Optimal Intron Analyses in the *Trimeresurus* Radiation of Asian Pitvipers. Syst. Biol. **55:** 57-62.

DALEBOUT, M. L., D. STEEL and C. S. BAKER, 2008 Phylogeny of the Beaked Whale Genus Mesoplodon (Ziphiidae: Cetacea) Revealed by Nuclear Introns: Implications for the Evolution of Male Tusks. Systematic Biology **57:** 857-875.

DMITRIEV, D. A., and R. A. RAKITOV, 2008 Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels. PLoS Computational Biology **4:** e1000113.

DRUMMOND, A., and A. RAMBAUT, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology **7:** 214.

DRUMMOND, A. J., B. ASHTON, S. BUXTON, M. CHEUNG, A. COOPER *et al.*, 2010 Geneious v5.1. Available from http://www.geneious.com.

DUGGER, K. M., D. G. AINLEY, P. O. B. LYVER, K. BARTON and G. BALLARD, 2010 Survival differences and the effect of environmental instability on breeding dispersal in an Adélie penguin meta-population. Proceedings of the National Academy of Sciences **107:** 12375-12380.

EDWARDS, S. V., and P. BEERLI, 2000 Perspective: Gene Divergence, Population Divergence, and the Variance in Coalescence Time in Phylogeographic Studies. Evolution **54:** 1839-1854.

EXCOFFIER, L., and H. E. L. LISCHER, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10:** 564-567.

FAY, J. C., and C. I. WU, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Molecular Biology and Evolution **16:** 1003-1005.

FLOT, J.-F., 2010 SeqPHASE: a web tool for interconverting PHASE input/output files and fasta sequence alignments. Molecular Ecology Resources **10:** 162-166.

FRIESEN, V. L., 2000 Introns, pp. 274-294 in *Molecular Methods in Ecology*, edited by A. J. BAKER. Blackwell Science Ltd, Oxford.

FRIESEN, V. L., B. C. CONGDON, M. G. KIDD and T. P. BIRT, 1999 Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. Molecular Ecology **8:** 2147-2149.

FRIESEN, V. L., B. C. CONGDON, H. E. WALSH and T. P. BIRT, 1997 Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. Molecular Ecology **6:** 1047-1058.

FU, Y.-X., 1997 Statistical tests of neutrality against population growth, hitchhiking and background selection. Genetics **147:** 915 - 925.

FU, Y. X., and W. H. LI, 1993 Statistical Tests of Neutrality of Mutations. Genetics **133:** 693-709.

HARPENDING, H. C., 1994 Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Human Biology **v66:** p591(510).

HARPENDING, H. C., S. T. SHERRY, A. R. ROGERS and M. STONEKING, 1993 The Genetic Structure of Ancient Human Populations. Current Anthropology **34:** 483-496.

HASEGAWA, M., H. KISHINO and T.-A. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution **22:** 160-174.

HUELSENBECK, J. P., F. RONQUIST, R. NIELSEN and J. P. BOLLBACK, 2001 Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science **294:** 2310-2314.

HUSON, D. H., and D. BRYANT, 2006 Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution **23:** 254-267.

JACOBSEN, F., N. R. FRIEDMAN and K. E. OMLAND, 2010 Congruence between Nuclear and Mitochondrial DNA: Combination of Multiple Nuclear Introns Resolves a Well-supported Phylogeny of New World Orioles (*Icterus*). Molecular Phylogenetics and Evolution **56:** 419-427.

KNOWLES, L. L., and W. P. MADDISON, 2002 Statistical phylogeography. Molecular Ecology **11:** 2623-2635.

KUHNER, M. K., 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics **22:** 768-770.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of Evolution in Ancient DNA from Adélie Penguins. Science **295:** 2270-2273.

LIBRADO, P., and J. ROZAS, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:** 1451-1452.

MARTIN, D. P., P. LEMEY, M. LOTT, V. MOULTON, D. POSADA *et al.*, 2010 RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26:** 2462-2463.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al.*, 2008 Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

NIELSEN, R., and J. WAKELEY, 2001 Distinguishing Migration from Isolation: A Markov Chain Monte Carlo Approach. Genetics **158:** 885-896.

PETERS, J. L., Y. N. ZHURAVLEV, I. FEFELOV, E. M. HUMPHRIES and K. E. OMLAND, 2008 Multilocus phylogeography of a holarctic duck: colonization of North America from Eurasia by Gadwall (*Anas strepera*). Evolution **62:** 1469-1483.

RITCHIE, P. A., and D. M. LAMBERT, 2000 A repeat complex in the mitochondrial control region of Adélie penguins from Antarctica. Genome **43:** 613-618.

RITCHIE, P. A., C. D. MILLAR, G. C. GIBB, C. BARONI and D. M. LAMBERT, 2004 Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. Molecular Biology and Evolution **21:** 240-248.

ROEDER, A. D., R. K. MARSHALL, A. J. MITCHELSON, T. VISAGATHILAGAR, P. A. RITCHIE *et al.*, 2001 Gene flow on the ice: genetic differentiation among Adélie penguin colonies around Antarctica. Molecular Ecology **10:** 1645-1656.

ROGERS, A., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution **9:** 552 - 569.

RONQUIST, F., and J. P. HUELSENBECK, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574.

RONQUIST, F., J. P. HUELSENBECK and P. VAN DER MARK, 2005 Mr Bayes 3.1 Manual. Available at http://mrbayes.scs.fsu.edu/manual.php.

ROSENBERG, N. A., and M. NORDBORG, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics **3:** 380-390.

SAMBROOK, J., E. F. FRITSCH and T. MANIATUS, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York.

SEUTIN, G., B. N. WHITE and P. T. BOAG, 1991 Preservation of avian blood and tissue samples for DNA analyses. Canadian Journal of Zoology **69:** 82-90.

SHEPHERD, L. D., C. D. MILLAR, G. BALLARD, D. G. AINLEY, P. R. WILSON *et al.*, 2005 Microevolution and mega-icebergs in the Antarctic. PNAS **102:** 16717-16722.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. Genetics **141:** 413-429.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations. Genetics **129:** 555-562.

STEPHENS, M., N. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68:** 978 - 989.

SUBRAMANIAN, S., D. R. DENVER, C. D. MILLAR, T. HEUPINK, A. ASCHRAFI *et al.*, 2009 High mitogenomic evolutionary rates and time dependency. Trends in Genetics **25:** 482-468.

TAJIMA, F., 1989a The Effect of Change in Population Size on DNA Polymorphism. Genetics **123:** 597-601.

TAJIMA, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585 - 595.

TAJIMA, F., 1993 Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis. Genetics **135:** 599-607.

TAMURA, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Molecular Biology and Evolution **9:** 678-687.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution **10:** 512-526.

TAMURA, K., D. PETERSON, N. PETERSON, G. STECHER, M. NEI *et al.*, 2011 MEGA5: Molecular Evolutionary Genetics Analysis using Likelihood, Distance, and Parsimony methods. . Molecular Biology and Evolution **28:** 2731-2739.

YU, L., P.-T. LUAN, W. JIN, O. A. RYDER, L. G. CHEMNICK *et al.*, 2011 Phylogenetic Utility of Nuclear Introns in Interfamilial Relationships of Caniformia (Order Carnivora). Systematic Biology **60:** 175-187.

6 Chapter Six

RECOVERING ANCIENT NUCLEAR INTRONS OF ADÉLIE PENGUINS USING SECOND-GENERATION DNA SEQUENCING

6.1 Abstract

While many studies utilize multiple genetic loci to address questions in molecular ecology and evolution, the field of ancient DNA still relies heavily on mitochondrial DNA markers, due to the low copy number and general difficulties associated with the recovery of nuclear DNA. With the advent of second-generation sequencing however, a number of methods have been developed that may be applicable to obtaining ancient nuclear sequence from a range of individuals in one sequencing run. One such method, direct multiplex PCR FLX sequencing, previously found to be effective for the recovery of multiple ancient mitochondrial genomes at one time, was applied in this study to obtain three nuclear introns from ancient Adélie sub-fossil bones up to 6,000 years old. Over 90% of all reads obtained belonged to contaminant sequences, most of which were bacterial in origin, belonging to microbial groups known to dominate Antarctic soils. Of the intron reads obtained, all belonged to adenylate kinase intron 5, but the results did not provide enough coverage to obtain the full intron for all the sequences. A subset of these sequences was used to compare modern and ancient AK1i5 sequences, and a slight shift in haplotype frequencies was detected. A molecular rate of evolution was also estimated for this locus, however, there was not enough signal in the data to provide a reliable estimate. On the whole, PCR-enrichment methods for obtaining ancient nuclear intron sequence may not be suitable, due to the biases introduced during primer design and a lower limit of 100bp for fragment sizes. Hybridization capture methods may offer a better approach for ancient nuclear population studies.

6.2 Introduction

Within the last twenty to thirty years, an increasing number of studies have taken advantage of historical or ancient samples from both extant or extinct species in a number of ways. Ancient DNA has provided us with an improved understanding of a number of extinct species, for example the wooly mammoth (BARNES *et al*. 2007; CAMPBELL *et al*. 2010; GREENWOOD *et al*. 1999; HAILE *et al*. 2009; KRAUSE *et al*. 2006; MILLER *et al*. 2008; ROGAEV *et al*. 2006; RÖMPLER *et al*. 2006), the New Zealand moa (BAKER *et al*. 2005; BUNCE *et al*. 2009; HUYNEN *et al*. 2010; HUYNEN *et al*. 2003), and Neanderthals, among many others (BRIGGS *et al*. 2009; GREEN *et al*. 2010; GREEN *et al*. 2006; GREEN *et al*. 2008; KRINGS *et al*. 1999; LALUEZA-FOX *et al*. 2011; LALUEZA-FOX *et al*. 2005; NOONAN *et al*. 2006; SERRE *et al*. 2004). Historical samples of species of conservation concern can make valuable contributions towards wildlife management, by establishing pre-decline levels of diversity, population sizes, and gene flow or providing a temporal context regarding connectivity between currently isolated populations, information that would be oftentimes impossible to obtain another way (LEONARD 2008).

Obtaining temporally distributed DNA sequences from one species has also allowed the estimation of evolutionary rates for mitochondrial DNA sequences without relying on fossil calibration points (HAY *et al*. 2008; LAMBERT *et al*. 2002; MILLAR *et al*. 2008a; SHAPIRO *et al*. 2004). Interestingly rates obtained in this fashion have been high compared to those obtained using more traditional methods (HO *et al*. 2005; HO *et al*. 2007). This could in part be due to evolutionary processes, though it has also been shown that the Bayesian MCMC analyses employed to generate these estimates can be biased due to low information content of aDNA (DEBRUYNE and POINAR 2009) or demographic model misspecification (NAVASCUÉS and EMERSON 2009). Hence, any attempt to generate a molecular rate from aDNA data needs to take these issues into consideration. Ancient DNA techniques are fraught with difficulties associated with low template quantities and high fragmentation of template DNA, as well as miscoding lesions. Cytosine residues can be deaminated and will then be identified as thymine by DNA polymerases, leading to an overestimation of C – T changes (GILBERT *et al*. 2007). This can have serious consequences for demographic

inferences (AXELSSON *et al*. 2008). Distinguishing between genuine mutations and the consequences of post-mortem damage is therefore of great importance for the reliable estimation of population history.

New DNA sequencing technologies (KNAPP and HOFREITER 2010; MILLAR *et al*. 2008b) are offering exciting new ways to overcome a number of the hurdles involved in ancient DNA research. Massively parallel amplification from small amounts of starting material has meant that ancient DNA, present in small amounts and highly fragmented, can be obtained in high enough coverage to identify and distinguish DNA damage from substitutions and sequencing error. Initially these technologies have been instrumental in obtaining mitochondrial and nuclear genomes of, for example, Neanderthals (GREEN *et al*. 2010; GREEN *et al*. 2006; GREEN *et al*. 2008; NOONAN *et al*. 2006), and mammoth (MILLER *et al*. 2008). With the development of tagging techniques in which pooled samples can be distinguished statistically post-sequencing (BINLADEN *et al*. 2007; MEYER *et al*. 2008b), population-level studies of ancient and modern specimens are increasingly more affordable as well. This method has been applied to obtaining full mitochondrial genomes successfully, in combination with different enrichment approaches, for example in killer whales (MORIN *et al*. 2010) and beavers (HORN *et al*. 2011). In combination with target enrichment by multiplex PCR (STILLER *et al*. 2009) or hybridization capture (BRIGGS *et al*. 2009; MARICIC *et al*. 2010), these technologies may be opening up multilocus population level studies of ancient specimens, incorporating nuclear DNA to a field still very much reliant on mitochondrial DNA.

In the present study, sub-fossil bones of Adélie penguins, radiocarbon dated and ranging in ages from 775 to 6075 ybp and used successfully for mitochondrial DNA studies previously, were used to obtain single-copy nuclear non-coding DNA. The main objective was to obtain sequence from three nuclear introns previously applied to modern Adélie penguins (*AK1i5*, *UCHL3* and *MPP4*; see Chapters 4 and 5) from sub-fossil bones of a range of ages, characterize variability for these introns compared to the present time, and, if possible, estimate a substitution rate for the regions amplified.

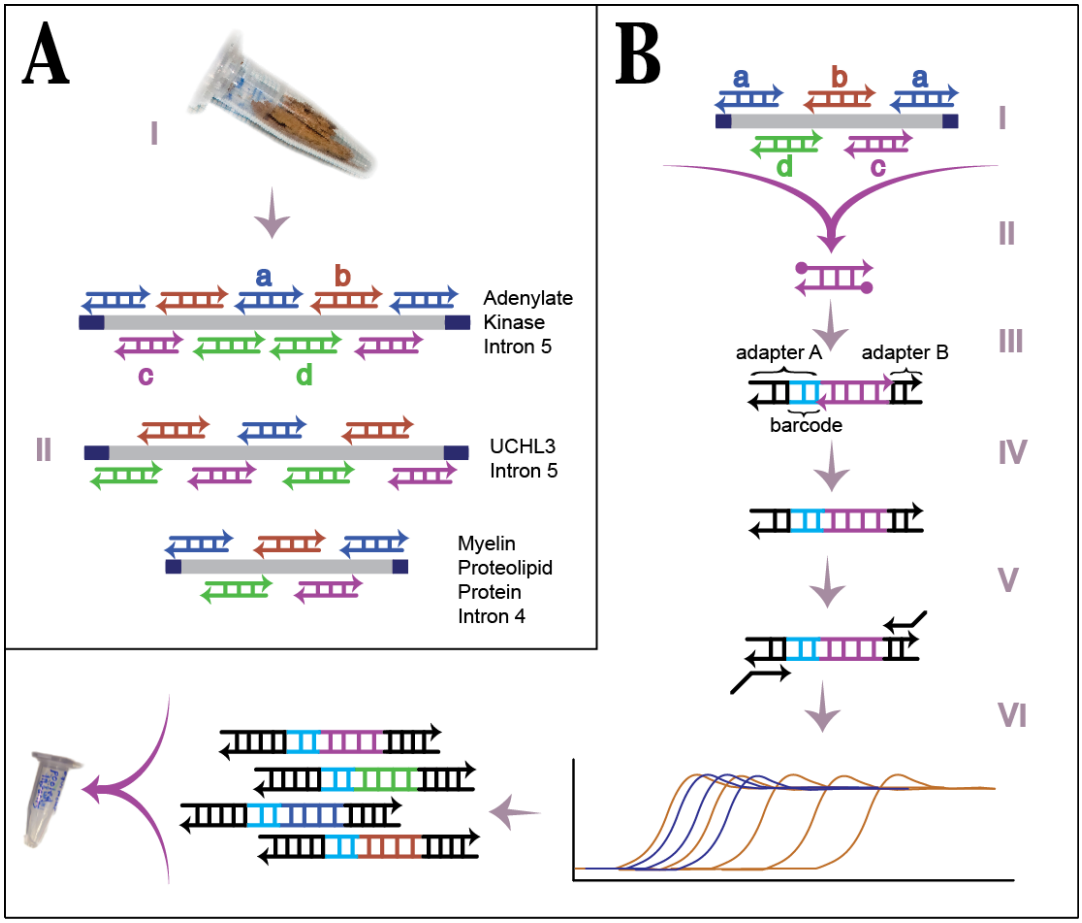


Figure 6.1: Overview of direct multiplex sequencing of ancient Adélie penguin nuclear intron products.

A. Each ancient sample was extracted following the Silica method (I) then subjected to multiplex PCR (II) in 4 reactions (a, b, c, d). Products were purified with SPRI beads and kept at -20°C until adapter ligation. Fragment number and position is shown for each of the three introns targeted in this study. **B.** After removal of primer dimers and excess reagents, multiplex products from each sample and replicate were pooled (I) and blunt-end repaired (II). A truncated barcoded FLX Titanium shotgun 'A' adapter and a truncated FLX Titanium shotgun 'B' adapter were ligated to the products (III), and then nicks were filled-in using a strand displacing polymerase (IV). The truncated adapters were extended to full length in an amplification step using 5' tailed primers (V). Double-stranded barcoded sequencing libraries were quantified by real-time PCR and pooled in equimolar ratios (VI) then subjected to FLX Titanium sequencing.

A method developed to obtain ancient mitochondrial genomes, combining multiplex PCR enrichment, tagging and sequencing on the FLX (Roche) sequencer (STILLER *et al*. 2009), was applied to obtain ancient nuclear Adélie penguin introns (Fig. 6.1). Multiplex PCR allows enrichment of multiple target fragments of DNA at one time. This is essential in ancient DNA studies as fragment sizes, especially for nuclear regions, tend to be on average between 50 and 150bp in length, and individual PCRs

of this fragment size would rapidly consume the limited amount of DNA typically available in an ancient extract (KRAUSE *et al*. 2006). Traditionally individual PCRs are then performed using the multiplex product. These singleplex products could then be either subjected to direct Sanger sequencing or pooled for second-generation sequencing. This has the advantage that pooling of different fragments can be done in equimolar ratios, thus obtaining nearly equal representations of sequencing reads across the amplicons. It has the disadvantage, however, of being very time consuming, particularly for longer regions obtained through a large number of very short amplicons. The multiplex product itself can also be used as a template for a second generation sequencing library, a more time efficient method (STILLER *et al*. 2009). Unequal efficiencies of different primer pairs can lead to unequal coverage of different amplicons; optimizing the number of cycles in PCR to be enough to lift the target above the contaminant background without creating a great disparity between poorly and well amplifying amplicons.

This method was characterized and implemented to obtain mitochondrial genomes and has not been used to obtain multiplexed single-copy ancient nuclear DNA. One added objective of this study was to gauge the applicability of this technique for the retrieval of single-copy nuclear DNA for a number of samples simultaneously. To this end, primers were designed for three nuclear introns to produce overlapping amplicons, multiplex PCRs were carried out in replicate along with negative controls, and the pooled products were subjected to tagging and sequencing on the FLX platform following the protocol described in Stiller *et al* (2009).

6.3 Methods

6.3.1 DNA Extractions from bone

Sub-fossil Adélie penguin bones, collected and C^{14} dated (LAMBERT *et al*. 2002; MILLAR *et al*. 2008a; RITCHIE *et al*. 2004; SHEPHERD *et al*. 2005) were chosen for the present study based on the apparent preservation of the sample, previous success

obtaining mitochondrial DNA from the sample, and adequate age distribution for further analyses. All ancient DNA extractions were performed in a facility dedicated to ancient DNA research, physically separated from other facilities handling DNA, and standard guidelines were followed to avoid contamination.

A total of 38 samples were chosen for extraction. 49 extractions were performed (for some of the samples, separate bones were extracted). Eight samples and 2 extraction negatives were processed at a time. Extractions were carried out following the Silica based method described in (ROHLAND and HOFREITER 2007a; ROHLAND and HOFREITER 2007b), with some small modifications. This method was shown to recover the most DNA without co-purifying common ancient DNA inhibitors (ROHLAND and HOFREITER 2007b), such as humic acids from samples covered in soil. Bone fragments or powder removed from the samples were, where possible, divided into two. Half of the fragments or powder was set aside and labeled so as to be available for replication work. The remainder was ground to fine powder and extracted. Bone pellets left after the first decalcification and extraction stage were kept at -20°C and re-extracted later if needed (17 were re-extracted). These second extractions from the bone pellets produced almost as much DNA as the first extractions. DNA was eluted from the silica pellet in 50μl of 0.001% Triton.

6.3.2 PCR verification of DNA extractions

Prior to PCR and second-generation sequencing, a verification step was carried out following DNA extraction to select extracts with amplifiable nuclear DNA of approximately 130bp. Previous work showed older samples sometimes failed to amplify nuclear products above 130bp (see Appendix Three).

A small internal fragment of the 5th intron of the Myelin Proteolipid Protein gene was chosen for this purpose. This was 124bp long plus primers. The primers MPPF and MPPRint1 (Chapter 4) were used at a concentration of 0.5μM along with dNTPs (0.2mM), MgCl2 (2mM), Buffer (1x), platinum Taq (1unit – all Invitrogen), BSA (0.6 mg/ml) and 1.5μl of undiluted DNA extract. PCRs were carried out in the modern DNA lab on an Applied Biosystems Veriti 96 well Thermocycler using the following

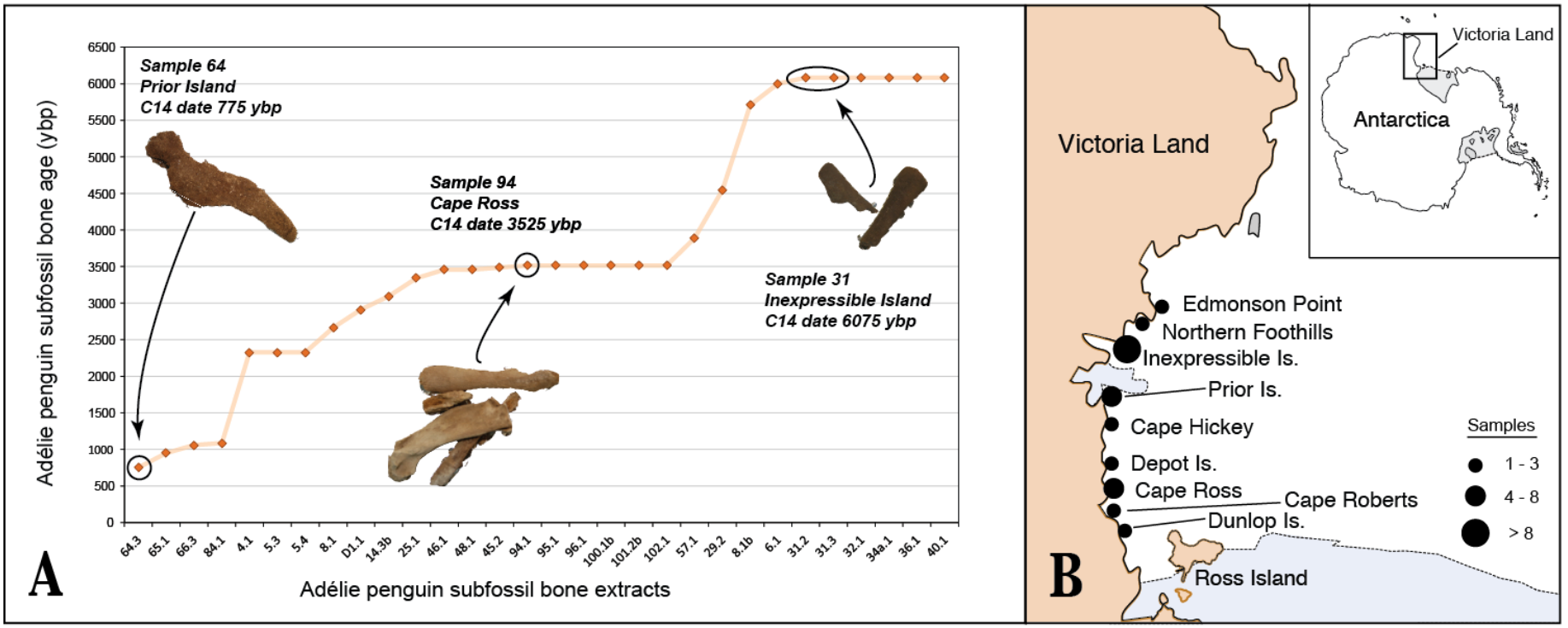


Figure 6.2: The age and geographical distribution of subfossil Adélie penguin bones used in this study.

A. Extraction number (X axis) and age of sub-fossil sample (Y axis). Three subfossil remains from which extracts were obtained are shown. **B**. Geographical distribution of abandoned penguin rookeries where sub-fossil bones were sampled. The number of sub-fossil bones used in this study are indicated approximately on the map (see key). The coordinates of each location and sub-fossil extracts (sfe) obtained from each location are the following: Edmonson Point (74°19'S, 165°04'E; sfe 84.1), Northern Foothills (74°41'S, 164°06'E; sfe 29.2), Inexpressible Island (74°54'S, 163°44'E; sfe 4.1, 5.3, 5.4, 25.1, 8.1b, 31.2, 31.3, 32.1, 34a.1, 36.1, 40.1), Prior Island (75°41'S, 162°52'E; sfe 64.3, 65.1, 66.3, 57.1), Cape Hickey (75°05'S, 162°38'E; sfe 45.2, 48.1), Depot Island (76°42'S, 162°57'E; sfe D1.1), Cape Ross (76°43'S, 162°59'E; sfe 94.1, 95.1, 96.1, 100.1b, 101.2b, 102.1), Cape Roberts (77°02'S, 163°10'E; sfe 14.3b), Dunlop Island (77°14'S, 163°28'E; sfe 8.1, 6.1).

program: 5' 94°C; 50 x [30" 94°C, 30" 54°C, 15" 72°C]; 10' 72°C. All further incubations and PCRs were carried out in the same thermocycler.

PCR products were visualized on a 2% agarose gel stained with ethidium bromide under UV light. Amplification products were checked for size against a 1kbplus ladder (Invitrogen) and were scored for presence or absence of the MPP intron fragment. Samples that amplified the product were selected for downstream work. Six extracts that were either duplicates (i.e. another working extract for that sample was available) or were the same age as another working extract were selected for multiplex PCR optimization steps. Thirty extracts were selected for the final multiplex amplification and FLX sequencing.

6.3.3 Designing internal primers for ancient DNA work

Initially, sequences obtained from the different introns were verified using NCBI Blast (*blastn*) and the closest hit was downloaded and included in sequence alignments to gauge variability within introns between closely related species as well as in flanking exons. If no intron sequence could be found, chicken, turkey or zebra finch flanking exonic sequences from Ensembl were added to the alignment. Most of the intron markers selected contained some flanking exon sequence, which, by alignment to known avian exonic sequences, permitted positive identification of the sequenced product.

Using the sequence reads obtained from modern samples, variable sites were identified within each intron marker (of those producing good sequence) and primers were designed flanking these sites within the intron to produce overlapping, smaller fragments that might amplify from ancient Adélie DNA extracts from sub-fossil bone. Primers were designed by eye from multiple sequence alignments visualized in Seaview v.4 (GOUY *et al*. 2010) and edited in Se-Al v2.0a11 (http://tree.bio.ed.ac.uk/software/seal/). The online program Primer3 (http://primer3.sourceforge.net/) was used to verify candidate primers by checking for unwanted self-hybridization that can outcompete binding to the DNA template, and

also to estimate melting temperatures. Where possible, primer length was constrained between 20 and 30bp, with melting temperatures above 40 and under 60°C and as close to each other as possible. Due to the small fragment sizes amplifiable from ancient DNA and the need for a minimum 30bp overlap there was often no way to design a perfect primer, and stringency was relaxed. Care was taken, where possible, to position priming sites outside of known variable positions. However, due to small numbers of sequenced samples it was not possible to rule out the presence of unknown variant positions. The program Amplify 3.1.4 (http://engels.genetics.wisc.edu/amplify/) was then used to simulate PCR, using a consensus sequence as a template, to detect potential nonspecific binding of primers within the target region, as smaller fragments are more abundant in ancient DNA and would amplify preferentially over our desired larger fragments.

The intron markers *Ak1i5*, *MPP4* and *UCHL3* (details in Chapter Four) were selected for targeting in ancient samples. Internal primers were designed to amplify small fragments, based on screening results from ancient samples (Appendix Three). Primers were designed to produce fragments over 100bp, as the magnetic beads used for downstream purification bind DNA 100bp and higher, and under 130bp.

6.3.4 Direct Multiplex PCR FLX Sequencing Methodology

Adapting a protocol specifically designed to work with first stage mitochondrial multiplex products from ancient DNA (STILLER *et al*. 2009) to nuclear intron fragments was used in order to minimize the amount of genetic material used while maximizing the sequence data obtained. A multiplex approach was necessary to avoid wasting ancient DNA sample unnecessarily, and a second-generation sequencing route was decided on to reduce the time required for direct sequencing and cloning. The GS FLX Titanium platform was chosen as the protocols for tagging of multiplexed samples are already in place, the sequence reads are the longest of any second-generation platform, and should in theory produce enough coverage to detect sequence variants and measure PCR, sequencing and ancient DNA damage from true substitutions.

Table 6.1: Multiplex primer groupings and information for the three different introns that were the subject of this study.

Group	Intron	Primer	Sequence (5'-> 3')	Position in Fig. 6.3	Product length (bp)	Multiplex amplification success
a	*AK1i5*	AKm1F	CTCGCGAGGTGAAGCAGGGAGA	Af1-r1	118	good
		AKm1R	TCCATTGGTGCCAGCAGCA			
c	*AK1i5*	AKm2F	ATTGCCATTGCCACCCACC	Af2-r2	120	good
		Akint5aR	TGGGACGGTGCCCGAGTTTAG			
b	*AK1i5*	AKm3F	CCAGCACAGATCCCATCCCAACA	Af3-r3	119	good
		AKm3R	GGGGGACGTCTCTCCAGGAT			
d	*AK1i5*	AKm4F	CCKGCCTTGGGGACAGCCT	Af4-r4	118	no
		AKm4R	AAGCCCTGGCTCTCTCACCYT			
a	*AK1i5*	AKm5F	GGATGTCCTCAGGCTGACAGA	Af5-r5	113	good
		AKm5R	TGGGATTCTGGGGGTCACCGG			
d	*AK1i5*	AKint5bF	ATGGTGTCCCTCTGTTCCTCTGC	Af6-r6	123	average
		AKint5cR	GGTCACAGATGTGCCCACAC			
b	*AK1i5*	AKint5cF	ACCGGGACACCGGAACAGTG	Af7-r7	115	average
		AKm7R	ACCACCTCTCCCCAAACCAAA			
c	*AK1i5*	AKm8F	GCAGAGGATRGTGTGGGCA	Af8-r8	121	good
		AKm8R	GGGCGATCTGTGGGCYGGG			
a	*AK1i5*	AKm9F	ACCTCTCCCCAAACCAAACCAGA	Af9-r9	130	average
		AKm9R	AAGGAGACGATGGTGAAACGC			
d	*UCHL3*	27356F[1]	GCTTGTGGGACAATTGGG	Cf1-r1	120	no
		16R1	TAACTGTGGGGAGGCACAGAG			
b	*UCHL3*	16F2	GAACTTTGGTAAGGTTGGGT	Cf2-r2	126	average
		16R6	ACCATTCATAAAGGATGATCTCT			
c	*UCHL3*	16F3	GGATATGTTTTGTACTACTTTCTCTG	Cf3-r3	129	no
		16R5	CCACAAACTAAAGCACTTGAAGT			
a	*UCHL3*	16F4	CTATTCAAAGAGCATTTCTACCTAT	Cf4-r4	128	no
		16R4	GCAATCAGCCTTCACTGTCT			
d	*UCHL3*	16F5	GTTGCAAACTGCAGCAGGTA	Cf5-r5	127	average
		16R3	CTGTAAGCTTTGAATGAAGTTGGAG			
b	*UCHL3*	16F6	TATCTTTAAGTCTTTCTTGGAATGA	Cf6-r6	127	average
		16R2	TCTTCACTGAAAAAGTTCCTAG			
c	*UCHL3*	16F7	ATGAAGTTGGAGTTAAAGCATGAC	Cf7-r7	123	no
		27356R[1]	CCTGAAGAGAGGGCCAAATA			
a	*MPPI4*	MPPF[2]	TACATCTACTTTAACACCTGGACCACCTG	Bf1-r1	124	no
		MPPintR1	AGTACCCGGTTCGCTGTCCAG			
d	*MPPI4*	22F2	TGTGCGGACGCCAGGATGTA	Bf2-r2	128	average
		22R4	ATAAAGCAGGGGCAAGTGCT			
b	*MPPI4*	22F3	TGGAGGAGAGGGTCCAGGTGGATA	Bf3-r3	123	good
		22R3	CCAGCCAGCTGCCTTGCTC			
c	*MPPI4*	22F4	GTGTGCATGACACATGTGGCT	Bf4-r4	122	good
		22R2	TGGAACGCTTTCCCTGGCAAG			
a	*MPPI4*	22F5	GAGGCTGAGCAAGGCAGCT	Bf5-r5	129	average
		MPPR[2]	TTGCAGATGGAGAGCAGGTTGGAGCC			

All primers were designed during this study except those indicated with a note ([1] Backström *et al*. 2008, [2] Friesen *et al*. 1999). Amplicon groupings refer to those detailed in Figure 6.2, and amplicon positions are indicated in Figure 6.3).

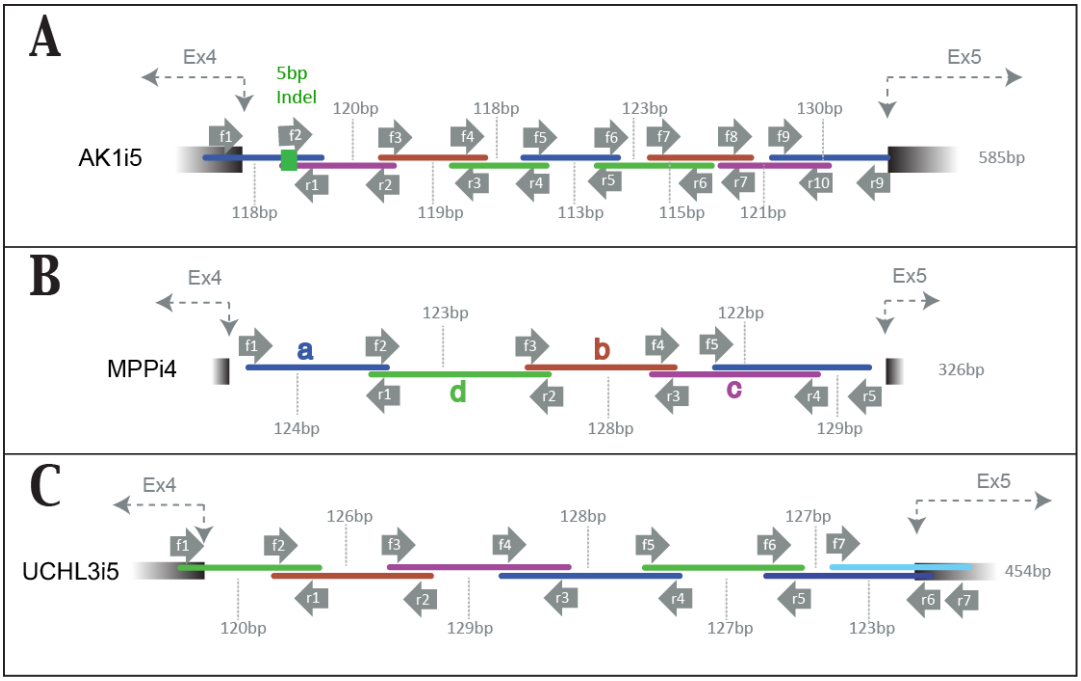


Figure 6.3: Multiplex primer positions and groupings.

Primer sequence details are found in Table 6.1. The four multiplex groups have been colour coded.

6.3.4.1 Multiplex PCR

Primer pairs covering three nuclear introns (adenylate kinase intron 5, ubiquitin carboxyl-terminal esterase L3 intron 5 and myelin proteolipid protein intron 4) with overlapping amplification products ranging in size from 118 and 130 bp in length were designed. These primers were initially split into two sets with no overlap within each set. PCR products were visually checked on a 3% agarose gel (2% agarose, 1% low-melt agarose) to verify the negative control and check for the presence of primer dimers. First stage multiplex runs were purified using Agencourt AMPure XP magnetic beads (see manufacturer's protocol) at a ratio of 1.8 beads per microlitre of product to eliminate all products under 100bp. The purified products were then diluted and used as starting template in singleplex PCRs to verify amplification of each original target from the first stage amplification. Verification was performed by gel electrophoresis on a 3% agarose gel (1% low melt agarose, 2% agarose).

Initial tests produced large amounts of primer dimer and biased amplification of certain products over others. In an effort to reduce these issues, primer concentrations

were reduced proportionally (concentrations of primer pairs that produced larger amplicon numbers were reduced relative to less-successful primer pairs), and primer groups were split into four instead of two. While amplification issues were reduced this way, six of the twenty-one primer pairs (16F-R4; MPPf-intR1; 16F3-R5; 16F7-27356R; Akm4F-R; 27356F-16R1) still failed to amplify during the second stage. These were tested individually on modern DNA samples and failed to amplify, suggesting problems with primer design probably due to the presence of unknown variable sites. They were nevertheless kept in the protocol, as perhaps even a small amount of amplified product amplified could be detected on the FLX Titanium platform.

Multiplex PCR was carried out in 20μl reactions containing 2.5mM $MgCl_2$, 0.25mM dNTPs, 100nM of appropriate primer mix, 1x Hi-Fidelity Buffer, 2 units of Platinum Taq HiFidelity (all Invitrogen), 5μl of template DNA, and 0.6mM BSA. DMSO and Betaine were initially included in PCR reactions but produced no beneficial effect on amplification so were not added to the final runs. The following thermal profile was used: 10' 94°C; 50 x [1' 95°C, 1'30" 57°C, 20" 68°C]; 10' 68°C.

All thirty samples and five extraction negative controls were amplified in eight reactions – four separate primer mixes each replicated independently once. The four products originating from the different primer mixes were pooled together at a later stage; however, the replicates were not pooled and were sequenced independently. Multiplex PCR products were purified as described above, eluted in 40μl TT Buffer (10mM Tris-HCl pH 8.0, 0.05% Tween20) and immediately placed in -20°C storage.

6.3.4.2 Tagged FLX Titanium Library Preparation

Sixty-five libraries were prepared following a slightly modified protocol described in (STILLER *et al*. 2009) (See Fig. 6.1 for an illustration of the protocol, Supplementary Methods, Appendix Three for the detailed protocol). Initial library preparation tests on modern DNA were carried out to verify the ligation protocol and efficiency of the

adapters. The final library preparations from the multiplexed products were carried out at the University of Otago, Department of Anatomy and Structural Biology.

In order to avoid costly titration steps on the FLX Titanium platform, and because library concentrations from ancient DNA are generally too low for traditional protocols, quantitative PCR was performed on the libraries following a slightly modified protocol detailed in (MEYER *et al*. 2008a) (Supplementary Methods, Appendix Three).

Once quantified, the barcoded libraries were pooled in equimolar ratios and concentrated using a Qiagen minelute column, eluting in 30μl 0.1xTE. This pooled library was first quantified on a NanoDrop 2000c spectrophotometer and subsequently on a Qubit Fluorometer (Invitrogen). The final concentration of the libraries was 27.8 ng/μl. Emulsion PCR was then set up with two copies of library per bead and submitted to the standard FLX Titanium protocol on 1/16th of a plate by the University of Otago High-Throughput DNA Sequencing Unit.

6.3.5 Analytical Methods

6.3.5.1 Raw FLX data filtering and aligning

Due to the nature of second-generation sequencing, an initial step in the analysis of the data produced is a non-negligible error-filtering and sequence alignment step. Large numbers of sequence reads are typically produced, and untagged reads, or reads with unassignable tags need to be eliminated first from the data. Next, 8bp tags and A and B adapters are identified on the sequence reads and trimmed off and a label is added to each sequence read name with a tag number for identification.

Sequences were loaded into Geneious Pro 5.0.4 (DRUMMOND *et al*. 2010) and separated according to tag number. Reference target sequences for the three introns, as well as references for each multiplex amplicon within the three introns, originating

from sequencing efforts in modern Adélie penguin populations, were loaded and used for assembly of sequence reads by tag and intron. Assembly parameters in Geneious were: Highest Sensitivity and Maximum Fine Tuning. Unmapped reads were saved as well. Assemblies for each intron-sample-replicate combination were assessed and primers trimmed by comparing to the reference amplicons (from which primers were previously trimmed). The assemblies were then visually inspected and edited in order to call a consensus sequences for each assembly, or two if heterozygote changes were found. Care was taken to verify potential changes through coverage consensus and by identifying miscoding lesions (G to A and C to T). The consensus assemblies were compared to their replicates. Changes were verified in replicate or discarded if not supported. One final super-consensus sequence for the intron was called, or two if two alleles were detected.

Separately, the taxonomic identity of each tagged read was queried by BLAST against the nr/nt database, and significant hits were recorded. All reads that gave a significant hit to an avian species were trimmed to remove primer sequence that might give rise to a false positive hit, and checked against the database again, and the results were updated to reflect this. BLAST results were separated into five sample age groups and one group consisting of the negative controls. The groups ranged in age from 750-1086 (1), 2328 (2), 2655-3485 (3), 3514 – 3888 (4) to 4543-6082 ybp (5).

6.3.5.2 Analysis of modern and ancient sequences

Modern and ancient introns were phased prior to analysis. Ancient intron sequences were phased based on FLX read data as described above. Modern heterozygote intron sequences used were phased by a combination of bioinformatics approaches. Length-variant heterozygotes were separated using CHAMPURU (FLOT 2007) and INDELLIGENT (DMITRIEV and RAKITOV 2008). If one heterozygote position was present, alleles were phased manually. Sequences with more than one heterozygote position present were subjected to analysis to statistically separate haplotypes based on probability, using the program PHASE (STEPHENS *et al*. 2001). Input files for PHASE were prepared using SeqPHASE (FLOT 2010). The most probable haplotypes

for each DNA sequence were selected for further analysis. Intron sequences were combined into different datasets for analysis. Separate modern and ancient intron datasets were generated, as well as a combined modern and ancient dataset. The number of modern samples included was similar to the number of ancient samples producing adequate coverage of the intron to avoid sample bias. Ancient datasets were further divided into one containing all consensus sequences obtained, and another including only those sequences providing >200bp coverage of the intron. Unless stated otherwise, all further analyses were performed using this dataset containing only sequences with >200bp coverage.

Best-fit models of sequence evolution were selected using the modeltest function implemented in MEGA 5.04 (TAMURA *et al*. 2011) for each intron alignment, based on the Akaike information criterion (AIC) (AKAIKE 1974), as well as transition – transversion ratios. A number of statistics were determined using the program ARLEQUIN 3.5.1.2 (default parameters, except for the substitution models and transition – transversion ratios found in MEGA 5.04) (EXCOFFIER and LISCHER 2010). PGDSpider 2.0.0.2 was used to generate the input files for each intron alignment. Genetic diversity measures were estimated (nucleotide diversity and mean number of pairwise differences among sequences) and their standard deviations. Different factors, such as population growth, selective sweeps, and background selection could affect patterns of DNA polymorphism. Two neutrality tests (Tajima's *D* statistic (TAJIMA 1989) and Fu's *Fs* (FU 1997)) were used to test whether the observed polymorphisms fit with neutral model expectations. The pairwise Φ_{st} for the intron between "modern" and "ancient" datasets was estimated and significance assessed. Haplotypes were determined as well, and these haplotype definitions were then used to create haplotype networks, to visualize phylogenetic relationships among intron haplotypes and alleles. Statistical parsimony haplotype networks were generated for the intron using the program TCS 1.2.1 (CLEMENT *et al*. 2000). For all loci, gaps were considered a 5th state. A chinstrap penguin sequence (*Pygoscelis antarcticus*) obtained for Chapter 5 was included as an outgroup. Modern and ancient haplotypes were partitioned into separate datasets to visualize the networks taking into account the temporal dimension of the data. Ancient sequences were split into three age groups; 500-2500 (1), 3000-4000 (2), 6000-6500 ybp (3).

6.3.5.3 Evolutionary rate estimates of intron sequences

Combining the full ancient intron dataset obtained and a modern intron dataset, an estimate of the evolutionary rate was obtained using BEAST (DRUMMOND and RAMBAUT 2007; DRUMMOND and RAMBAUT 2003). Genealogies were calibrated by adding tip age data. The default parameters were used, except for clock rate, in which the prior distribution was set to uniform (initial value 4.6×10^{-8} s/s/y). A strict clock model was applied, and a constant population size assumed. The HKY + gamma + four gamma rate categories model was used. The MCMC settings were 100,000,000 generations, sampled every 1000 generations. Convergence of the estimate was verified in TRACER 1.5 (http://beast.bio.ed.ac.uk/Tracer) to calculate the effective sample size (ESS). ESS values of 100-200 or greater indicate convergence had been achieved.

6.4 Results

6.4.1 FLX Output, Assembly and Coverage

A total of 38,700 reads were obtained on 1/16th of a plate using FLX Titanium chemistry. Of this total, 38,338 of these reads were tagged, confirming a very high (99%) success rate for our tagging protocol. Assembly results indicated a low presence of the target amplicons generated during multiplex PCRs (approximately 6% of reads). Of the three introns targeted, only adenylate kinase intron 5 reads were detected with any significance. Initial assemblies to myelin proteolipid protein intron four indicated one amplicon was sequenced; however, closer inspection showed this was due to mispriming during PCR and amplification of a contaminant sequence. A mistake was found in which the reverse complement of primer MPPR was ordered and used as both "MPPR" and "22R3". This primer acted as a forward primer, located in the 3' flanking 5th exon, amplifying contaminant sequences. Within adenylate

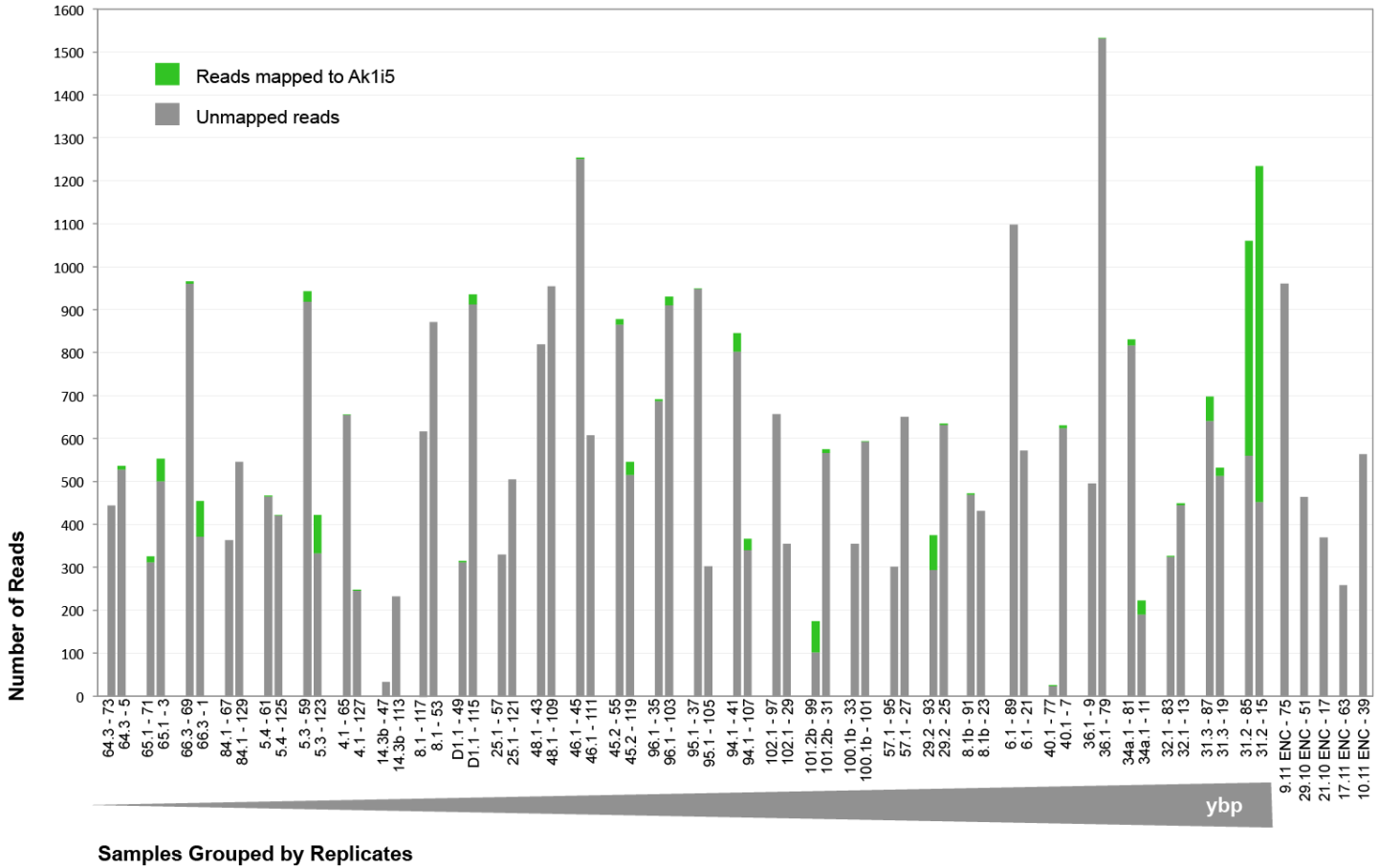


Figure 6.4: FLX read distribution across tags and replicates.

The total number of reads obtained per tag/replicate, split to show number of reads mapped to adenylate kinase intron 5. Samples are ordered on the X axis by increasing age (ybp).

Table 6.2 Coverage and distribution of FLX reads for adenylate kinase intron 5

Samples are ordered by increasing age and grouped by replicate. Coverage is color coded and detailed in the key below the table.

Sample Age (ybp)	Sample and Tag Number	Adenylate kinase intron 5 amplicons									Total Reads per Tag		% of Total	
		1	2	3	4	5	6	7	8	9	Not Assembled	Assembled	Not Assembled	Assembled
750	64.3 - 73	0	0	0	0	0	0	0	0	0	444	0	100.00	0.00
750	64.3 - 5	0	0	0	8	0	0	0	0	0	528	8	98.51	1.49
949	65.1 - 71	0	11	0	1	0	0	0	2	0	312	14	95.71	4.29
949	65.1 - 3	0	25	0	6	4		6	6	5	501	52	90.60	9.40
1057	66.3 - 69	0	2	1	2	0	0	0	0	0	961	5	99.48	0.52
1057	66.3 - 1	0	55	0	9	7	0	7	5	0	372	83	81.76	18.24
1086	84.1 - 67	0	0	0	0	0	0	0	0	0	363	0	100.00	0.00
1086	84.1 - 129	0	0	0	0	0	0	0	0	0	546	0	100.00	0.00
2328	5.4 - 61	0	2	0	0	0	0	0	0	0	465	2	99.57	0.43
2328	5.4 - 125	0	1	0	0	0	0	0	0	0	421	1	99.76	0.24
2328	5.3 - 59	0	11	0	1	3	0	6	3	0	919	24	97.45	2.55
2328	5.3 - 123	0	43	0	33	1	0	7	5	0	333	89	78.91	21.09
2328	4.1 - 65	0	0	0	1	0	0	0	0	0	655	1	99.85	0.15
2328	4.1 - 127	0	0	0	1	0	0	0	1	0	246	2	99.19	0.81
2655	14.3b - 47	0	0	0	0	0	0	0	0	0	33	0	100.00	0.00
2655	14.3b - 113	0	0	0	0	0	0	0	0	0	232	0	100.00	0.00
2911	D1.1 - 49	0	2	0	1	0	0	0	0	0	312	3	99.05	0.95
2911	D1.1 - 115	0	23	0	0	0	0	0	0	0	913	23	97.54	2.46
3340	25.1 - 57	0	0	0	0	0	0	0	0	0	330	0	100.00	0.00
3340	25.1 - 121	0	0	0	0	0	0	0	0	0	505	0	100.00	0.00
3456	48.1 - 43	0	0	0	0	0	0	0	0	0	819	0	100.00	0.00
3456	48.1 - 109	0	0	0	0	0	0	0	0	0	954	0	100.00	0.00
3456	46.1 - 45	0	0	0	0	0	1	2	0	0	1251	3	99.76	0.24
3456	46.1 - 111	0	0	0	0	0	0	0	0	0	608	0	100.00	0.00
3485	45.2 - 55	0	10	0	2	0	1	0	0	0	865	13	98.52	1.48
3485	45.2 - 119	0	18	0	9	0	0	3	1	0	515	31	94.32	5.68
3514	96.1 - 35	0	4	0	0	0	0	0	0	0	687	4	99.42	0.58
3514	96.1 - 103	0	17	0	2	0	0	0	1	0	910	20	97.85	2.15
3514	95.1 - 37	0	1	0	0	0	0	0	0	0	948	1	99.89	0.11
3514	95.1 - 105	0	0	0	0	0	0	0	0	0	303	0	100.00	0.00
3514	94.1 - 41	0	34	0	4	1	0	0	4	0	803	43	94.92	5.08
3514	94.1 - 107	0	16	0	6	4	0	0	0	0	340	26	92.90	7.10
3514	102.1 - 97	0	0	0	0	0	0	0	0	0	657	0	100.00	0.00
3514	102.1 - 29	0	0	0	0	0	0	0	0	0	355	0	100.00	0.00
3514	101.2b - 99	5	63	0	3	0	0	1	1	0	102	73	58.29	41.71
3514	101.2b - 31	0	7	0	0	1	0	0	0	0	567	8	98.61	1.39
3514	100.1b - 33	0	0	0	0	0	0	0	0	0	355	0	100.00	0.00
3514	100.1b - 101	0	0	0	0	1	0	0	0	0	593	1	99.83	0.17
3888	57.1 - 95	0	0	0	0	0	0	0	0	0	302	0	100.00	0.00
3888	57.1 - 27	0	0	0	0	0	0	0	0	0	650	0	100.00	0.00
4543	29.2 - 93	0	69	1	3	0	0	0	8	0	294	81	78.40	21.60
4543	29.2 - 25	0	3	0	0	0	0	0	0	0	632	3	99.53	0.47
5706	8.1b - 91	0	0	0	2	1	0	0	0	0	469	3	99.36	0.64
5706	8.1b - 23	0	0	0	0	0	0	0	0	0	431	0	100.00	0.00
5706	8.1 - 53	0	0	0	0	0	0	0	0	0	619	0	100.00	0.00
5706	8.1 - 117	0	0	0	0	0	0	0	0	0	873	0	100.00	0.00
5997	6.1 - 89	0	0	0	0	0	0	0	0	0	1098	0	100.00	0.00
5997	6.1 - 21	0	0	0	0	0	0	0	0	0	572	0	100.00	0.00
6082	40.1 - 77	0	2	0	0	0	0	0	0	0	24	2	92.31	7.69
6082	40.1 - 7	0	7	0	0	0	0	0	0	0	624	7	98.89	1.11
6082	36.1 - 9	0	0	0	0	0	0	0	0	0	495	0	100.00	0.00
6082	36.1 - 79	0	1	0	0	0	0	0	0	0	1532	1	99.93	0.07
6082	34a.1 - 81	0	9	0	5	0	0	0	0	0	817	14	98.32	1.68
6082	34a.1 - 11	0	32	0	0	0	0	0	1	0	190	33	85.20	14.80
6082	32.1 - 83	0	2	0	0	0	0	0	0	0	325	2	99.39	0.61
6082	32.1 - 13	0	4	0	0	0	0	0	0	0	445	4	99.11	0.89
6082	31.3 - 87	0	34	0	10	4	0	6	3	0	641	57	91.83	8.17
6082	31.3 - 19	0	12	0	2	0	1	0	4	0	513	19	96.43	3.57
6082	31.2 - 85	0	462	0	5	5	0	23	5	0	560	500	52.83	47.17
6082	31.2 - 15	0	769	0	1	0	0	9	3	0	452	782	36.63	63.37
NC	9.11 ENC - 75	0	0	0	0	0	0	0	0	0	961	0	100.00	0.00
NC	29.10 ENC - 51	0	0	0	0	0	0	0	0	0	464	0	100.00	0.00
NC	21.10 ENC - 17	0	0	0	0	0	0	0	0	0	370	0	100.00	0.00
NC	17.11 ENC - 63	0	0	0	0	0	0	0	0	0	259	0	100.00	0.00
NC	10.11 ENC - 39	0	0	0	0	0	0	0	0	0	564	0	100.00	0.00
TOTAL READS PER REGION		5	1751	2	117	32	3	70	53	5	**36199**	**2038**		
% of Total Reads	**38237**	0.01	4.58	0.01	0.31	0.08	0.01	0.18	0.14	0.01	**94.67**	**5.33**		

COVERAGE KEY	0	1 to 5	6 to 20	21 to 50	51 to 200	200+

kinase intron 5, amplicon coverage was skewed as well. Approximately 86% of *AK1i5* reads corresponded to the second amplicon fragment (Akm2F-Akint5aR) (Table 6.2). PCR singleplex results (Table 6.1) from the multiplex optimization were not a good indicator of coverage obtained from the FLX sequencing run. Certain fragments that amplified well in singleplex following first stage multiplex were not represented, particularly those for marker *MPP4*. For *AK1i5*, the first fragment amplified well consistently in singleplex PCR reactions yet was absent from FLX results. Also, the fourth fragment did not amplify well in singleplex, and yet was one of the most well represented sequences from FLX results. Primers designed for *UCHL3* did not amplify their target sequences well in singleplex. The number of sequences available at the time of primer design for this intron was lower than for *AK1i5* and *MPP4*, therefore it is possible that certain variable sites were not known and could have affected primer annealing.

6.4.2 Analysis of modern and ancient AK1i5 sequences

Of the thirty sub-fossil bones extracted and sequenced, eighteen contained enough *AK1i5* reads to generate consensus sequences of varying lengths. Ten of these produced total sequence length between 200-310bp; the remaining eight contained less than 200bp of sequence (Table 6.5). No one sample produced full coverage of the intron, in general three segments of the intron were sequenced (Fig. III.5, Appendix Three). The first segment, corresponding to AKm2F-Akint5aR, was the most successfully sequenced; however, this segment was for the most part invariable except for isolated changes. The second and third segments, including regions containing several closely clustered SNPs identified in modern Adélie penguins, were more variable, though less well covered. Four sites found to be variable in modern penguins were also variable in ancient penguins (sites 369, 393, 408, 437 when aligned to Chinstrap penguin, Fig. III.5, Appendix Three). These sites were found to be variable in previous work on modern introns (sites 367, 391, 406, 435, as labeled in Chapters 4 & 5). There were a number of isolated changes found only in one ancient sequence. One site, however, did show changes among most ancient sequences that were not present in the modern sequences, namely a single-nucleotide

change from T to G or A at site 243 (241 in Chapter 5). The youngest of the ancient sequences obtained, 65.1 (949 ybp) contained the modern T base. Another difference was observed at bases 169-170, an insertion deletion polymorphism not present in modern sequences. Certain ancient sequences also had either a C or a CG inserted when compared to modern sequences. This insertion was not present in the Chinstrap penguin sequence, and also did not appear to be related to age. As a string of three cytosine residues and a string of three guanine residues surrounded this insertion event, this change may be an artifact due to second-generation sequencing error. These two sites (the SNP and indel) were not taken into account for network construction and haplotype determination due to the fact that a number of sequences contained missing data at those sites (detailed further below).

Including all the ancient intron sequences obtained reduced the number of sites common to all sequences to 77. In modeltest results from MEGA 5 this implied obtaining a different evolutionary model than the one selected for modern sequences, or for ancient sequences including only those containing over 200bp of sequence (Table 6.3). The log likelihood for this model compared to those selected for the other datasets was also larger, indicating the fit was not as good. As a result, all further analyses were performed without the eight samples containing less than 200bp of sequence. The evolutionary model selected for modern and ancient sequences was the Tamura-3-parameter model, with invariant sites if ancient sequences were included (due to the high proportion of missing loci) (TAMURA 1992), the same as that obtained with a larger dataset (including unambiguously phased sequences) in Chapter 5.

Summary statistics between modern and ancient sequences were in general comparable (Table 6.4). Haplotypic and nucleotide diversity appeared higher in ancient samples. However, this is most likely due to the small number of sites included in the ancient analysis (198) that included most of the variable sites of the alignment. Compared to the results for *AK1i5* in Chapter 5, with a greater number of sequences, modern sequences here presented lower diversities, most likely due to a lower sample size. Tajima's D was non-significant, indicating a departure from the neutral theory could not be supported. Fu's Fs was negative and significant but small

Table 6.3: Modeltest results for modern and ancient AK1i5 Adélie datasets

Dataset	Model	lnL	Ti/Tv	A	T	C	G
Modern	T92	-784	2.67	0.174	0.174	0.326	0.326
Ancient all	JC + I (0.96)	-470	0.5	0.25	0.25	0.25	0.25
Ancient cut	T92	-596	3.34	0.184	0.184	0.316	0.316
M & A cut	T92+I (0.94)	-879	1.39	0.178	0.178	0.322	0.322

Evolutionary models were selected based on the AIC criterion, in MEGA. Datasets: Modern sequences, all ancient sequences obtained, all ancient sequences obtained >200bp, modern and ancient sequences >200bp. T92 = Tamura 3 parameter model. JC = Jukes Cantor model. I = invariant sites. lnL = log likelihood of each model. Ti/Tv = transition / transversion ratio. A/T/C/G = frequency of each nucleotide in the alignment.

Table 6.4: Summary statistics and neutrality tests for modern and ancient AK1i5 Adélie datasets

Data	N_{sam}	N_{seq}	N_h	N_{loci}	N_{loci}^*	Ti	Tv	Pol	*h*	Nucleotide Diversity	θ (π)	D	F_s	F_{st}
M	10	16	6	525	505	4	2	6	0.7417 +/- 0.1037	0.003438 +/- 0.002368	1.736229 +/- 1.1931	0.59814	**-0.52718**	n/a
A	9	17	7	525	198	4	0	4	0.8603 +/- 0.0503	0.007894 +/- 0.005539	1.563030 +/- 1.096764	0.93356	-1.28101	n/a
B			9											0.09

Summary statistics generated in Arlequin. Dataset used was modern sequences plus ancient sequences >200bp. M = modern. A = ancient. B = both. N_{sam} = number of samples. N_{seq} = number of sequences. N_h = number of haplotypes. N_{loci} = number of loci. N_{loci}^* = number of usable loci. Ti = transitions. Tv = transversions. *h* = haplotypic diversity. D = Tajima's D (nonsignificant). F_s = Fu's F_s, M = significant $p<0.05$, A = nonsignificant, $p=0.085$.

Table 6.5: Ancient and Modern Samples included in Analyses

Ancient	Age (ybp)	Length (bp)	Haplotype	Modern	Age (ybp)	Length (bp)	Haplotype	Ch5 Haplotype
65.1	949	296	b1, a8	T67	0	525	1	1
66.3	1057	304	a2, b3	T315	0	525	1	13
4.1	2328	160		T02	0	525	1	11
5.3	2328	310	b1, a3	T10*	0	525	a2, b1	5, new
5.4	2328	77		T18	0	525	a,b1	1, 5
D1.1	2911	121		T69*	0	525	4	new
45.2	3485	282	1	T08*	0	525	a2, b3	2, new
101.2b	3514	261	2	T162	0	525	a3, b4	2, 23
96.1	3514	171		T09*	0	525	a1, b5	1, new
94.1	3514	263	3	T13*	0	525	a1, b6	1, new
95.1	3514	77						
29.2	4543	264	a4, b9					
8.1b	5706	105						
34a.1	6082	226	b4, a7					
31.3	6082	285	a,b4					
31.2	6082	307	b2, a4					
32.1	6082	77						
40.1	6082	77						

The age of each sample is shown in years before present, as well as length of sequence obtained, and haplotypes generated from each sample. If the sample produced heterozygote sequence, haplotypes are shown for alleles a and b. For modern sequences, haplotypes following the definitions in Chapter 5 are listed, unless the sample contained new haplotypes (not defined) from samples used in Chapter 4 but not 5 (samples marked with an asterisk).

Haplotype	41	179	190	393	406	408	437
1	A	T	G	G	C	A	G
2	.	.	.	.	.	G	.
3	.	.	.	A	.	.	.
4	.	.	.	A	.	.	A
5	T	.	.	.	.	G	.
6	.	.	.	.	A	.	
7	.	.	A	A	.	.	A
8	.	C	.	.	.	.	.
9	.	.	.	.	.	G	A

Figure 6.5: Defining nucleotide changes for haplotypes of modern and ancient *AK1i5* sequences.

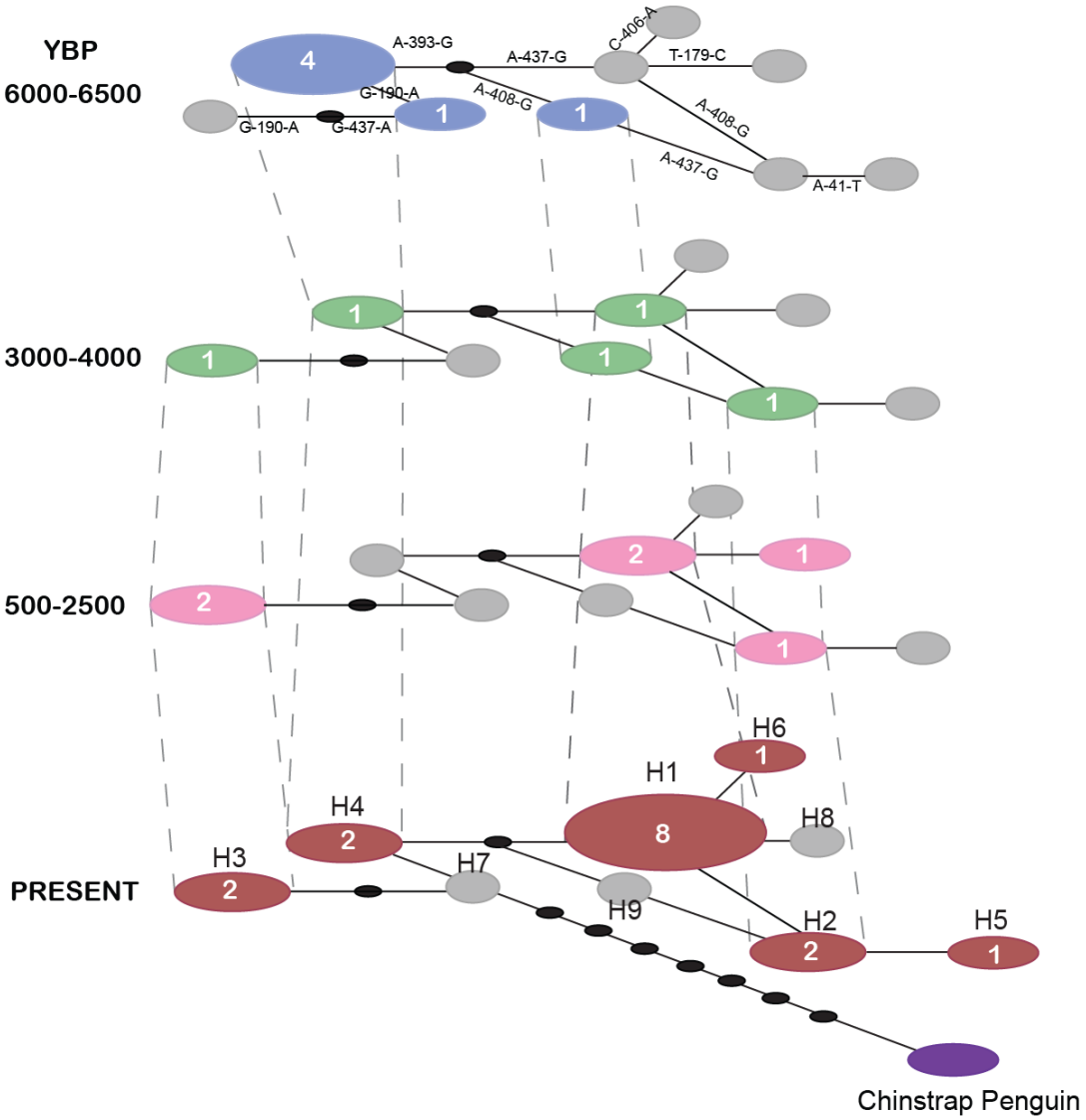


Figure 6.6: Temporal haplotype network for modern and ancient Adélie *AK1i5* sequences.

Haplotype numbers (H1-9) refer to Table 6.5. Number of sequences represented for each haplotype and layer are illustrated in white. "Missing" haplotypes not present in any samples included are black circles. Grey circles are haplotypes found in the samples included but not the age layer.

for modern sequences, and negative but non-significant for ancient sequences. Again, this difference to the results from Chapter 5 (F_s negative, large and significant) could also be due to sample size.

Haplotypes for the combined modern and ancient intron dataset were generated in Arlequin v. 3.5 (EXCOFFIER and LISCHER 2010), excluding ambiguous sites. This implies that the haplotypes shown here do not correspond to those found in Chapter 5 using the full intron sequence for *AK1i5*. In particular, the exclusion of the length-

variant polymorphic 5bp TGCCA repeat (not covered in ancient sequences obtained) affects haplotype determination. Table 6.5 lists the sequences used in this analysis and their haplotype. Haplotype definitions are shown in Figure 6.5. Most haplotypes were shared between ancient and modern sequences. Haplotypes 7, 8 and 9 were exclusive to ancient sequences, and 5 and 6 were exclusive to modern sequences. Figure 6.6 illustrates the connections between the nine haplotypes through time, as well as a probable position for the Chinstrap penguin root. Interestingly, the position of this root is closest to Haplotype 7, which is exclusive to the oldest time layer (Fig. 6.6). The four time layers show other differences as well. Haplotype 1, the most prevalent in modern sequences, becomes less abundant in the ancient age layers, and is entirely absent from the oldest layer. Haplotype 4 presents the opposite effect, being most prevalent in the oldest sample though still present in modern samples. Ashift in haplotype frequencies from those most closely related to H4 in ancient samples, to those more closely related to H1 is observed. However, sample sizes are small for each ancient layer. Further sampling is needed to characterize variation at each layer.

An evolutionary rate of 1.71×10^{-6} s/s/y (4.65×10^{-7} – 3.24×10^{-6} s/s/y) was estimated using BEAST (ESS = 816).

6.4.3 BLAST results of contaminant reads

BLAST results for the sequencing reads showed that for the total reads as well as for all age groups and negative extracts (Fig. 6.7), 73-89% of all reads were unassignable, that is, they did not produce any significant hit ($p<0.05$). Four percent of all reads produced a significant avian hit. The difference of 2% between the assembly results and BLAST results can be explained by the primer trimming carried out on sequences obtaining a significant avian hit in a first analysis to avoid false positive results; in some cases fragments were thus too small to produce a significant result. Age group results were updated to include positive avian hits identified from assemblies. A total of 13% of all reads produced a significant hit in bacteria, while mammals contributed 2%, plants 1% and the rest of reads were distributed among different groups (1%).

The highest percentage of bird reads was produced in the oldest age group, 5 (13%); the rest of age groups contained between 1% (3) and 7% (2). The high result from age group 5 is due to the high result from sample 31.2 (both replicates, tags 85 and 15) (Table 6.2). This sample produced the best total reads in assembly by far, despite being from the oldest age group. However, all reads were from one fragment so in terms of coverage, it was among the least successful samples.

More detailed results of the distribution of the significant reads are illustrated in Fig. 6.8. Percentages illustrated on the figure and in the text below refer to the percentage of the total significant reads. Human contamination varied from 1 to 6% of significant reads depending on the age group. Plant hits were made up almost completely by *Arabidopsis lyrata* originating from Cape Ross, samples 94.1, 96.1, 100.1b and 101.2b from age group 4 (Fig. 6.2). *A. lyrata* reads were present in only one of the two replicates for each of these samples. *Trichoplax adhaerens* (Metazoa: Placozoa) significant hits originated from Inexpressible Island (sample 25.1) and Depot Island (sample D1.1), but only in one of the two replicates for each of these, from age group 3. *M. mazeii* (Archaea) significant hits were from both replicates of sample 32.1, Inexpressible Island, age group 5.

With all results combined, the majority of reads were from Bacteria. Actinobacteria (34.8%) and Proteobacteria (17.6%) phyla made up nearly all of these. Among Actinobacteria, 5.1% of significant reads grouped within Actinosynnemataceae, of which *Actinosynnema mirum* made up all the hits. Within the same suborder (Pseudonocardineae), *Pseudonocardiaceae* were 2% (mostly *Amycolatopsis mediterranei*). The next most common family of Actinobacteria was Mycobacteriaceae (3.8%), including mainly *Mycobacterium smegmatis* (2.3%). Streptomycetaceae representatives amounted to 3.6%, mainly *Streptomyces bingchenggensis* and *S. griseus*. Three families dominated representatives from the suborder Micrococcineae: Dermabacteraceae, Micrococcaceae and Sanguibacteraceae. The family Dermabacteraceae, composed of *Brachybacterium faecium* and a few other congeneric species made up 3.1%. Micrococcaceae representatives amounted to 2.7% (approximately half were *Micrococcus luteus*). Sanguibacteraceae, composed solely of *Sanguibacter keddieii* made up 2.3%.

Geodermatophilaceae (suborder Frankineae; all *Geodermatophilus obscurus*) made up 2.4%. Several other families were represented under 2%; for example Catenulisporaceae (1.7%, *Catenulispora acidiphila*), Conexibacteraceae (*Conexibacter woesei*, 1.5%), and Nocardioidaceae (1.4%), among others.

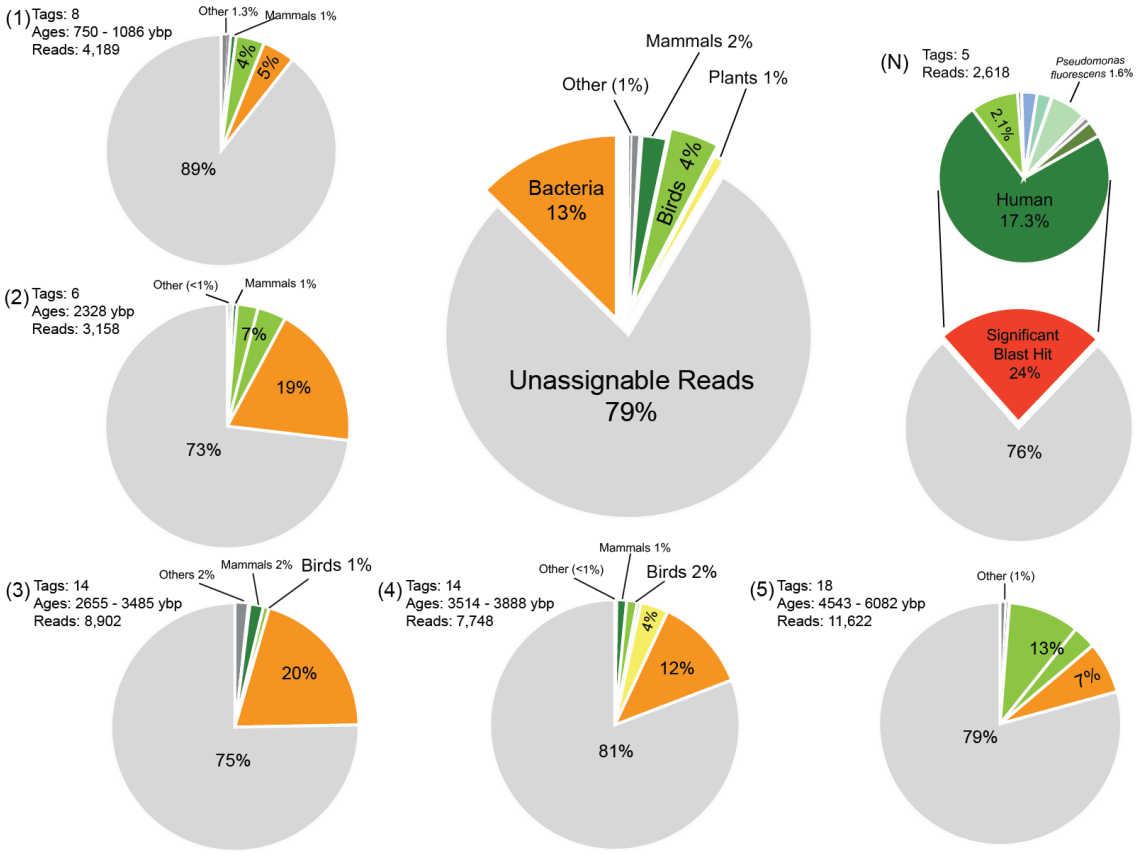


Figure 6.7: Distribution of BLAST hit results for FLX sequencing reads.

Pie charts representing the proportion of reads either unassignable (grey), belonging to birds (light green), human (dark green), bacteria (orange) among others. Results shown for all the reads pooled (center pie chart, for reads split into five age groups and negative controls.

Proteobacteria hits were mainly γ (7.2%) and α (5.1%) proteobacteria. Among γ-proteobacteria representatives, Halomonadaceae were most common (3%), made up of *Chromohalobacter salexigens* and *Halomonas elongata* hits. Within α-proteobacteria, the family Rhodobacteraceae was the most abundant (3.6%), including *Paracoccus denitrificans* and *Ketogulonicigenium vulgare*. Other bacteria classes

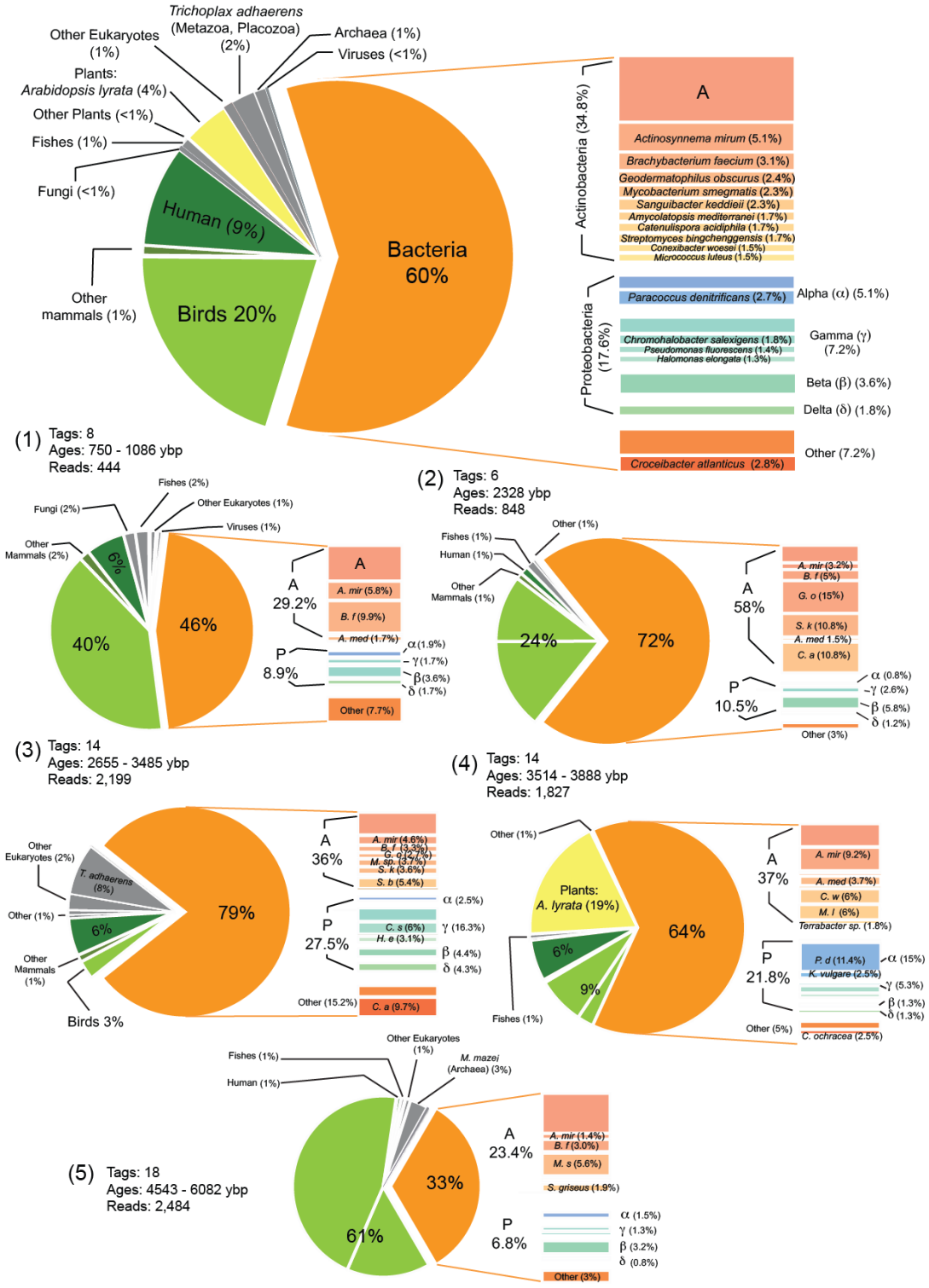


Figure 6.8: Distribution of significant BLAST hits for FLX sequencing reads.

The proportion of groups making up the significant BLAST hits for all FLX reads (top pie chart) as well as split into five age groups is shown.

made up 7.2% of the total significant reads. The most abundant family of these other classes was Flavobacteriaceae (3.5%), from Bacteroidetes, due to *Croceibacter atlanticus* (present in two samples 96.1 and 95.1, Cape Ross) (2.8%).

When the bacterial reads were divided into age groups, Actinobacteria was still the most represented phylum, followed by Proteobacteria, for all groups. For age groups 3 and 4 this was true by a smaller margin than for groups 1, 2 and 5. *Actinosynnema mirum* was present in all age groups, as was *Brachybacterium faecium*, though percentages of this species fluctuated (lowest in group 4, 0.8% of significant reads). *Geodermatophilus obscurus* was present in group 2 mainly, from samples 5.4 and 4.1, Inexpressible Island. In group 3, the three samples from Cape Hickey all contained *G. obscurus* reads (48.1, 46.1, 45.2). Group 5 contained less than ten reads of *G. obscurus*, from three samples. *Sanguibacter keddieii* was most prevalent in groups 2 (samples 5.4 (tag 61) and 5.3 (tag 59) (Inexpressible Island)) and 3 (samples 14.3b (Cape Roberts), D1.1 (Depot Island), and 45.2 (Cape Hickey). Mycobacteriaceae reads were negligible in age groups 1 and 2, and abundant in 3 (3.9%), 4 (2.6%) and 5 (6.4%). Streptomycetaceae presented a similar pattern, negligible in groups 1 and 2 but present in 3 (5.8%), 4 (1.6%) and 5 (4.9%). Micrococcaceae also exhibited the same pattern, only represented by a handful of reads in groups 1 and 2, but present in 3 (3.7%), 4 (6.2%), and to a much lesser extent, 5 (0.7%). Pseudonocardiaceae (mostly *Amycolatopsis mediterranei*) were present in all age groups (0.7-3.8%), though most abundant in group 4 (3.8%). *Conexibacter woesei* reads originated almost exclusively from group 4.

Proteobacteria reads varied between groups. A-proteobacteria dominated for group 4 (15%) due to a high number of *Paracoccus denitrificans* reads. In groups 1 and 2, betaproteobacteria dominated. In group 3, γ-proteobacteria was prevalent, with a high proportion of *Chromohalobacter salexigens* and *Halomonas elongata*. In group 5, β-proteobacteria representatives were more common.

The negative extracts, also subjected to the same experimental protocol as the ancient samples, showed different results (Fig. 6.7). Of the 24% significant BLAST hits, the majority was human (17.3% of the total reads; approximately 70% of the significant

reads). Bacteria, which formed the majority of reads in the four out of five age groups (Fig 6.8.) were present in only very small numbers. Approximately 2.1% of the total reads (8.8% of significant reads) were avian in origin. However, when these reads were mapped to intron reads, none corresponded to any of the multiplex amplicons and appear to represent spurious amplification products. Approximately 1.6% of the total reads were *Pseudomonas fluorescens* (γ-proteobacteria), which did not figure significantly in any of the five age groups.

6.5 Discussion

In the present study, three introns were targeted from ancient sub-fossil Adélie penguin samples using a multiplex PCR enrichment followed by direct tagging and sequencing on a second-generation FLX sequencing platform. This has not been attempted before, to the best of our knowledge, and therefore represents a first attempt to obtain a large amount of nuclear sequence data from thirty ancient samples in parallel. As a result of this, inevitably this study could, and did, indicate that the method implemented would need adjusting to obtain the best results. While the data collected were less than anticipated (on average +/- 200bp coverage for locus *AK1i5* for ten samples instead of full coverage of three introns for thirty samples), ancient nuclear DNA for a number of ancient samples up to 6500 years of age were obtained successfully, and valuable results were obtained that will be of use in future research using second-generation sequencing.

6.5.1 Direct Multiplex FLX Sequencing

The number of total reads obtained for ancient intron fragments was very small compared to the total number of reads (2038 out of over 38,000). This could be increased by sequencing on a larger portion of an FLX plate, however, as is discussed below, this would not be efficient due to the high number of contaminant sequences in the library. Prior to further sequencing, the method implemented needs modification.

Of the three introns targeted, only adenylate kinase intron 5 target fragments were recovered. Overall this marker performed best in singleplex PCR tests during multiplex PCR optimization. Fragment size was not related to success, as the average length of fragments that were sequenced and those that were not was the same (approximately 117bp). It could be that redesigning primers to amplify smaller fragments could marginally improve these results, however, the bigger issue is where to position primers in order to obtain overlapping reads, while avoiding variable positions in priming sites. Primers were designed based on variable sites observed in up to twenty modern sequences. Variability of these introns is high in Adélie penguins (see Chapter 5) and therefore it is possible that primers could have been positioned in such a way that variable sites would interfere with priming sites. It is also possible that ancient intron sequences could contain variable sites not present in the modern sample. Aside from single nucleotide changes, both *AK1i5* and *UCHL3* have been shown to contain insertion deletion polymorphisms (Chapters 4 and 5). Locus *UCHL3* in particular was shown to contain two separate instances of indels in Adélie penguins, and unknown indels would certainly interfere with primer annealing. It is apparent, therefore, that one possible flaw in this method could be the reliance on multiple small internal intron fragments and the difficulty of finding enough invariant sites to position primers to amplify these fragments. Therefore, and importantly, methods not reliant on PCR may be more appropriate for the recovery of ancient nuclear introns.

Almost 94% of all reads obtained through FLX sequencing were contaminants (that is, not belonging to the target amplicons). This is comparable to ancient DNA studies using second-generation shotgun sequencing methods, for example Neanderthal genomic DNA sequencing (GREEN *et al*. 2006; NOONAN *et al*. 2006). Noonan *et al*. (2006) reported recovering 7,880 Neanderthal sequence reads from a total of 1.47 million (0.005%). The present study generated 6% of sequence reads belonging to target intron sequences, showing an improvement compared to shotgun ancient genomic sequencing. If compared to multiplex PCR enriched second-generation sequencing of ancient mitochondrial targets, however, the sequencing success of this study is much lower. Stiller *et al*. (2009) reported upwards of 60% of reads on target. The implications for obtaining ancient nuclear intron sequence using this method are

discussed below. Among the fragments that were successfully sequenced, there was an amplification bias of two fragments over all the rest. Reducing the number of cycles from fifty could help address this, however, the number of reads originating from the amplicons was still far below the background contamination present in the samples. This method is potentially not applicable as such to nuclear introns due to their very low copy. Singleplex PCRs following the first multiplex stage would have certainly increased read numbers and coverage of all three introns, however this would have been time and resource consuming.

6.5.2 Ancient Adélie adenylate kinase intron 5 sequences

Ancient and modern *AK1i5* sequences did not differ significantly from each other, as shown by similar estimates for evolutionary models, diversity indices, base pair frequencies, and a non-significant F_{st}. Several haplotypes were specific to either ancient or modern sequences, however, and a shift in haplotype frequencies was observed through the construction of a temporal haplotype network. This slight shift is indicative of micro-evolutionary change in Adélie penguins. A study from one Adélie penguin colony (Inexpressible Island) at two time points 6,000 years apart detected significant gene frequency shifts at nine microsatellite loci (SHEPHERD *et al*. 2005). It would be of interest to extend sampling to more nuclear intron loci and individuals, particularly from the oldest layer included in this study (6,000 years approximately) or older, to further investigate this result.

The mutation rate estimate obtained from BEAST was high compared to expectations (10^{-9} expected, Chapter 5). Certain introns may be evolving slightly faster or slower than the average rate of evolution for noncoding loci, however, estimates of the rate reported here would benefit from additional studies. The mutation rate of the mitochondrial control region in Adélie penguins has been estimated using ancient DNA and pedigree methods and found to be 0.55 substitutions per site per million years, or 5.5×10^{-7} s/s/y. An evolutionary rate of 1.71×10^{-6} s/s/y for a nuclear non-coding region is most likely an overestimate, as it is unlikely that the intron studied here has a rate ten times higher than that of the mitochondrial control region,

generally thought to accumulate substitutions more quickly than nuclear introns (FRIESEN 2000). This overestimation could be due to a lack of signal in the data, which may cause BEAST to overestimate the mutation rate, particularly in sequences with low nucleotide diversity (DEBRUYNE and POINAR 2009; HO *et al*. 2011). This lack of signal may in part be due to low mutation rates for non-coding single copy nuclear DNA, implying few changes have accumulated in 6000 years in Adélie penguins. Another potential factor in this low signal could be due to sequence length obtained. Only two fragments of one intron were obtained with any reasonable coverage, and short fragments will contain fewer informative sites. This estimate, however, represents the first attempt to estimate a rate of evolution for a nuclear non-coding region from ancient DNA sequences, and as such has great value, particularly by helping understand what further research is needed to achieve a more robust estimate. To improve on this estimate using ancient DNA methodology, in the future the sampling needs to be extended to individuals over 6000 years of age (up to 40,000) and over longer non-coding intron sequences. However, any future attempt to apply this method to a rate estimation of nuclear DNA may be subject to a certain degree of overestimation due to model violation, e.g., the presence of recombination (NAVASCUÉS *et al*. 2010).

6.5.3 Contaminant Sequences

Of the reads obtained, when subjected to BLAST analysis, most (79%) were not significantly similar to any available sequence on GenBank. This may be due to a combination of short read sizes and sequencing of species potentially not sequenced before. This high proportion of non-significant hits is typical of ancient DNA second-generation sequencing runs (GREEN *et al*. 2006). DNA extractions of Adélie subfossil bones were carried out in a dedicated ancient DNA laboratory, adhering to strict protocols to minimize contamination. As negative extractions were performed simultaneously (without bone material), these offer a way to gauge contamination. In this study, extraction negative controls were also subjected to multiplex PCR, tagging and FLX titanium sequencing. These negative controls produced slightly lower than average number of reads compared to the extracts from ancient bone. BLAST results

of these negative controls showed a similar proportion of unidentifiable reads. The identifiable reads consisted mainly of human DNA, *Pseudomonas fluorescens*, and reads that apparently mapped to the myelin proteolipid protein, but when compared to the reference intron *MPP4* it was found that the reads were primer dimer fragments or spurious amplification products. Reads from bacteria were in the minority. This differs to the contaminant reads from all pooled amplicon sequences, which contained a much higher proportion of bacterial reads than human. This suggests that contaminants within the negative controls originated during any or all of the various steps prior to sequencing, but in any case not from bone samples used for DNA extraction. This gives us confidence that there has been little or no cross-contamination or contamination from modern Adélie penguin PCR products. Contaminant reads from Adélie penguin intron amplicon tagged reads, therefore, are likely to have originated from the bone samples themselves.

Within Antarctica, ornithogenic soils below current and abandoned Adélie penguin colonies are the only soils south of the Antarctic circle with high concentrations of organic matter, originating from inputs in the form of penguin guano, eggshell, feathers, and other bird remains (SPEIR and ROSS 1984). Microbial biomass, respiration, and other indices are higher in these soils than mineral soils studied, though bacterial diversity does not appear to be any higher (AISLABIE *et al*. 2009). Few fungi, plants or Archaea have been found in Antarctic soils, which are dominated by bacteria. The bacterial composition varies by geographic region, presence or absence of Adélie penguins, and depth in the soil profile (AISLABIE *et al*. 2009). Uric-acid degrading bacteria are present in large numbers, more so in currently occupied colonies with fresh inputs of guano. Endospore producers are also quite common, a characteristic that allows a bacterium the ability to survive inhospitable conditions (AISLABIE *et al*. 2009). Acidobacteria, Actinobacteria and Bacteroidetes phyla were well represented in Ross Sea mineral soils (AISLABIE *et al*. 2008), while Firmicutes (similar to *Oceanobacillus*) and γ-proteobacteria were more common in ornithogenic soils (AISLABIE *et al*. 2009; AISLABIE *et al*. 2008). The composition of contaminant reads from the sixty tagged amplicon pools showed a clear dominance of bacteria, and very few plant, virus, archaea, or other, among the significant BLAST hits, consistent with the general profile of Antarctic soils. The distribution of phyla within bacteria

for the tagged samples showed an abundance of Actinobacteria, followed by Proteobacteria, mainly γ-proteobacteria. This pattern is more indicative of mineral soils; soil composition data for each site from which ancient samples were obtained could provide an interesting additional dimension to our understanding of these contaminant reads. The majority of the families identified correspond to families previously found in Antarctic soils, though the proportions varied somewhat. No clear age-related compositional change was found at present, except for slight variations in families present. For example, in older samples, there was a greater proportion of Mycobacteriaceae, Streptomycetaceae and Micrococcaceae compared to younger samples. These and any other changes could be due to differences in bacterial composition in the depth of the soil rather than age of the sample, as these two factors are likely to be correlated. Due to the large proportion of unassignable reads and potential mispriming during PCR enriching certain bacterial groups over others, not much can be said at present regarding the proportional representation of bacterial contaminants compared to those known for Antarctic soils. In general, however, it is clear that the contaminants found in this study fit with the general profile of Antarctic soils, and differ from that found for the negative controls, which suggests sample-processing conditions were for the most part clean and sequences generated are ancient in origin. A high presence of soil-living Actinobacteria (particularly Actinomycetales) has also been reported previously for other ancient DNA second-generation sequencing studies (GREEN *et al*. 2006), further supporting the authenticity of the sequences found during this study.

6.5.4 Conclusions

Ancient single-copy nuclear Adélie penguin DNA belonging to the fifth intron of the adenylate kinase gene was recovered and sequenced using second-generation sequencing in this study. Sequence data revealed a shift in haplotype frequencies from the oldest samples, compared to modern *AK1i5* sequences. This study illustrates the potential of second-generation sequencing for recovering large amounts of nuclear data for population level studies using ancient DNA. As it is also the first attempt to

recover population level nuclear data from ancient samples, valuable insight into the use of second-generation sequencing for this purpose was also gained.

Direct multiplex PCR FLX sequencing methodology requires some adjustment to recover ancient nuclear sequence more efficiently. Low copy number of ancient nuclear DNA implies a high number of PCR cycles is needed, however a high number of cycles creates an amplification bias of certain amplicon fragments over others, and endogenous contaminant sequences still outnumber amplified sequences. Increasing coverage by sequencing across a larger section of an FLX plate would help increase the number of target sequences obtained. This would also reduce the possibility of obtaining sequences originating from a damaged DNA strand containing a miscoding lesion – increased coverage would improve our ability to identify DNA damage. An issue with relying on PCR for enrichment is the necessity of finding adequate priming sites and fragment length, shown to be non-trivial in the present case. Also, a PCR enrichment method targets only those templates above a certain size. DNA capture via hybridization, in which target DNA is hybridized with specific probes immobilized on a microarray or on beads, may be a better option for ancient nuclear sequences (KNAPP and HOFREITER 2010). Utilizing a capture method would ensure the smaller fragment sizes, generally more common in an ancient DNA extract, would not be discarded. Also, contaminant sequences would be greatly reduced. The current study was limited by a lower fragment size of 100bp including primers due to the purification method used. Only a handful of ancient DNA studies have used capture methods so far (e.g. (BRIGGS *et al*. 2009). One caveat to applying this method is that library preparation is carried out from aDNA extracts, which would require optimizing for very low copy numbers – particularly in the case of nuclear introns. Library amplification would be needed prior to hybridization (KNAPP and HOFREITER 2010). Further research on these methods needs to be carried out before they can be applied to population-level ancient nuclear studies. However, capture methods may be best suited for nuclear aDNA studies when compared to PCR methods due to the much shorter average fragment length obtainable when compared to mtDNA studies.

Second-generation sequencing methods hold great potential for extending ancient DNA population studies by incorporating nuclear markers as well as mitochondrial.

Despite the difficulties highlighted in the present study, the promise of multi-locus nuclear approaches in ancient DNA is quickly becoming a reality. Current methodology is improving continuously through being tested, as the present study shows.

6.6 References

AISLABIE, J., S. JORDAN, J. AYTON, J. L. KLASSEN, G. M. BARKER *et al.*, 2009 Bacterial diversity associated with ornithogenic soil of the Ross Sea region, Antarctica. Canadian Journal of Microbiology **55:** 21-36.

AISLABIE, J. M., S. JORDAN and G. M. BARKER, 2008 Relation between soil classification and bacterial diversity in soils of the Ross Sea region, Antarctica. Geoderma **144:** 9-20.

AKAIKE, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19:** 716-722.

AXELSSON, E., E. WILLERSLEV, M. T. P. GILBERT and R. NIELSEN, 2008 The Effect of Ancient DNA Damage on Inferences of Demographic Histories. Molecular Biology and Evolution **25:** 2181-2187.

BAKER, A. J., L. J. HUYNEN, O. HADDRATH, C. D. MILLAR and D. M. LAMBERT, 2005 Reconstructing the tempo and mode of evolution in an extinct clade of birds with ancient DNA: The giant moas of New Zealand. PNAS **102:** 8257-8262.

BARNES, I., B. SHAPIRO, A. LISTER, T. KUZNETSOVA, A. SHER *et al.*, 2007 Genetic Structure and Extinction of the Woolly Mammoth, Mammuthus primigenius. Current Biology **17:** 1072-1075.

BINLADEN, J., M. T. P. GILBERT, J. P. BOLLBACK, F. PANITZ, C. BENDIXEN *et al.*, 2007 The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. PLoS ONE **2:** e197.

BRIGGS, A. W., J. M. GOOD, R. E. GREEN, J. KRAUSE, T. MARICIC *et al.*, 2009 Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. Science **325:** 318-321.

BUNCE, M., T. H. WORTHY, M. J. PHILLIPS, R. N. HOLDAWAY, E. WILLERSLEV *et al.*, 2009 The evolutionary history of the extinct ratite moa and New Zealand Neogene paleogeography. Proceedings of the National Academy of Sciences **106:** 20646-20651.

CAMPBELL, K. L., J. E. E. ROBERTS, L. N. WATSON, J. STETEFELD, A. M. SLOAN *et al.*, 2010 Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. Nature Genetics **42:** 536-540.

CLEMENT, M., D. POSADA and K. CRANDALL, 2000 TCS: a computer program to estimate gene genealogies. Molecular Ecology **9:** 1657 - 1659.

DEBRUYNE, R., and H. N. POINAR, 2009 Time Dependency of Molecular Rates in Ancient DNA Data Sets, A Sampling Artifact? Systematic Biology **58:** 348-360.

DMITRIEV, D. A., and R. A. RAKITOV, 2008 Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels. PLoS Computational Biology **4:** e1000113.

DRUMMOND, A., and A. RAMBAUT, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology **7:** 214.

DRUMMOND, A. J., B. ASHTON, S. BUXTON, M. CHEUNG, A. COOPER *et al.*, 2010 Geneious v5.1. Available from http://www.geneious.com.

DRUMMOND, A. J., and A. RAMBAUT, 2003 BEAST version 1.3, pp. Oxford University Press, Oxford.

EXCOFFIER, L., and H. E. L. LISCHER, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10:** 564-567.

FLOT, J.-F., 2007 CHAMPURU 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Molecular Ecology Notes **7:** 974-977.

FLOT, J.-F., 2010 SeqPHASE: a web tool for interconverting PHASE input/output files and FASTA sequence alignments. Mol Ecol Resour **10:** 162 - 166.

FRIESEN, V. L., 2000 Introns, pp. 274-294 in *Molecular Methods in Ecology*, edited by A. J. BAKER. Blackwell Science Ltd, Oxford.

FU, Y.-X., 1997 Statistical tests of neutrality against population growth, hitchhiking and background selection. Genetics **147:** 915 - 925.

GILBERT, M. T. P., J. BINLADEN, W. MILLER, C. WIUF, E. WILLERSLEV *et al.*, 2007 Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. Nucl. Acids Res. **35:** 1-10.

GOUY, M., S. GUINDON and O. GASCUEL, 2010 SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution **27:** 221-224.

GREEN, R. E., J. KRAUSE, A. W. BRIGGS, T. MARICIC, U. STENZEL *et al.*, 2010 A Draft Sequence of the Neandertal Genome. Science **328:** 710-722.

GREEN, R. E., J. KRAUSE, S. E. PTAK, A. W. BRIGGS, M. T. RONAN *et al.*, 2006 Analysis of one million base pairs of Neanderthal DNA. Nature **444:** 330-336.

GREEN, R. E., A.-S. MALASPINAS, J. KRAUSE, A. W. BRIGGS, P. L. F. JOHNSON *et al.*, 2008 A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. Cell **134:** 416-426.

GREENWOOD, A. D., C. CAPELLI, G. POSSNERT and S. PÄÄBO, 1999 Nuclear DNA Sequences from Late Pleistocene Megafauna. Molecular Biology and Evolution **16:** 1466-1473.

HAILE, J., D. G. FROESE, R. D. E. MACPHEE, R. G. ROBERTS, L. J. ARNOLD *et al.*, 2009 Ancient DNA reveals late survival of mammoth and horse in interior Alaska. PNAS **106:** 22352-22357.

HAY, J. M., S. SUBRAMANIAN, C. D. MILLAR, E. MOHANDESAN and D. M. LAMBERT, 2008 Rapid molecular evolution in a living fossil. Trends in Genetics **24:** 106-109.

HO, S. Y. W., R. LANFEAR, M. J. PHILLIPS, I. BARNES, J. A. THOMAS *et al.*, 2011 Bayesian Estimation of Substitution Rates from Ancient DNA Sequences with Low Information Content. Systematic Biology **60:** 366-375.

HO, S. Y. W., M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2005 Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Molecular Biology and Evolution **22:** 1561-1568.

HO, S. Y. W., B. SHAPIRO, M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2007 Evidence for Time Dependency of Molecular Rate Estimates. Systematic Biology **56:** 515-522.

HORN, S., W. DURKA, R. WOLF, A. ERMALA, A. STUBBE *et al.*, 2011 Mitochondrial Genomes Reveal Slow Rates of Molecular Evolution and the Timing of Speciation in Beavers (*Castor*), One of the Largest Rodent Species. PLoS ONE **6:** e14622.

HUYNEN, L., B. J. GILL, C. D. MILLAR and D. M. LAMBERT, 2010 Ancient DNA reveals extreme egg morphology and nesting behavior in New Zealand's extinct moa. PNAS **107:** 16201-16206.

HUYNEN, L., C. D. MILLAR, R. P. SCOFIELD and D. M. LAMBERT, 2003 Nuclear DNA sequences detect species limits in ancient moa. Nature **425:** 175-178.

KNAPP, M., and M. HOFREITER, 2010 Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. Genes **1:** 227-243.

KRAUSE, J., P. H. DEAR, J. L. POLLACK, M. SLATKIN, H. SPRIGGS *et al.*, 2006 Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature **439:** 724-727.

KRINGS, M., H. GEISERT, R. W. SCHMITZ, H. KRAINITZKI and S. PÄÄBO, 1999 DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. PNAS **96:** 5581-5585.

LALUEZA-FOX, C., A. ROSAS, A. ESTALRRICH, E. GIGLI, P. F. CAMPOS *et al.*, 2011 Genetic evidence for patrilocal mating behavior among Neandertal groups. PNAS **108:** 250-253.

LALUEZA-FOX, C., M. L. SAMPIETRO, D. CARAMELLI, Y. PUDER, M. LARI *et al.*, 2005 Neandertal Evolutionary Genetics: Mitochondrial DNA Data from the Iberian Peninsula. Molecular Biology and Evolution **22:** 1077-1081.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of Evolution in Ancient DNA from Adélie Penguins. Science **295:** 2270-2273.

LEONARD, J. A., 2008 Ancient DNA applications for wildlife conservation. Molecular Ecology **17:** 4186-4196.

MARICIC, T., M. WHITTEN and S. PÄÄBO, 2010 Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE **5:** e14004.

MEYER, M., A. W. BRIGGS, T. MARICIC, B. HÖBER, B. HÖFFNER *et al.*, 2008a From micograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Research **36:** e5.

MEYER, M., U. STENZEL and M. HOFREITER, 2008b Parallel tagged sequencing on the 454 platform. Nature Protocols **3:** 267-278.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al.*, 2008a Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

MILLAR, C. D., L. HUYNEN, S. SUBRAMANIAN, E. MOHANDESAN and D. M. LAMBERT, 2008b New developments in ancient genomics. Trends in Ecology & Evolution **23:** 386-393.

MILLER, W., D. I. DRAUTZ, A. RATAN, B. PUSEY, J. QI *et al.*, 2008 Sequencing the nuclear genome of the extinct woolly mammoth. Nature **456:** 387-390.

MORIN, P. A., F. I. ARCHER, A. D. FOOTE, J. VILSTRUP, E. E. ALLEN *et al.*, 2010 Complete mitochondrial genome phylogeographic analysis of killer whales (Orcinus orca) indicates multiple species. Genome Research **20:** 908-916.

NAVASCUÉS, M., F. DEPAULIS and B. C. EMERSON, 2010 Combining contemporary and ancient DNA in population genetic and phylogeographical studies. Molecular Ecology Resources **10:** 760-772.

NAVASCUÉS, M., and B. C. EMERSON, 2009 Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. Molecular Ecology **18:** 4390-4397.

NOONAN, J. P., G. COOP, S. KUDARAVALLI, D. SMITH, J. KRAUSE *et al.*, 2006 Sequencing and Analysis of Neanderthal Genomic DNA. Science **314:** 1113-1119.

RITCHIE, P. A., C. D. MILLAR, G. C. GIBB, C. BARONI and D. M. LAMBERT, 2004 Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. Molecular Biology and Evolution **21:** 240-248.

ROGAEV, E. I., Y. K. MOLIAKA, B. A. MALYARCHUK, F. A. KONDRASHOV, M. V. DERENKO *et al.*, 2006 Complete Mitochondrial Genome and Phylogeny of Pleistocene Mammoth *Mammuthus primigenius*. PLoS Biology **4:** e73-e80.

ROHLAND, N., and M. HOFREITER, 2007a Ancient DNA extraction from bones and teeth. Nature Protocols **2:** 1756-1762.

ROHLAND, N., and M. HOFREITER, 2007b Comparison and optimization of ancient DNA extraction. BioTechniques **42:** 343-352.

RÖMPLER, H., N. ROHLAND, C. LALUEZA-FOX, E. WILLERSLEV, T. KUZNETSOVA *et al.*, 2006 Nuclear Gene Indicates Coat-Color Polymorphism in Mammoths. Science **313:** 62.

SERRE, D., A. LANGANEY, M. CHECH, M. TESCHLER-NICOLA, M. PAUNOVIC *et al.*, 2004 No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. PLoS Biology **2:** 0313-0317.

SHAPIRO, B., A. J. DRUMMOND, A. RAMBAUT, M. C. WILSON, P. E. MATHEUS *et al.*, 2004 Rise and Fall of the Beringian Steppe Bison. Science **306:** 1561-1565.

SHEPHERD, L. D., C. D. MILLAR, G. BALLARD, D. G. AINLEY, P. R. WILSON *et al.*, 2005 Microevolution and mega-icebergs in the Antarctic. PNAS **102:** 16717-16722.

SPEIR, T. W., and D. J. ROSS, 1984 Ornithogenic soils of the Cape Bird adelie penguin rookeries, Antarctica. Polar Biology **2:** 207-212.

STEPHENS, M., N. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68:** 978 - 989.

STILLER, M., M. KNAPP, U. STENZEL, M. HOFREITER and M. MEYER, 2009 Direct multiplex sequencing (DMPS) -- a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Research **19:** 1843-1848.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585 - 595.

TAMURA, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Molecular Biology and Evolution **9:** 678-687.

TAMURA, K., D. PETERSON, N. PETERSON, G. STECHER, M. NEI *et al.*, 2011 MEGA5: Molecular Evolutionary Genetics Analysis using Likelihood, Distance, and Parsimony methods. . Molecular Biology and Evolution **28:** 2731-2739.

7 Chapter Seven

CONCLUSIONS AND PERSPECTIVES

7.1 Introduction

The notion of model organisms has changed in recent times. This term was typically used for species such as zebra fish, mouse, *Drosophila melanogaster*, etc, i.e. species for which very large amounts of broadly biological data, including genomic data, are available. However this concept is clearly in need of revision. At best, the distinction between model species and others is now less clear. The study of molecular ecology and evolution in 'non-model organisms' has become increasingly diverse, as advances in molecular techniques and statistics applied to biological scenarios have improved. However, most studies in these fields, working on poorly characterized species or species present only in small numbers, rely heavily on work carried out in the 'model organisms' of the past. While this has certainly conveyed huge benefits, a new generation of extensively characterized model organisms is the next development in this rapidly growing field. Certainly these studies can tell us more about the genetic diversity of wild populations of species, subject to more natural, complex population history than the classical model organisms living in highly controlled experimental environments.

Adélie penguins have been the subject of research in many different fields. Antarctica, home to Adélie penguins, has also been extensively researched. This combination has proven powerful, as understanding the changing climate of Antarctica has helped researchers understand the population history of the species while at the same time, patterns of occupation through time has helped researchers understand past climates and sea-ice coverage. Continuous molecular research on Adélie penguins has provided not only a good understanding of the population history of this species, but has yielded a more general understanding of evolutionary processes themselves. In this thesis, new findings shed light on Adélie penguin biology, as well as on the application of molecular markers to population genetics and phylogenetics in general.

7.2 Thesis Summation

In this thesis, I sought to extend our understanding of Adélie penguin population history, and concurrently with this objective, address more general questions of molecular ecology and evolution. Extensive research has been carried out using mitochondrial DNA markers in this species, providing a large number of samples and sequences to work with and a 'tried and tested' mutation rate estimated for the mitochondrial *HVRI* region. Generally, these conditions are not met in other species. Using these resources, in **Chapter Two** I addressed the question of whether population size and mitochondrial genetic diversity in the Adélie penguin were positively correlated. Studies have suggested that mitochondrial diversity may be subject to recurrent selective sweeps due to linkage with loci under selection (BAZIN *et al*. 2006). These could mimic the effects of population size reduction. Using robust population count data for colonies of widely varying sizes and genetic diversity estimates for these colonies I found a positive correlation between population size and mitochondrial control region diversity, supporting the use of mitochondrial DNA for population inference for species up to the population sizes of Adélie penguins. Population sizes estimated from genetic diversity were on the whole smaller than current sizes obtained from direct counts. In addition, coalescence times indicated the origins of the lineages comprising the populations during the last glacial maximum.

During this period ice-free areas would have been greatly reduced around Antarctica. This is important given the requirement of the species for ice-free nesting sites.

Despite the many advantages of using mitochondrial DNA markers for phylogeography and population studies, the mitochondrial genome does represent only one genetic signature and, increasingly, independent assessments using multiple nuclear loci are being implemented. Agreement or disagreement between mitochondrial and nuclear markers can refine our understanding of population history thanks to the different coalescence times of these markers, and also can help us distinguish between the effects of selection and drift. Drift should affect all genomic regions, whereas natural selection will be localized. In phylogenetics, incorporating multiple loci from different, unlinked genomic regions helps us approach the true species tree more closely, rather than individual gene trees. In Adélie penguins, the majority of the research carried out on the molecular ecology of the species has used mitochondrial DNA markers. Nuclear DNA markers, aside from microsatellites, have not been used to complement mitochondrial studies. In **Chapter Four**, I investigated published primer pairs amplifying nuclear introns to identify nuclear markers that could be applied to phylogeography and phylogenetics of penguins. Six introns were identified from this study that could prove to be useful markers (3015bp). Four of these markers have been used in other avian molecular ecology studies (*AK1i5*, *MPP4*, *ODC6*, *ACTB2*), and two have not at present been used in any studies (*HMG2* and *UCHL3*). Initial variability of these markers showed a SNP density of one SNP per 72 bp, comparable to that found in kelp gull (*Larus dominicanus*) (DE MENDONÇA DANTAS *et al*. 2009). Insertion-deletion polymorphisms were also present in several of the markers obtained (*AK1i5*, *UCHL3*), which complicated direct sequencing efforts. Aside from providing tested markers for Adélie penguin population genetics, a minimum of five other markers that could be amenable for this purpose were also identified. In **Chapter Five**, five of these intron markers (*AK1i5*, *MPP4*, *ODC6* and *HMG2*) were then used to assess evidence for population structure, in particular to investigate whether the highly diverged Antarctic (A) and Ross Sea (RS) mitochondrial lineages are present in the nuclear genome. These lineages were not detected in the five introns studied, and no other evidence for population structure was found. Longer coalescence times and lower mutation rates for nuclear versus

mitochondrial regions imply that these populations were most likely not separated long enough for nuclear genes to show this divergence. This provides support for previous research. Also, recombination and biparental inheritance after gene flow between the two putative refugial last glacial maximum populations would have served to mask this divergence. Population expansion was detected for these markers, providing evidence for a signature older than that previously found in mitochondrial markers, indicative of a species that has undergone recurring expansions and declines coupled with changes in Antarctic climate. The phylogenetic utility of four intron markers (*AK1i5*, *UCHL3*, *MPP4* and *ODC6*) was also assessed in this chapter. Representatives of all extant penguin genera, including most penguin species, were amplified for the four introns using one touchdown PCR program, and sequenced in both directions. The markers tested in this thesis are therefore immediately applicable to phylogeography of other penguin species, as they cross-amplify in all species tested. A phylogeny generated from the four concatenated intron datasets (1926 bp) recovered the majority of splits between penguin species with significant support, when compared to a previous phylogeny (BAKER *et al*. 2006). Tip resolution among *Eudyptes* species, however, was low, and *Pygoscelis* species appear basal to the whole penguin phylogeny. Previous work relying on mitochondrial genes and a nuclear exon, *RAG-1*, showed *Aptenodytes* species to be basal. More nuclear intron sequence as well as a number of well-placed outgroup species should help clarify if this is indicative of an incongruence. Individually, none of the markers could adequately resolve the phylogeny. A substitution rate was also estimated, by calibrating a concatenated intron-only genealogy using two putative times to the most recent common ancestor for extant penguins, 40 million years ago (BAKER *et al*. 2006) and 15 million years ago (CLARKE *et al*. 2007). Population parameters estimated using the evolutionary rate derived from the younger divergence time appeared to be the most parsimonious, indicating that the 40 million year old divergence time is most likely far older than the true divergence time of extant penguins. In **Chapter Six**, a protocol developed to obtain full mitochondrial genomes in one multiplex PCR step followed by tagging and sequencing on the FLX second generation sequencing platform was applied to the recovery of three nuclear introns from Adélie penguin ancient samples up to 6500 years old. Internal multiplex primers were designed for this purpose and thirty sub-fossil bone samples were processed in replicate, along with five extraction

negative controls. This represented the first attempt to obtain ancient nuclear population data using second-generation sequencing. Ancient nuclear DNA recovered was low in comparison to mitochondrial DNA recovered in a previous study using this method, and background contamination out-represented target nuclear DNA. These results are of importance for future research using ancient nuclear DNA for population-level studies. PCR-based enrichment prior to second-generation sequencing may not be the best approach to obtain ancient nuclear intron sequences, due to fragment size limits and primer design constraints. Hybridization capture methods may offer a more effective way to recover population-level ancient nuclear DNA. The contaminant sequences recovered corresponded mostly to bacteria and presented similar composition profiles to those found for Antarctic soils. This confirms the origin of the contaminants as belonging to the sub-fossil bones and not due to laboratory mishandling, which provides confidence in the authenticity of the results obtained. The ancient intron sequences recovered presented a small number of sequence differences when compared to modern sequences, however summary statistics were on the whole similar between different time periods. A shift in haplotype frequencies from the oldest samples to modern was also detected.

7.3 Future Perspectives

The work carried out for this thesis could be carried forward in a number of ways. This thesis presents the first ever use of nuclear intron markers for Adélie penguin population genetics, and therefore focused on sample sizes large enough to identify promising markers and address some population questions. Increasing sampling by adding one or more intron markers, and sequencing at least ten individuals from a range of Adélie penguin colonies throughout their range, would allow a more detailed analysis of the structure of Adélie penguins at nuclear loci. It would be interesting to see whether the relationship found between population size and mitochondrial diversity would hold true for nuclear regions, though the lack of structure found at nuclear loci most likely precludes this. Adding more intron markers and penguin species to the phylogenetic study, as well as suitable outgroup species, should come close to resolving the penguin phylogeny, which could then be compared more

thoroughly to the mostly-mitochondrial phylogeny published by Baker *et al* (2006). Also, the possibilities of applying the introns identified in this thesis to other penguin species are wide-ranging. No research has been carried out on other penguin species using intron markers, aside from broader avian phylogenetic studies. For example, ancient and modern mitochondrial DNA analysis discovered a sister species of yellow-eyed penguin (*Megadyptes antipodes*) that went extinct in New Zealand following human settlement (BOESSENKOOL *et al*. 2008). Further characterization of this extinct species using nuclear markers would provide an independent molecular viewpoint. As samples available for this species are ancient in origin, directly sequencing nuclear introns would be problematic. An approach similar to that described in Chapter Six using second-generation sequencing could circumvent these problems. However, modifying this approach would be necessary for obtaining ancient nuclear intron data from this or the Adélie penguin, or, very likely, any other species. Nuclear DNA is fragmented and less abundant than mitochondrial DNA and a one stage multiplex PCR approach does not provide enough amplification products to raise the target sequences above the background contaminants inherent to ancient samples. Sequencing more of the library on a larger section of an FLX plate is necessary to increase coverage of the target sequences. On the other hand, rather than enriching ancient DNA extracts via PCR, a hybridization capture-based enrichment protocol could be developed to allow more efficient targeting of nuclear regions of interest and would not require labor-intensive primer design or size selection caused by constraints to regions between priming sites. As for the use of ancient samples to generate an evolutionary rate for nuclear regions, similar to that obtained previously for mitochondrial DNA regions in Adélie penguins (LAMBERT *et al*. 2002; MILLAR *et al*. 2008; RITCHIE *et al*. 2004; SUBRAMANIAN *et al*. 2009), future research needs to obtain more extensive genomic sequences as well as extend sampling to bones older than 6500 ybp. However, in parallel, the statistical methodology used to estimate these rates needs to be updated to be able to take into account recombination, among other things.

On the whole, this study illustrates the promise of new technologies and combined research approaches to improve our understanding of the ecology and evolution of 'non-model' species, and hopefully sheds light on future directions for this research.

7.4 References

BAKER, A. J., S. L. PEREIRA, O. P. HADDRATH and K.-A. EDGE, 2006 Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. Proceedings of the Royal Society Series B **273:** 11-17.

BAZIN, E., S. GLÉMIN and N. GALTIER, 2006 Population size does not influence mitochondrial genetic diversity in animals. Science **312:** 570-571.

BOESSENKOOL, S., J. J. AUSTIN, T. H. WORTHY, P. SCOFIELD, A. COOPER *et al*., 2008 Relict or colonizer? Extinction and range expansion of penguins in southern New Zealand. Proceedings of the Royal Society B: Biological Sciences **276:** 815-821.

CLARKE, J. A., D. T. KSEPKA, M. STUCCHI, M. URBINA, N. GIANNINI *et al*., 2007 Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. PNAS **104:** 11545-11550.

DE MENDONÇA DANTAS, G. P., R. GODINHO, J. S. MORGANTE and N. FERRAND, 2009 Development of new nuclear markers and characterization of single nucleotide polymorphisms in kelp gull (*Larus dominicanus*). Molecular Ecology Resources **9:** 1159-1161.

LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al*., 2002 Rates of Evolution in Ancient DNA from Adélie Penguins. Science **295:** 2270-2273.

MILLAR, C. D., A. DODD, J. ANDERSON, G. C. GIBB, P. A. RITCHIE *et al*., 2008 Mutation and Evolutionary Rates in Adélie Penguins from the Antarctic. PLoS Genetics **4:** e1000209.

RITCHIE, P. A., C. D. MILLAR, G. C. GIBB, C. BARONI and D. M. LAMBERT, 2004 Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. Molecular Biology and Evolution **21:** 240-248.

SUBRAMANIAN, S., D. R. DENVER, C. D. MILLAR, T. HEUPINK, A. ASCHRAFI *et al*., 2009 High mitogenomic evolutionary rates and time dependency. Trends in Genetics **25:** 482-468.

1 APPENDIX ONE

THE MOLECULAR ECOLOGY OF THE EXTINCT NEW ZEALAND HUIA

OPEN ACCESS Freely available online

PLoS one

The Molecular Ecology of the Extinct New Zealand Huia

David M. Lambert[1,2]*, Lara D. Shepherd[2], Leon Huynen[3], Gabrielle Beans-Picón[3], Gimme H. Walter[4], Craig D. Millar[5]

1 Griffith School of Environment and School of Biomolecular and Physical Sciences, Griffith University, Nathan, Australia, **2** Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Auckland, New Zealand, **3** Institute of Natural Sciences, Massey University, Auckland, New Zealand, **4** School of Biological Sciences, The University of Queensland, Brisbane, Australia, **5** Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

Abstract

The extinct Huia (*Heteralocha acutirostris*) of New Zealand represents the most extreme example of beak dimorphism known in birds. We used a combination of nuclear genotyping methods, molecular sexing, and morphometric analyses of museum specimens collected in the late 19th and early 20th centuries to quantify the sexual dimorphism and population structure of this extraordinary species. We report that the classical description of Huia as having distinctive sex-linked morphologies is not universally correct. Four Huia, sexed as females had short beaks and, on this basis, were indistinguishable from males. Hence, we suggest it is likely that Huia males and females were indistinguishable as juveniles and that the well-known beak dimorphism is the result of differential beak growth rates in males and females. Furthermore, we tested the prediction that the social organisation and limited powers of flight of Huia resulted in high levels of population genetic structure. Using a suite of microsatellite DNA loci, we report high levels of genetic diversity in Huia, and we detected no significant population genetic structure. In addition, using mitochondrial hypervariable region sequences, and likely mutation rates and generation times, we estimated that the census population size of Huia was moderately high. We conclude that the social organization and limited powers of flight did not result in a highly structured population.

Citation: Lambert DM, Shepherd LD, Huynen L, Beans-Picón G, Walter GH, et al. (2009) The Molecular Ecology of the Extinct New Zealand Huia. PLoS ONE 4(11): e8019. doi:10.1371/journal.pone.0008019

Editor: Richard Cordaux, University of Poitiers, France

Received June 16, 2009; **Accepted** October 19, 2009; **Published** November 25, 2009

Funding: This research was supported by Griffith University, the Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: d.lambert@griffith.edu.au

Introduction

Huia (*Heteralocha acutirostris*) had the most extreme sex-linked bill dimorphism known in birds [1–3]. Male Huia were thought to have short, stout bills, whereas females were characterised by long curved bills, about a third longer than those of males. Males and females had such distinctive bill morphologies that they were originally described as different species [4,5]. Observations by early naturalists suggested that the species was territorial, that juveniles lived with adults, and that family members cooperated in foraging [6].

The closest relatives of the Huia are the North Isand Saddleback (*Philesturnus carunculatus*) and the North Island Kokako (*Callaeas cinereus*). Like the Huia, these species are also wattlebirds and are characterised by a pair of colourful, fleshy wattles; strong feet; and short, rounded wings [7]. The pre-human distribution of Huia bones in caves, dunes, and middens indicates that they were once common throughout the North Island of New Zealand but were absent from the South Island [8]. Following Polynesian settlement, the species declined. Further reduction ensued as hunting pressure increased, partly because Huia tail feathers became fashionable among Europeans, especially after the Duke of York (later King George V) wore one in his hatband. Huia bills were also commonly used as brooches and these became increasingly popular in the late nineteenth century. Increased hunting, clearance of lowland forest, and the introduction of predators finally led to the Huia's extinction. The last confirmed sighting was in 1907 [9]; however, evidence suggests that the species survived until the 1930s [9 ; Lovis pers. comm.].

Little is known about the behaviour or social structure of Huia, apart from limited observations made by early naturalists. Buller [6] observed that Huia inhabited thick forest and moved mainly on foot 'by a series of bounds or jumps'. Colenso [10] suggested that Huia were social birds, and Buller [6] noted that they were almost always found in pairs and sometimes in groups of four or more. Potts [11] observed Huia young accompanying what he assumed to be their parents for a considerable time after fledging. He gave an account of four juveniles, barely distinguishable from adults, still being fed by their parents. Moorhouse [12] suggested that Huia were highly territorial based on Buller's [6] observation that pairs were attracted by imitations of their call, a trait also common to the Saddleback and Kokako [7]. Their social organisation and limited powers of flight suggests that Huia are likely to have exhibited a high level of population genetic structure.

We examined museum samples collected from both sexes and from a number of locations (Figure 1) to understand the population genetic structure of Huia and the nature of their sexual dimorphism. We used several nuclear genotyping markers isolated from extant Saddleback to amplify ancient DNA from Huia. Using rigorous ancient DNA methodologies, we were able to determine the genotypes of a number of individuals unambiguously. To determine the relationship between bill morphology and sex in Huia, we used molecular sexing methods and correlated these results with morphometric measurements.

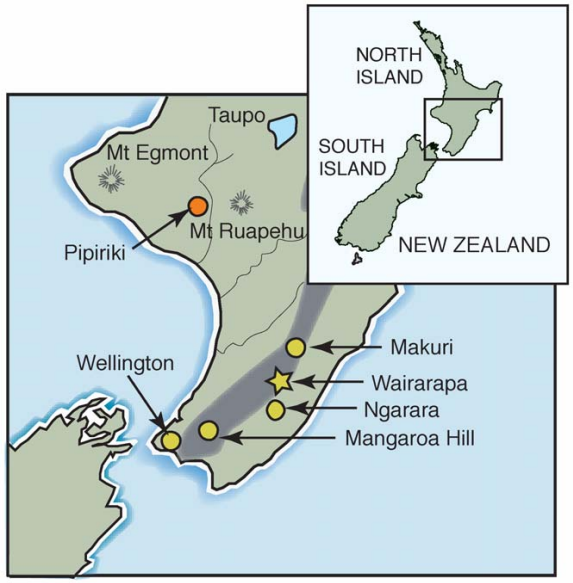


Figure 1. Provenance of Huia samples used in this study. Precise sample locations are indicated by circles and approximate locations by stars. Other place names are for reference only.
doi:10.1371/journal.pone.0008019.g001

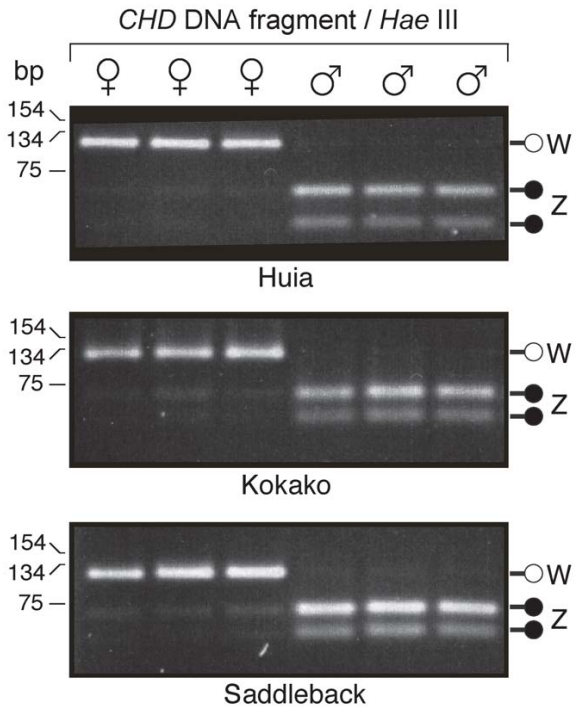


Figure 2. Huia sexing using semi-nested PCR with *CHD* primers p2/p5 followed by p2/p3. Amplification was followed by digestion of the Z *CHD* fragment with the restriction enzyme *Hae*III. The W chromosome *CHD* fragment in Huia is 121 bp long and the Z chromosome fragments are 65 and 56 bp in length, after enzyme digestion. Known-sex relatives of Huia, the North Island Kokako and Saddleback, were used to verify the test. Molecular weight maker sizes at shown at left.
doi:10.1371/journal.pone.0008019.g002

Results

We determined the molecular sex of 38 Huia specimens, as described in the Methods section. We show that Z and W chromosome sequences can be amplified in Huia and, along with digestion of these products with the restriction enzyme *Hae* III, allowed ZZ males to be distinguished from ZW females (Figure 2). Discrimination of the sexes was aided by the preferential amplification of the W *CHD* locus which was also recorded for Huia's close relatives, the Kokako and the Saddleback. We routinely find this with other avian sexing work (unpublished data), especially when using nested PCRs. This preferential W loci amplification fortuitously provided us with very clear sexing results – either W loci amplification only, or just Z loci amplification. Of the 38 specimens successfully sexed, 17 were shown to be males and 21 were females. Morphometric data show that Huia males had an average beak length of 52.8 mm, whereas female beaks were typically much longer, with an average of 77.9 mm. This difference was highly significant using a two-tailed t-test 22 df, assuming unequal variance ($P<0.0001$). In contrast, the mean female beak depth (13.9 mm) was significantly smaller than that of males (16.9 mm) using a two-tailed t-test 33 df and assuming unequal variance ($P<0.0001$). Variation in the beak length of males was relatively small (SD ±3.8 mm), while in females it was much larger (SD ±18.2 mm). To test the null hypothesis that there was no difference in the variance between male and female beak lengths, we conducted a one-tailed *F*-test. The test statistic is the ratio of two sample variances, and the difference was highly significant 20, 16 df ($P<0.0001$). In addition to beak length, females showed a significantly greater variance in beak depth (one-tailed *F*-test 20, 16 df; $P = 0.012$). Moreover, four DNA-sexed females were indistinguishable from males in terms of beak length (Figure 3; these individuals are indicated by black triangles). Another DNA-sexed female was intermediate between the males and shorter beaked female Huia (Figure 3), and was labeled as a "juvenile" in the museum collection (white triangle).

Genotyping

Eighteen of twenty-five Huia samples (72%) amplified for four or more of the six polymorphic microsatellite DNA loci used. All loci were in Hardy-Weinberg equilibrium (Table 1). No linkage was observed between loci following adjustment of the significance level for multiple comparisons with a Bonferroni correction [13]. Loci exhibited moderate levels of variation with 2–10 alleles per locus, and expected heterozygosities ranged from 0.437–0.766, with a mean of 0.637 across all loci (Table 1). The mean number of loci amplified per sample was 5.7 of the 6 loci. The mean shared allele distance between presumably non-related Huia was calculated to be 0.553 across all Huia individuals, except those in a putative family (AV2745, AV2746, and AV2747; Table 2).

Genotyping Error

Data for consensus genotypes are presented in Table 2. The ADO_{μ} (frequency of allelic dropout) of each locus varied from 3.2% to 20% (Table 1) with a mean of 12.9%. Seven replications were carried out for each locus so that the probability of obtaining false homozygotes was negligible ($P<0.0001$). Across all loci, the longer of the two alleles in a heterozygote was significantly more likely to not amplify (20 longer alleles verses 7 shorter alleles

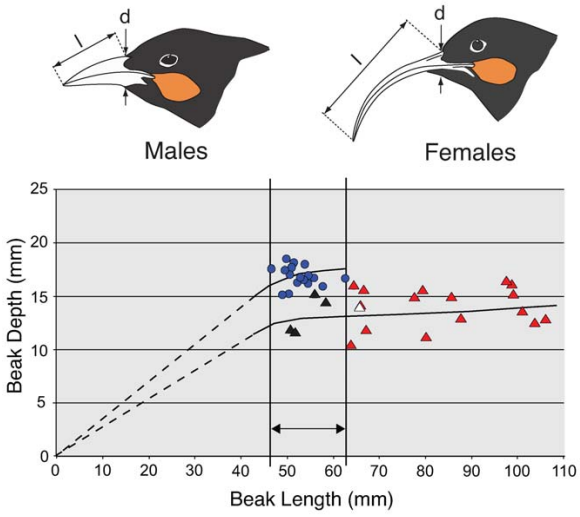


Figure 3. The relationships between beak length and depth for 17 male and 21 female Huia. Arrows indicate where measurements were made. Sex was determined using molecular methods. Blue circles represent males, red triangles indicate females, and black triangles represent the four DNA-sexed females that fall within the beak length range of males. Another DNA-sexed female, which was labelled 'juvenile', is indicated by a white triangle. Possible growth curves for males and females are shown by the black lines.
doi:10.1371/journal.pone.0008019.g003

'dropped out'; Chi squared test $\chi^2 = 6.24$, d.f. = 1, p = 0.01). Allelic dropout rates also differed significantly among samples (Chi squared test $\chi^2 = 36.11$, d.f. = 23, p = 0.040). Three false alleles were observed in the dataset (rate of occurrence = 0.7%).

Allelic dropout is thought to be a significant problem for genetic census data from degraded DNA samples such as those analysed here. This can result in false genotypes and hence can cause overestimation of population size [14]. In contrast, the effect of allelic dropout on estimations of population structure is probably less of a problem [15], although it has not been rigorously investigated [16]. The microsatellite genotyping error rates determined for the Huia dataset are within the range of those encountered in other studies of samples with low template concentrations [reviewed in 17]. Allelic dropout was a more common form of genotyping error in this dataset than the occurrence of false alleles. The probability of false homozygotes at each locus owing to allelic dropout was calculated to be negligible after seven replicates were averaged for each sample. However, this conclusion is based on the assumption that all individuals have equal dropout rates. This is probably violated in most datasets. Consequently, a few undetected dropouts may remain in the Huia dataset. However, the number is likely to be minimal because low quality samples were identified and removed prior to analysis. This approach has been found to be reliable in decreasing genotyping errors in other studies [14,18].

Estimating Genetic Diversity and Population Size of Huia

We recorded eight mitochondrial haplotypes among the 21 Huia individuals examined (Figure 4). The resulting dataset did not show extensive levels of artifactual mutations; samples 0.37386 and LB4568 only have two C>T singletons that could have arisen as a result of post mortem deamination of cytosine residues. If artifactual, these singletons would lead to a false overestimation of the extent of population expansion, and would consequently provide false census size estimates. However, these two samples were sequenced in the forward and reverse direction from independent amplifications and both sequence reads cover the singletons in question. Therefore, we are confident that they represent real variation. We used these data to obtain an estimate of genetic diversity (θ) of 0.011846 (95% credibility interval 0.005804–0.031121) from 199 bp of mitochondrial hypervariable region sequence. Mitochondrial data were used in this analysis because mutation rates for the hypervariable region are better known. The relationship between θ and the effective population size of breeding females is given by the expression $N_e(f) = \theta/2\mu$, where μ is the mutation rate per base pair per generation and generation time is defined as the average age at which a female reproduces. A generation time of 6.3 years was used in our analyses, and was based on the generation times of the related North Island Kokako [19] and Saddleback [20]. We used a mutation rate of 5.5×10^{-7} mutations per site per year for the hypervariable region, based on a pedigree study in Adélie penguins [21], as well as the phylogenetic estimate of 2.1×10^{-7} mutations per site per year. The former resulted in a mutation rate per generation of 3.47×10^{-6} for Huia and the latter 2.65×10^{-7}. Using the above values, $N_e(f)$ for Huia was estimated at 1709–4477. The range of estimates of the overall

Table 1. Genetic diversity measures and genotyping errors at six microsatellite DNA loci amplified from Huia.

Locus	Pca01	Pca05	Pca12	Pca13	Pca16	K9/K10	Overall
Number of alleles (N_A)	4	3	10	2	5	6	30
Allele size range (bp)	178–186	131–135	113–130	157–159	114–127	69–85	-
H_O	0.714	0.583	0.750	0.348	0.762	0.818	0.662
H_E	0.692	0.494	0.766	0.437	0.713	0.723	0.637
Hardy-Weinberg p values	0.201	0.967	0.777	0.896	0.610	0.649	-
Allelic dropout: Longer allele missing	6	0	6	2	2	4	20
Allelic dropout: Shorter allele missing	1	1	0	2	1	2	7
ADO_μ	0.179	0.032	0.133	0.200	0.088	0.146	0.129
FA_μ	0.024	0.019	0	0	0	0	0.007

The number of alleles (N_A), their size ranges, observed and expected heterozygosities and genotyping errors (allelic dropout rate: ADO_μ, and rate of occurrence of false alleles: FA_μ) are given.
doi:10.1371/journal.pone.0008019.t001

Table 2. Huia microsatellite DNA genotypes obtained following the 'multiple tubes' approach.

	Likely Genotype at Locus					
Sample Name	**Pca01**	**Pca05**	**Pca12**	**Pca13**	**Pca16**	**K9/K10**
AV2727	182/182 (7/7)	135/133 (2/2)	121/119 (2/2)	157/157 (7)	-	81/72 (2/3)
						81/81 (1/3)
AV37493A	184/182 (2/3)	135/133 (2/2)	121/119 (2/2)	157/157 (7/7)	121/119 (1/4)	72/72 (7/7)
	186/184/182 (1/3)				*119/119 (1/4)*	
					121/121 (2/4)	
AV1083	186/182 (2/3)	135/133 (2/2)	119/115 (2/3)	157/157 (7/7)	119/113 (2/2)	72/83 (1/3)
	186/186 (1/3)		*115/115 (1/3)*			*72/72 (1/3)*
						83/83 (1/3)
AV21283	Not consistently scorable	133/133 (7/7)	130/116 (2/2)	157/157 (7/7)	-	72/72 (7/7)
HBH	186/186 (7/7)	135/133 (3/3)	119/115 (3/3)	157/157 (7/7)	121/117 (2/2)	72/72 (7/7)
AV374934B	184/182 (2/2)	135/133 (2/2)	121/115 (2/2)	159/157 (2/2)	-	81/72 (2/2)
AV2745*	184/184 (7/8)	133/133 (8/9)	121/119 (3/3)	157/157 (7/7)	119/117 (2/2)	85/72 (2/2)
	186/184 (1/8)	**133/131 (1/9)**				
AV2244	-	133/131 (2/3)	119/115 (2/3)	-	117/117 (7/7)	85/72 (2/2)
		133/133 (1/3)	*115/115 (1/3)*			
AV2747*	184/182 (2/4)	135/133 (2/2)	121/119 (2/2)	159/157 (2/2)	119/119 (7/7)	85/72 (2/2)
	182/182 (2/4)					
AV2283	186/178 (2/2)	135/135 (7/7)	121/115 (2/2)	159/157 (2/4)	117/117 (7/7)	81/81 (7/7)
				157/157 (2/4)		
AV2744	184/182 (2/3)	135/133 (2/2)	122/115 (2/4)	159/157 (2/2)	121/119 (2/2)	85/72 (2/3)
	182/182 (1/3)		*115/115 (2/4)*			*72/72 (1/3)*
AV2746*	186/182 (2/3)	135/135 (7/7)	121/115 (2/2)	159/159 (7/7)	121/119 (2/2)	72/69 (2/2)
	182/182 (1/3)					
AV1082	184/178 (2/2)	135/133 (2/2)	129/121 (2/2)	157/157 (7/7)	117/113 (2/2)	83/72 (1/3)
						83/83 (1/3)
						72/72 (1/3)
AV2245	-	133/133 (8/8)	119/117 (3/3)	159/157 (2/4)	121/119 (2/2)	74/69 (2/3)
				159/159 (2/4)		*69/69 (1/3)*
AV1078	184/182 (2/2)	135/133 (2/2)	115/113 (2/2)	157/157 (7/7)	121/117 (2/2)	81/74 (2/2)
AV1085	186/184 (2/2)	133/133 (7/7)	130/115 (2/2)	159/157 (2/2)	121/119 (2/2)	-
AV21289	186/184 (2/2)	133/133 (7/7)	115/115 (7/7)	157/157 (7/7)	121/117 (2/2)	85/72 (2/2)
AV36838	184/182 (2/3)	133/133 (7/7)	126/121 (2/4)	157/157 (7/7)	127/119 (2/2)	81/72 (2/2)
	182/182 (1/3)		*121/121 (2/4)*			

False alleles are shown in bold text. False genotypes resulting from allelic dropout are shown in italic text. – indicates no amplification; * indicates three members of the same putative family.

doi:10.1371/journal.pone.0008019.t002

effective population size (N_e) of Huia using this generation time and these mutation rate variables was 3419–8954 breeding adults, assuming an equal sex ratio. The ratio of effective to census population size (N_c) in wild populations is often quite low and it has been suggested that a N_e: N_c ratio of 0.1 is appropriate [22]. This resulted in an expected census population size of 34,187 birds for the higher mutation rate. If we use the slower molecular rate of 2.1×10^{-7} [23], the mean estimate of the census population size is 89,538. Table 3 gives details of the credibility intervals for all these estimates. This range of population size estimates is best described as 'moderately high'. Estimates of growth rate (g) were consistently positive and large (~1111), with confidence intervals excluding zero (419–2776), indicative of an expanding population (http://evolution.gs.washington.edu/lamarc/).

Huia–Saddleback Comparisons

Huia exhibited a greater mean number of alleles per locus than did its relative, the North Island Saddleback, when polymorphic loci alone were considered (Table 4). Huia also had a higher mean number of alleles per locus when all loci that were polymorphic in either Huia or Saddleback (i.e., all loci in Table 4) were measured. The mean expected heterozygosity was higher in Huia (0.637) than in Saddleback (0.559, [24] and the length distribution of alleles at each locus was more continuous. For example, at locus Pca12, Huia possessed ten alleles differing in size by 1–4 bp, whereas only two alleles were detected in Saddleback, which differed by 16 bp.

Huia averaged a greater number of alleles per locus over all polymorphic loci than Saddleback, despite fewer Huia samples being genotyped. Moreover, Huia had higher levels of heterozy-

Sample Number	Variable Site(s)
	11
	7788933
	0167836
AV1076	A------
AV1082	----C--
AV1083	--A----
AV1078	----C--
AV1126	----C--
AV1079	--A----
AV1085	-------
AV2744	-------
AV2745	-------
AV2746	--A----
AV2747	--A----
LB4567	---TC--
LB4568	-------
LB4577	-------
0.37386	-T--CA-
RMNH110.109	-------
RMNH110.108	A------
AMNH669775	A------
AMNH669774	----C-C
AMNH669772	----CA-
HBH	----C--
Conserved Sequence	GCGCTGT

Figure 4. DNA nucleotide variation in 199 bp of the mitochondrial hypervariable region (HVRI) among 21 Huia.
doi:10.1371/journal.pone.0008019.g004

gosity than the Hen Island Saddleback population, from which all contemporary populations derive [24]. These results are unexpected because microsatellite loci are commonly longer and more variable in the species from which they are derived [25]. However, this is consistent with the fact that North Island Saddleback lost a considerable portion of their genetic variation through a population bottleneck [24].

Discussion

We suggest that the large variation in bill length and depth of female Huia, along with our molecular sexing data, indicate that female Huia may have began life with short stout bills indistinguishable from those of males and that their beaks grew longer over their lifetime. The primary evidence for this is that we recorded four female Huia that were DNA sexed, and these individuals had bills indistinguishable from males in terms of length. In addition a single Huia specimen labeled as a 'juvenile' was sexed as a female but had a very short bill. Several processes could explain the greater coefficient of variation in female bill length. For example, bill length might be controlled by W chromosome-specific loci and females consequently show much greater variation, or alternatively female bill growth may have been influenced by other factors such as diet. There is evidence for both these processes in avian species [26,27].

Our data show that Huia were characterised by a high level of genetic variation. Two genetic studies of Kokako reported low levels of genetic structuring among modern populations for both mtDNA [19] and microsatellite DNA loci [28]. However the level of genetic variation in Kokako, prior to their reduction in numbers, remains unknown. Although Kokako adults remain in the same territory for many years, it has been suggested that juveniles disperse early to find mates and/or establish territories [29], and that this is responsible for the recorded low levels of genetic structure. In contrast, early observations of Huia behaviour, such as apparent site fidelity and the presence of family groups, suggest that Huia would exhibit a high level of population genetic structure. To test this idea, we examined six polymorphic nuclear loci in Huia. Since many of the museum specimens examined in this study did not have detailed location data, a series of analytical methods were used to detect population differentiation independent of geographic information. The most likely number of populations in the Huia dataset of genotypes was estimated using a Bayesian clustering approach implemented in STRUCTURE [30]. This analysis suggested no evidence of population differentiation among the Huia samples examined (Table 5). Population structure was also investigated using PARTITION [31]. Using this algorithm, the tree plot of Huia genotype data also provided no evidence of population subdivision, as no well-supported clusters were separated by long branches. The

Table 4. The number of alleles (N_A) at microsatellite DNA loci isolated from North Island Saddleback that are polymorphic in North Island Saddleback and/or Huia.

Locus	N_A	
	Saddleback	**Huia**
Pca01	1	4
Pca02	4	1
Pca05	3	3
Pca08	3	1
Pca10	2	1
Pca12	2	10
Pca13	1	2
Pca14	3	1
Pca15	3	1
Pca16	1	5
Mean N_A/locus ± SE	2.30±0.33	3.00±0.91
Mean N_A/polymorphic locus ± SE	2.86±0.26	5.00*±1.15

*The mean N_A/polymorphic locus in Huia includes locus K9/K10 isolated from Kokako.
doi:10.1371/journal.pone.0008019.t004

Table 3. Long-term population size estimates of Huia based on mitochondrial hypervariable region diversity.

Mutation rate per base pair per generation (6.3 years)	Genetic diversity θ mean (95% credibility interval)	$N_e(f)$ (θ/2μ)	N_e ($N_e(f)$ ×2)	N census size (N_e:N_c ratio of 0.1)
3.47×10−6	**0.011846** (0.005804–0.031121)	**1709** (838–4491)	**3419** (1675–8982)	**34188** (16750–89815)

doi:10.1371/journal.pone.0008019.t003

Table 5. Estimated posterior probabilities, P(K/X), of K, the number of Huia populations.

K	lnP(X/K)	P(K/X)
1	−335.2	0.925
2	−338.1	0.051
3	−339.9	0.009
4	−339.4	0.014
5	−343.8	0.002

The estimated probability of the data, lnP(X/K), averaged over four independent runs for each K.
doi:10.1371/journal.pone.0008019.t005

plot of Bayesian probability levels versus generation time also indicated no evidence of population structure, i.e. the probability level declined gradually to a single cluster of individuals. Both analyses suggest a lack of population genetic structure in Huia, despite known provenance samples coming from locations up to 300 km apart. However, the area from which these samples came was covered in continuous lowland forest prior to human settlement, which may have allowed extensive mixing of resident populations.

Our estimate of the population size of Huia is relevant to the period prior to the human settlement of New Zealand and, depending on likely mutation rates and generation times, indicates a 'moderate' historical population size of 34,187 to 89,539 individuals. It is likely, however, that the pre-human population size was higher than our estimate, as our analyses show a period of population expansion. In combination, these factors suggest moderate to high Huia population densities. As a consequence, gene flow within the population is likely to be significant and will thereby promote genetic homogeneity within the species. In addition, populations of moderate size, such as that of Huia, typically exhibit a level of genetic inertia that may also have contributed to the overall genetic homogeneity of the species. This finding is in contrast to our original expectation that Huia populations would have exhibited high levels of population structure.

In conclusion, our molecular sexing results provide evidence that female Huia bills develop from a male-like condition. We suggest that young females may have been indistinguishable from males in terms of bill length, and that the typical differences developed over the lifetime of the individuals. Mitochondrial DNA variation in Huia allowed us to estimate the likely historical census population size for this species. Despite the suggestions from early naturalists that Huia were highly territorial and that the species had limited dispersal capability, no evidence for population genetic structure was detected using a range of nuclear loci. Generally, this work illustrates the potential of both sex-linked and autosomal DNA sequences to improve understanding of the population dynamics and molecular ecology of an extinct species. Furthermore, as many species of conservation concern become rarer in the wild, scientific programmes might benefit from using older specimens in museums. This would have less impact on modern populations of these species and would make use of a valuable museum resource. Hence, similar ancient DNA methods to those used here could be applied to scientific attempts to better understand the causes of species' decline.

Materials and Methods

Samples

Twenty-four Huia footpad samples were provided by the Canterbury Museum, 13 from the Auckland Museum, seven from the Naturalis Museum (Leiden), 24 from the American Museum of Natural History, two from Macleay Museum, two from the Australian Museum, and one sample from a private collector (Table 6). Exact provenance data were not known for 46 of these samples; the collection locations of the eight samples used for microsatellite analysis, plus two labelled as 'possibly Pipiriki', are illustrated in Figure 1.

Ancient DNA Methods

Extraction of ancient Huia DNA was performed in a dedicated ancient DNA laboratory physically separated from where contemporary DNA and PCR products were handled. Decontamination was routinely carried out by UV-irradiation and sodium hypochlorite washes. Approximately two mm^2 of Huia footpad tissue was removed and cut into several pieces using a sterile razor blade. Huia DNA was extracted by incubating footpad fragments overnight at 50°C in 2.5 ml of extraction buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 10 mM EDTA), 250 µl of 10% SDS, 15 µl of 200 mg/ml dithiothreitol (DTT), and 25 µl of 50 mg/ml Proteinase-K. Samples were then extracted with phenol followed by chloroform: isoamyl alcohol (24:1) and then concentrated by centrifugation through a Vivaspin-30 (Viva Science, U.K.) membrane. Negative extraction controls were included with every 6–12 sample extractions. All mitochondrial sequences were obtained by PCR as outlined below and sequenced in the forward and reverse direction from multiple independent amplifications. Huia HVRI sequences are deposited in the GenBank database with accession numbers GU176413–GU176433.

Huia Mitochondrial Hypervariable Region Sequences

One hundred and ninety-nine bp of the Huia hypervariable region I (HVRI) region were amplified from 21 individuals (Figure 4) as outlined below using primers huiaIF (5′-ATAAACCCAAGTGATCCTACCT) and huiaIIR (5′-TTGAGTAGCTCGGTTCTCGTGA). Amplification products were purified by centrifugation through Sephacryl™ S200HR (GE Healthcare), sequenced using ABI BigDye® Terminator v3.1 chemistry, and analysed using an ABI 3730 Genetic Analyzer. Sequences were edited and aligned in Sequencher™ 4.6 (Gene Codes Corporation).

Molecular Sexing of Huia and Morphometrics

The following primers were used to amplify Z and W chromosome sequences in Huia and thereby distinguish ZZ males from ZW females: p2, p3 [32], and a novel primer designed for Huia p5 (5′-GTAGGAGCAGAAGATATTCTG). These amplify a region of the sex-linked chromodomain-helicase-DNA gene (*CHD*) [32; Figure 2]. Amplifications were carried out in 10 µl reaction mixes containing 50 mM Tris-HCl pH 8.8, 20 mM $(NH_4)_2SO_4$, 1 mg/ml bovine serum albumin (BSA), 2.5 mM $MgCl_2$, 100 µM of each dNTP, 0.3 U of AmpliTaq® DNA Polymerase (Perkin Elmer), 40 ng of each primer, and approximately 1 ng of DNA. DNA was amplified using a Hybaid OmniGene thermal cycler. The initial amplification profile was 1×94°C 2 min, then 10×94°C for 20 sec, 55°C for 20 sec, and 72°C for 20 sec followed by 35×94°C for 20 sec, 52°C for 20 sec, and 72°C for 20 sec. A second PCR was carried out by adding approximately 0.3 µl from the initial amplification mix to a fresh

Table 6. Huia samples used in this study.

Museum Number	Location	Collector and Presentation Date	Morphological Sex	Molecular Sex (# times sexed)	Beak	
					L	D
AV1076	Wairarapa	Buller, 1891	Female	Female (2)	99.03	16.17
AV1078	Makuri	1892	Male	Male (1)	55	17
AV1079	Ngarara	Buller, 1891	Male	Male (1)	54.01	18.16
AV1081	-	-	Female	Female (1)	77.83	14.98
AV1082	Wairarapa	Buller, 1892	Male	Male (1)	50.09	18.61
AV1083	-	-	Female	Female (1)	51	11.9
AV1085	-	-	Male	Male (3)	50.48	17.3
AV1087	-	Moorhouse	Male	Male (3)	50.85	17.69
AV1126	-	-	Male	Male (2)	52.5	16.44
AV2244	-	Parker	Female	Female (1)	85.93	14.95
AV2245	-	Parker	Male	Male (1)	46.8	17.71
AV2283	-	Parker	Male	Male (1)	55.44	16.94
AV2727	-	Codmor	Female	Female (1)	67.35	11.9
AV2729	-	O'Connor	Male	Female (5)	56.16	15.27
AV2744	Wellington	-	Male	Male (3)	54.71	16.33
AV2745	Mangaroa Hill	Len Harris, 1885	Male	Male (1)	51.29	18.11
AV2746	Mangaroa Hill	Len Harris, 1885	Female	Female (3)	87.98	13
AV2747	Mangaroa Hill	Len Harris, 1885	Female (juvenile)	Female (2)	66.2	14.13
AV21283	-	P. Hall	Male	Male (1)	49.78	17.52
AV21289	-	-	Female	Female (2)	106.4	12.93
AV36838	-	F.Grimwood, 1870s Gifted by a North Island Maori Chief	Female	Female (3)	79.65	15.68
AV37493A	Possibly Pipiriki	Mrs F. Stewart	Male?	Female (2)	64	10.5
AV37493B	Possibly Pipiriki	Mrs F. Stewart	Female	Female (4)	104	12.5
HBH	-	-	Female	Female (2)	66	14
AV1070	-	-	Female	Female (2)	52	11.7
LB4564	North Island	-	Male	Male (3)	62.7	16.3
LB4565	North Island	-	Male	Female (1)	68.0	15.5
LB4567	Ruahine Range	C.E. Clarke, 20 Aug 1931	Male	Male (1)	58.2	15.9
LB4568	Ruahine Range	C.E. Clarke, 20 Aug 1931	Female	Female (1)	99.2	16.5
LB4571	North Island	S.H. Mountford, 1941	Male	Male (1)	54.3	16.5
LB4572	North Island	S.H. Mountford	Male	Male (1)	51.0	15.3
LB4573	North Island	S.H. Mountford	Female	Female (1)	64.2	15.8
LB4575	North Island	S.H. Mountford	Female	Female (1)	59.5	14.3
LB4576	North Island	C.A. Fleming	Female	Female (1)	80.3	11.7
LB4577	North Island	C.A. Fleming	Male	Male (1)	52.0	16.8
LB9213	-	J.A. Rentoul, 03 Dec 1969	Female	Female (1)	100.7	13.7
LB9215	-	-	Male	Male (1)	49.0	15.2
LB9217	-	-	Female	Female (3)	99.5	15.21
RMNH110.108	Rimutaka, Hills, Wellington	Travers, 1898	Male			
RMNH110.109	Rimutaka Hills, Wellington	Travers, 1898	Female			
AMNH669772	Wellington	1892	Male			
AMNH669774	Makuri Ranges		Male			
AMNH669775	Makuri Ranges		Male			

Museum numbers and presentation details if known and sex (morphological and molecular) are given. HBH = specimen obtained from Hastings Boys' High School; RMNH = Rijksmuseum van Natuurlijke Historie; AMNH = American Museum of Natural History, Beak length (L) and depth (D) are measured in millimetres.
doi:10.1371/journal.pone.0008019.t006

reaction mix. This mix was then cycled as follows: 1×94°C 2 min then 10×94°C for 20 sec, 55°C for 20 sec, and 72°C for 20 sec, followed by 25×94°C for 20 sec, 52°C for 20 sec, and 72°C for 20 sec. For restriction enzyme digestion, 1.2 µl of React 2 (Gibco-BRL), approximately 1 U of *Hae* III, and water was added directly to the PCR mix to a final volume of 12 µl. This mix was incubated at 37°C for 30 min before 3 µl of the digest was electrophoresed in 1.0%LE/1.5%MS agarose (Boehringer Mannheim) in TBE buffer. The gel was stained with ethidium bromide and the DNA was visualised over UV light. We also collected data on bill length and width using Vernier callipers accurate to 0.01 mm. Details of samples are provided in Table 6.

DNA products amplified with p2/p5 from samples AV2729, LB4565, AV1078, and LB4576 were cloned using a TOPO TA Cloning kit® (invitrogen) and several clones from each bird were sequenced. As expected when using these primers, only the *CHD* W allele could be isolated from the females. The *CHD* Z allele was isolated from male Huia AV1078 (Figure 5) in a similar way.

Nuclear Genotypes of Huia

Seventeen dinucleotide microsatellite DNA loci that had been isolated from a North Island Saddleback genomic library [24; T. King & D. Lambert, unpublished data] and eight loci isolated from a Kokako microsatellite library [28,33] were screened for polymorphism in Huia. PCR amplification was performed in 10 µl volumes containing 0.5 units of Taq polymerase (Roche), 200 µM of each dNTP, 0.8 pmol of each primer, 1.5 mM $MgCl_2$, 1x PCR buffer (50 mM Tris pH 8.8, 20 mM $(NH_4)_2SO_4$), and approximately 1 ng of extracted DNA. Samples were amplified by initial denaturation at 94°C for 4 minutes, followed by 35 cycles of 94°C for 45 seconds, then 50–55°C for 50 seconds followed by 72°C for one minute, with a final extension time of 72°C for 5 minutes. For each locus, PCR products amplified from seven Huia DNA samples were initially size fractionated in 1.0%LE/2.0%MS agarose to determine heterozygosity. Fluorescent dyes were used to label the reverse primer of each primer pair. Specifically Pca05, Pca12, Pca13, and Pca16 were labelled with 6-FAM; Pca01 with VIC; and K10 with HEX. Microsatellite DNA loci were amplified using the method described above and a negative control was included with every set of 6–12 reactions. Single-locus PCR reactions were pooled within samples where possible. Pooled reactions were genotyped using an ABI 3730 DNA Genetic Analyzer and visualised using Genescan. A sample standard was included with each genotyping run to account for between-run variation.

Genotyping Errors

Two techniques were used to avoid genotyping errors. Firstly, an initial screen of the quality of the samples was performed from the first round of PCR reactions [14]. Samples that amplified at fewer than three of the six loci were omitted from the study. For the remaining samples, the 'multiple tubes' method was used [34,35]. This approach involves multiple independent PCR amplifications of each locus to produce a consensus genotype. Putative homozygous genotypes were obtained seven times in order to discount allelic dropout (ADO_μ) and both alleles of putative heterozygotes were detected twice in order to rule out false alleles. In addition, a subset of four Huia samples were independently extracted and amplified for three loci (Pca01, Pca05, and Pca12) at an ancient DNA facility at the University of Auckland.

Analytical Methods

Measures of genetic variation, including observed and expected heterozygosities (H_O and H_E, respectively) and number of alleles, were calculated in Arlequin 2000 (http://cmpg.unibe.ch/software/arlequin/software/2.000/doc/faq/faqlist.htm). GENEPOP version 3.4 was used to test for linkage disequilibrium and deviation from Hardy-Weinberg parameters [36]. The sequential Bonferroni correction was applied to adjust the level of significance for multiple tests [13]. False alleles can occur in heterozygous and homozygous genotypes. Therefore, the rate of false allele formation is estimated across all positive amplifications. In contrast, allelic dropout can only be detected in heterozygotes; therefore, the allelic dropout rate is calculated using only the positive amplifications of heterozygotes. The rate of allelic dropout (ADO_μ) and the occurrence of false alleles (FA_μ) was calculated for each locus and across all loci using the equations below, as recommended by Broquet & Petit [17] (these calculations included the genotype data independently replicated at the University of Auckland).

Allelic dropout was calculated using the equation $ADO_\mu = D_j/A_{hetj}$, where D_j = the number of amplifications of locus j where an ADO event is observed, and A_{hetj} = the number of positive amplifications of heterozygotes. The number of false alleles (FA_μ) was determined using $FA_\mu = F_j/A_j$, where F_j = the number of amplifications at locus j where a false allele is observed, and A_j = the total number of amplifications (both hetero- and homozygotes).

The probability of false homozygotes at each locus after repeated PCR reactions (P) was calculated using the equation $P = (K) \times (K/2)^{n-1}$ [37] where K = the ADO_μ at each locus and n is the number of repeated amplifications; in this work, n = 7.

Consensus genotypes obtained from the multiple tubes method were used to examine population structure in Huia. The lack of provenance for the majority of samples prevented the application of traditional population genetic analyses such as F-statistics. Instead, two Bayesian clustering methods that do not require prior

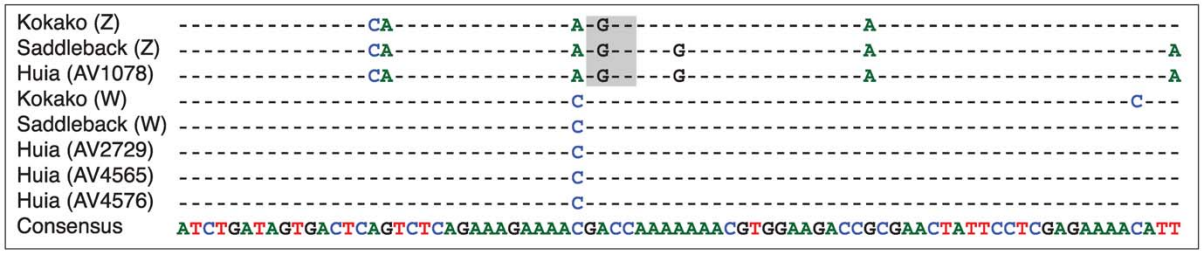


Figure 5. Sequence line-up of W and Z chromosome *CHD* sequences spanning the sex-specific *Hae*III restriction enzyme recognition sequence (GGCC) (grey box).
doi:10.1371/journal.pone.0008019.g005

population information to partition samples into genetic groups were used to detect any possible genetic structuring in the Huia microsatellite dataset: STRUCTURE 2.1 [30,38] and PARTITION [31]. Neither method requires the population of origin for individual samples, or even the number of sampled populations (K) to be known. Both methods identify clusters of individuals that are in Hardy-Weinberg and linkage equilibrium, but differ in their treatment of admixed individuals [39].

STRUCTURE was used with no input of prior population information in relation to individual samples, and admixture was assumed. Allele frequencies among clusters were considered to be independent to prevent overestimation of cluster number [38]. Four independent analyses of K = 1–5 were performed using 10^6 MCMC repetitions with the first 50,000 repetitions being discarded as 'burn-in' following visual confirmation that equilibrium had been reached. To select the optimal K, the posterior probabilities of the data, P(X/K) were calculated from the mean estimate log-likelihood of each K (lnP(X/K)).

PARTITION was applied only to the Huia samples that possessed a full complement of genotype data because missing data are not permitted in this software package. The parameter μ (the prior probability distribution on K) was set at 1, i.e. equal probabilities of each K were assumed, and the parameter θ (the prior distribution of alleles in the ancestral population) was varied from 1 to 20. The maximum number of source populations was changed with each analysis from 4 to 8. Estimates of the posterior probabilities were made after 50,000 observations of the Markov chain, with the first 5,000 observations omitted as 'burn-in'.

Shared allele distances (1minus half the average number of shared alleles per locus) between Huia were calculated online (http://www2.biology.ualberta.ca/jbrzusto/sharedst.php). The mean number of alleles (N_A) per locus and mean expected heterozyosity (H_E) were compared between Huia and Saddleback. Data from 41 individuals from the Hen Island population of Saddleback [24] were used in these comparisons as all contemporary Saddleback populations derive from this island.

Estimating Genetic Diversity and Effective Population Size in Huia

The genetic diversity parameter (θ) was calculated using a Bayesian framework [LAMARC v2.1.2; 40] that uses a coalescent approach to obtain a joint estimate of various population genetic parameters such as genetic diversity, growth, migration, and recombination rates. We estimated θ from 21 individuals for a 199 bp fragment of the mitochondrial hypervariable region. To ensure that the Bayesian estimate of θ was robust, we performed a number of repeat analyses. Fourteen preliminary analyses were conducted with a range of starting parameters (e.g. sample size and sampling increment). Posterior probability distributions were compared between runs to assess whether we converged on an estimate of θ. We also assessed convergence by calculating the effective sample size (ESS) (using the program Tracer v1.4, (http://beast.bio.ed.ac.uk/Tracer). An ESS of 100–200 has been suggested to indicate convergence. Our estimates were well above this value, in the thousands or greater. After preliminary analyses were completed, a final estimate of θ was performed from 10 replicates, each with the following starting parameters: $\theta = 0.015$, and a linear prior of 0.0001–3. Two initial chains sampled 5000 trees with a sampling increment of 40, of which the first 7000 trees sampled were discarded, followed by 4 final chains to produce the estimate, in which, after a burn-in of 5000 trees, every 50^{th} tree of 5×10^6 trees was sampled. Adaptive chain heating was used (chain temperatures 1, 1.1, 1.2, and 1.3). We also performed separate LAMARC analyses to test for signatures of exponential growth or shrinkage using the same searching strategy.

Acknowledgments

The authors thank Te Papa (Alan Tennyson); Canterbury Museum (Geoff Tunnicliffe and Paul Scofield); Te Manawa Museum (Pamela Lovis); the American Museum of Natural History (Paul Sweet and Joel Cracraft); Naturalis, National Museum of Natural History, Leiden (Caroline Pepermans and Hein van Grouw); Macleay Museum, the University of Sydney (Margaret Humphreys); Australian Museum (Les Cristidis and Walter Boles); Natural History Museum, Vienna (Ernst Bauernfeind and Anita Gamauf) and Rhys Cullen for permission to sample Huia skins. Thanks, too, to the numerous individuals from various institutions who responded to our enquiries about huia material. We also thank Dee Denver, Ed Minot, Hugh Paterson, and Leon Perrie for comments on the manuscript and Vivian Ward for graphics.

Author Contributions

Conceived and designed the experiments: DML CDM. Performed the experiments: LS LJH GBP. Analyzed the data: DML LS LJH GBP. Contributed reagents/materials/analysis tools: GW. Wrote the paper: DML CDM.

References

1. Selander RK (1966) Sexual dimorphism and differential niche utilization in birds. Condor 68: 113–151.
2. Selander RK (1972) Sexual selection and dimorphism in birds. In: Campbell B, ed. Sexual Selection and the Descent of Man 1897–1971. Chicago: Aldine. pp 180–230.
3. Burton PJK (1974) Anatomy of head and neck in the huia (*Heteralocha acutirostris*) with comparative notes on other Callaeidae. Bulletin British Museum Natural History (Zoology) 27: 1–48.
4. Gould JA (1837) Synopsis of the birds of Australia and adjacent Islands. London: Gould.
5. Lack D (1971) Ecological Isolation in Birds. Oxford: Blackwell Scientific.
6. Buller WL (1888) A History of the Birds of New Zealand. London: Van Vorst.
7. Heather BD, Robertson HA (1998) The Field Guide to the Birds New Zealand. Oxford: Oxford University Press.
8. Worthy TH, Holdaway RN (2002) The Lost World of the Moa. Christchurch: Canterbury University Press.
9. Phillips WJ (1963) The Book of the Huia. Christchurch: Whitcombe and Tombs.
10. Colenso W (1887) A description of the curiously deformed bill of a huia (*Heteralocha acutirostris*, Gould) an endemic New Zealand bird. Transactions and Proceedings of the New Zealand Institute 19: 140–145.
11. Potts T (1885) Oology of New Zealand. New Zealand Journal of Science 2: 373–484.
12. Moorhouse RJ (1996) The extraordinary bill dimorphism of the huia (*Heteralocha acutirostris*): sexual selection or intersexual competition. Notornis 43: 19–34.
13. Rice WR (1989) Analyzing tables of statistical tests. Evolution 43: 223–225.
14. Paetkau D (2003) An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12: 1375–1387.
15. Creel S, Spong G, Sands JL, Rotella J, Zeigle J, et al. (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12: 2003–2009.
16. Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology and Evolution 20: 136–142.
17. Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology 13: 3601–3608.
18. Hung C-M, Li S-H, Lee LL (2004) Faecal DNA typing to determine the abundance and spatial organisation of otters (*Lutra lutra*) along two stream systems in Kinmen. Animal Conservation 7: 301–311.
19. Double M, Murphy S (2000) Genetic variation within and among populations of North Island kokako. Science and Research Internal Report 176. Wellington: Department of Conservation.
20. Armstrong DP, Davidson RS, Perrott JK, Roygard J, Buchanan L (2005) Density-dependent population growth in a reintroduced population of North Island saddlebacks. Journal of Animal Ecology 74: 160–170.

21. Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, et al. (2008) Mutation and evolutionary rates in Adélie penguins from the Antarctic. Public Library of Science Genetics 4: 1–9.
22. Frankham R (1995) Effective population size/adult population size ratios in wildlife: a review. Genetical Research 66: 95–107.
23. Wenink PW, Baker AJ, Tilanus MG (1994) Mitochondrial control-region sequences in two shorebird species, the turnstone and the dunlin, and their utility in population genetic studies. Molecular Biology and Evolution 11: 22–31.
24. Lambert DM, King T, Shepherd LD, Livingston A, Anderson S, et al. (2005) Serial population bottlenecks and genetic variation: translocated populations of the New Zealand saddleback (*Philesturnus carunculatus rufusater*). Conservation Genetics 6: 1–14.
25. Ellegren H, Primmer CR, Sheldon BC (1995) Microsatellite 'evolution': directionality or bias. Nature Genetics 11: 360–362.
26. Grosler AG (1990) The variable niche hypothesis revisited: an analysis of intra- and inter-specific differences in bill variation in *Parus*. In: Blondel J, Gosler A, Lebreton JD, McCleery R, eds. Population Biology of Passerine Birds: an Integrated Approach. Berlin: Springer. pp 167–174.
27. Grosler AG, Carruthers TD (1994) Bill size and niche breadth in the Irish coal tit *Parus ater hibernicus*. Journal of Avian Biology 25: 171–177.
28. Hudson QJ, Wilkins RJ, Waas JR, Hogg ID (2000) Low genetic variability in small populations of New Zealand kokako *Callaeas cinerea* wilsoni. Biological Conservation 96: 105–112.
29. Innes J, Flux I (1999) North Island recovery plan 1999-2009. Threatened Species Recovery Plan 30. Wellington: Department of Conservation.
30. Pritchard JK, Stephens M, Donnelly P (2000) Inferences of population structure using multilocus genotype data. Genetics 155: 945–959.
31. Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetical Research 78: 59–77.
32. Griffiths R, Double MC, Orr K, Dawson JG (1998) A DNA test to sex most birds. Molecular Ecology 7: 1071–1076.
33. Hudson QJ (1999) Genetic variation within and among populations of kokako (*Callaeas cinerea*). MSc thesis. University of Waikato, Department of Biological Sciences.
34. Navidi W, Arnheim N, Waterman MS (1992) A multiple-tube approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. American Journal of Human Genetics 50: 347–359.
35. Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, et al. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24: 3189–3194.
36. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248–249.
37. Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6: 861–868.
38. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
39. Pearse DE, Crandall KA (2004) Beyond Fst: analysis of population genetic data for conservation. Conservation Genetics 5: 585–602.
40. Kuhner MK (2006) LAMARC (version 2.0): maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768–770.

II APPENDIX TWO

SUPPLEMENTARY MATERIAL CHAPTER FIVE
INDIVIDUAL INTRON PHYLOGENETIC TREES

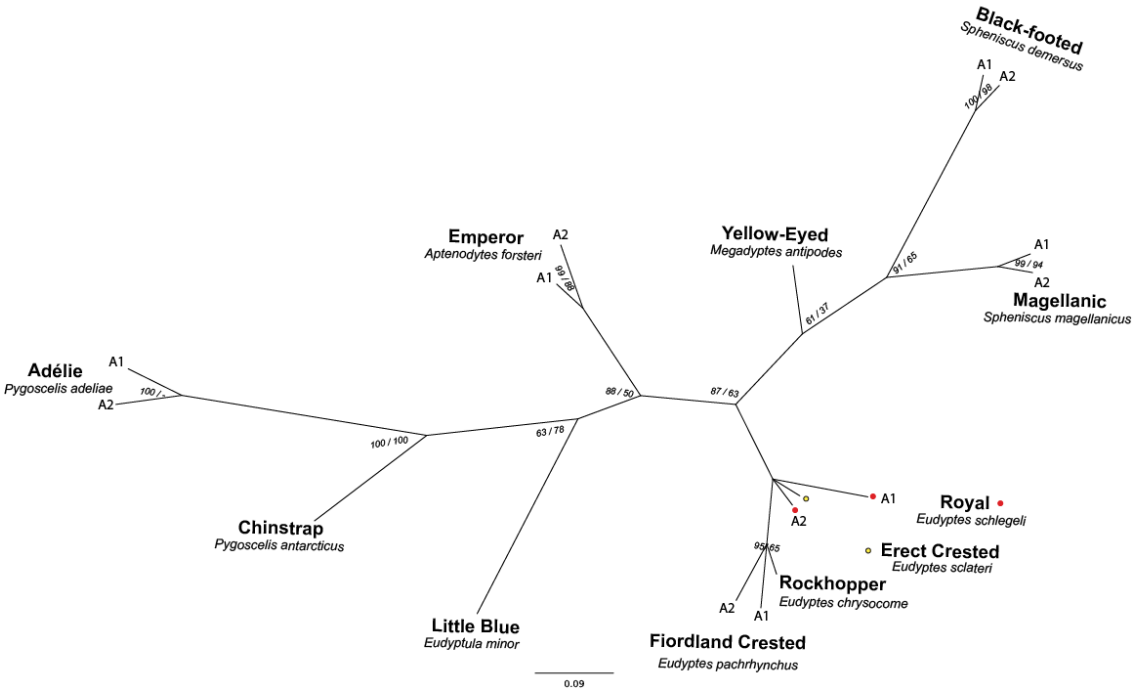


Figure II.1 Unrooted Bayesian modern penguin phylogenetic consensus tree for locus *UCHL3*. Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. The scale indicates a branch length equivalent to 0.09 substitutions per site.

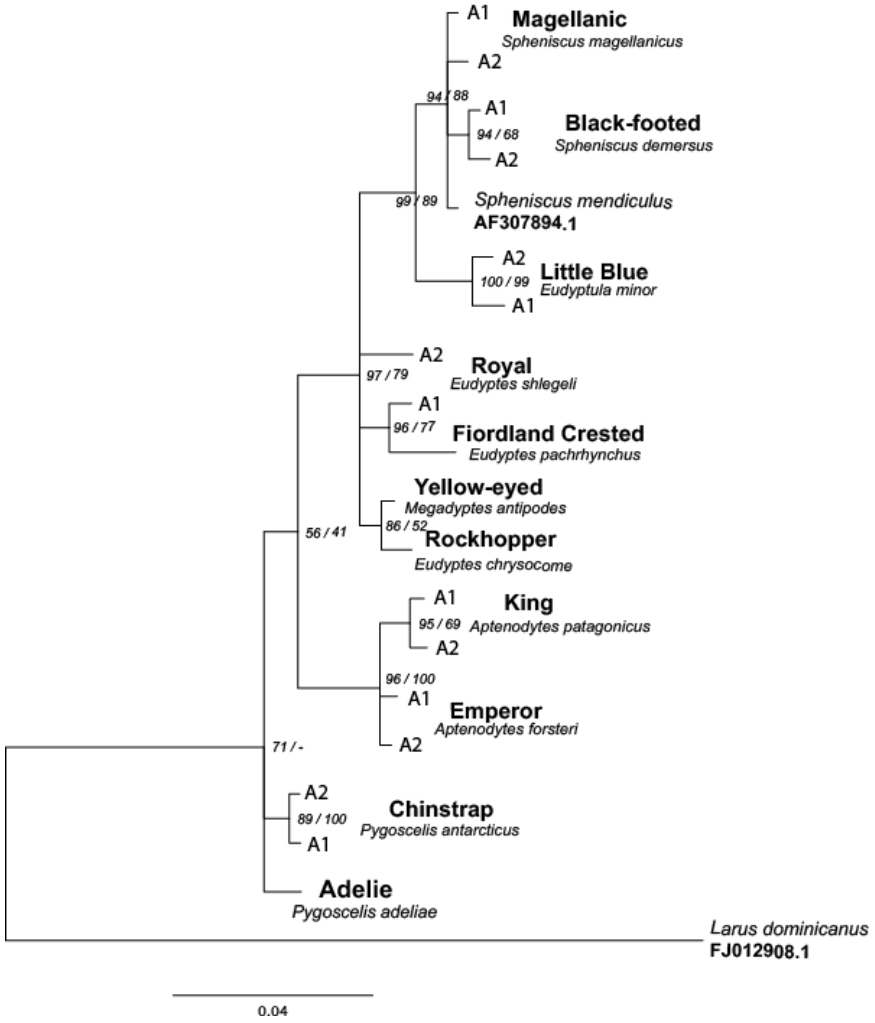


Figure II.2: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *AK1i5*. Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. The scale indicates a branch length equivalent to 0.04 substitutions per site. NCBI accession number for the outgroup is shown.

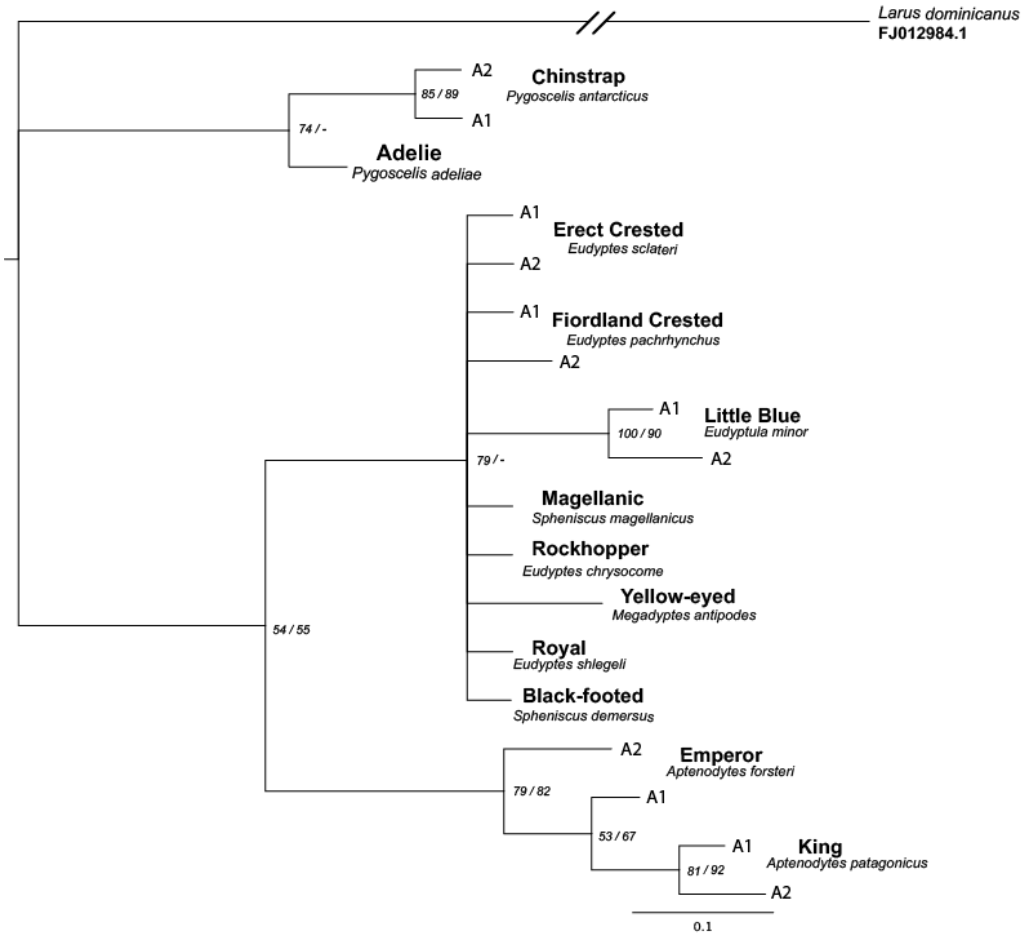


Figure II.3: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *MPP4*. Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. The scale indicates a branch length equivalent to 0.01 substitutions per site. NCBI accession number for the outgroup is shown.

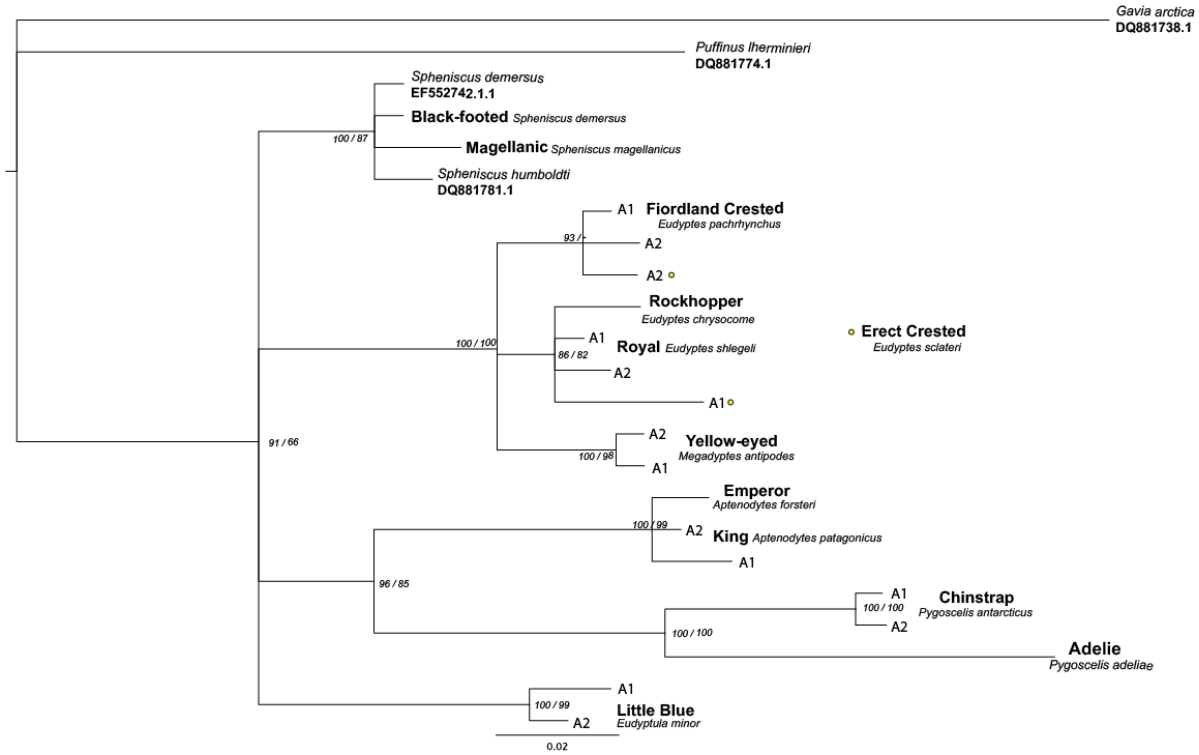


Figure II.4: Rooted Bayesian modern penguin phylogenetic consensus tree for locus *MPP4*. Bayesian/maximum likelihood support (Bayesian posterior probabilities and bootstrap scores) is shown for each node. The scale indicates a branch length equivalent to 0.02 substitutions per site. NCBI accession numbers for the outgroups are shown.

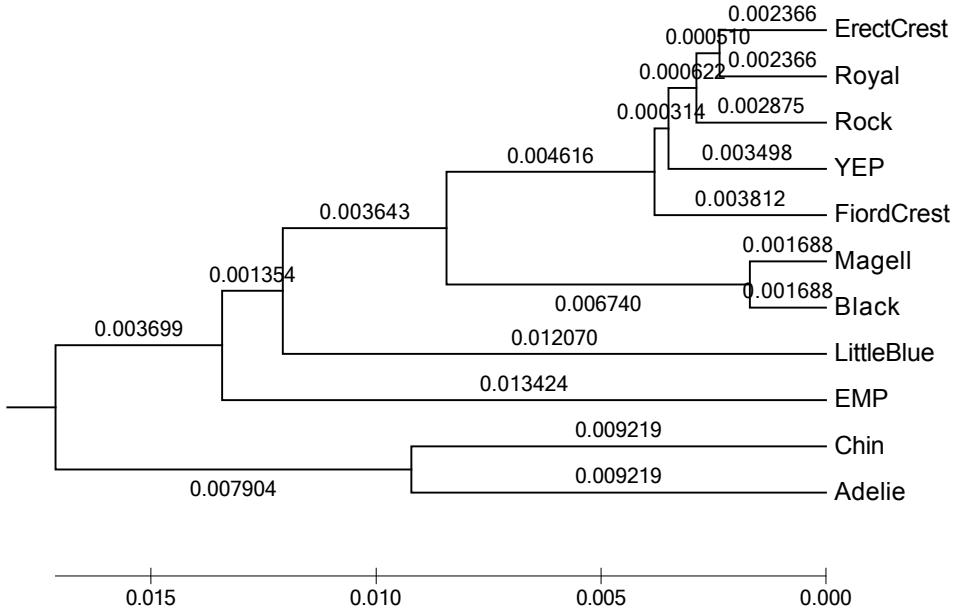


Figure II.5: UPGMA phylogenetic tree for the four concatenated intron-only dataset.

Tree generated in MEGA 5, and used for an estimate of the intron evolutionary rate for Chapter 5.

III Appendix Three

SUPPLEMENTARY MATERIAL FOR CHAPTER SIX

EXTENDED METHODS

1.1 Direct Sequencing of Ancient Myelin Proteolipid Protein Intron Four

Prior to applying the FLX protocol, primers were designed and tested to obtain locus *MPP4* directly through Sanger sequencing and as a way to gauge size ranges for amplifiable nuclear intron sequence in ancient Adélie penguin subfossil bone extracts. Internal primers were designed for *MPP4* following the same primer design methodology as described in Chapter Six, except that fragment sizes ranged from between 124 and 139bp including primers, amplifying the intron in four fragments, designed to be carried out as two multiplex PCR reactions followed by four singleplex reactions. Primer pairs were tested on ancient Adélie penguin DNA samples using similar optimization strategies as described in Chapter Four. Ancient samples used for this purpose were a subset of the 30 used in Chapter Seven. All extractions followed the protocol described in Chapter Six. Purification of PCR products was carried out as described in previous chapters, unless non-specific bands were present, and then the correct product was excised from a 2% agarose gel and purified using (ZYMO).

Table III.1 Internal primers for *MPP4*

Locus	Primer Name	Sequence (5'-3')	Location
MPP4	MPPFint1	GGACGCCAGGATGTACGG	Bf2
	MPPintR1	CTGGACAGCGAACCGGGTACT	Br4
	MPPintR2	CATTTTAGGGCTCTTCAGCACTTG	Br3
	MPPFint2	CTGCAGGAGGCTGAGCAAG	Bf4
	MPPFint3	GGATAAGGCTGGGATAAAGCA	Bf3
	MPPintR4	ATGCGGAGAGCCAGGTGAG	Br2

Location refers to the location of each primer within the intron, as shown on Fig. III.1.

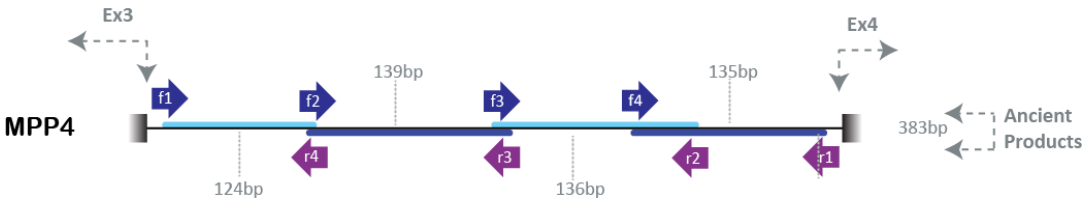


Figure III.1: Schematic representation of positions of *MPP4* primers

Primers designed within *MPP4* and product sizes. Primer labels refer to those in Table C.1, except for Bf1 (22F, Table 4.1, Chapter 4) and Br1 (22R, Table 4.1, Chapter 4), which were obtained from the literature as described in Chapter 4.

The different fragments of *MPP* showed varying success from sequencing in ancient Adélie samples (Fig. C.2). The 5' most fragment (f1-r4 in Fig. C.1, 124bp including primers) sequenced in all the samples screened, except 14.2 and 10.2, which produced unreadable sequence. The other fragments, 135-139bp in length, did not always sequence well, particularly in older samples. Following these results, second stage primer design aimed to amplify fragments under 130bp. As this was an initial screen to gauge nuclear fragment sizes in Adélie ancient DNA, sequencing efforts were not fully comprehensive for the samples tested. As a result, though four different changes were observed in four of the sequences, these were not verified by further sequencing and could be the result of error or DNA damage.

Primer design within *MPP* was successful and intron fragments were obtained from ancient samples tested ranging in age from <500 to >6000 years old. It was observed, however, that fragments over 130bp in length did not amplify as readily for older samples. As a result of this observation, primers were redesigned to cover all of *AK1i5*, *MPP4*, and *UCHL3* in overlapping multiplex sets of fragments sized between 100-130bp.

Sequencing of the 9 ancient samples for *MPP* showed a few changes in sequence, none of which were verified (all stemmed from 1x coverage). In order to verify whether these changes were the result of error, miscoding lesions or genuine changes, further cloning and sequencing would need to be carried out. Obtaining enough ancient sequence data for a sample is quite costly in terms of DNA, especially when wishing to obtain data for more than one marker. A multiplex approach helps to diminish this cost, but identifying variants and verifying them requires time-consuming cloning and sequencing for each fragment. Next-generation sequencing

technologies, however, offer a way of obtaining enough sequence data to address issues of sequence error, miscoding lesions and variation detection while utilizing low amounts of precious ancient DNA extract. This approach is described in Chapter 6.

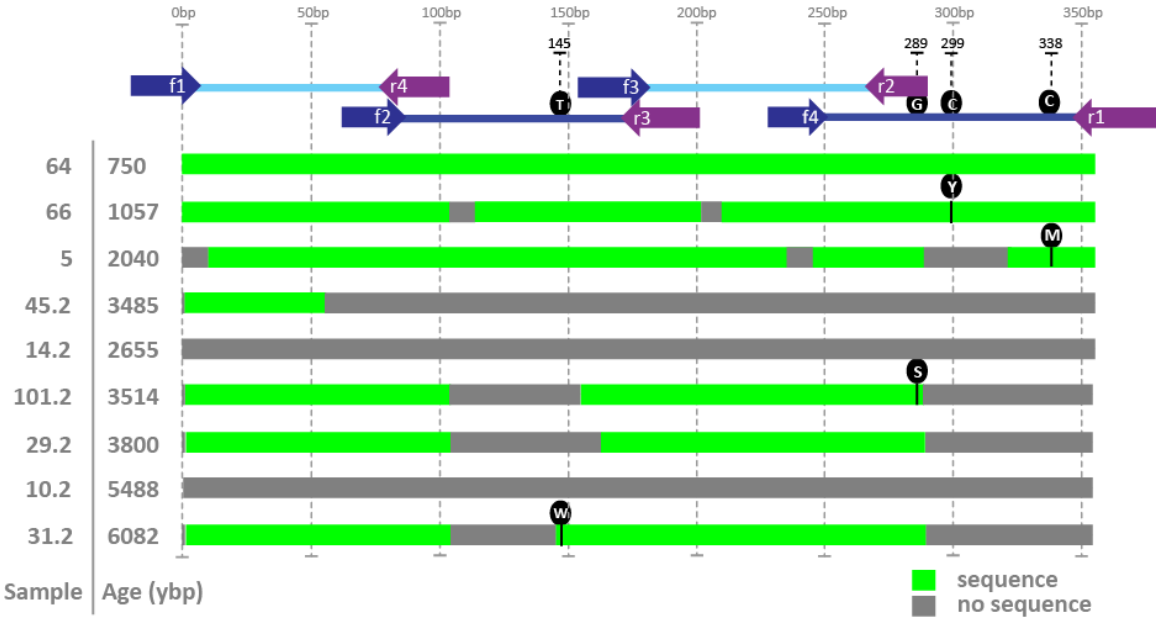


Figure III.2: *MPP4* direct sequencing results in four fragments from ancient Adélie subfossil bone samples.

Primers are labeled according to the key in Fig. III.1/Table III.1. Position along the consensus sequence of *MPP* trimmed of F/R primers is marked. Base changes observed are marked on the sample of origin and the consensus bases are marked on the primer map. All lengths are proportional to the actual lengths.

III.4 FLX Tagging Protocol and Library Quantification

Sixty-five 8bp barcoded adapters were designed by taking 8bp sequences, each with at least 3 differences between them, and adding them to a truncated FLX Titanium Shotgun 'A' adapter (see figure III.3) to give a 26bp oligo. The 8bp tags were designed by Matthias Meyer and used with his permission. Adapters were ordered with phosphothiorate bonds between the first and last four nucleotides as it has been observed to improve ligation efficiency (M. Knapp, *pers. comm.*). 12bp complementary oligos (which include the reverse complement of the 8bp tag plus TCAG segment of the adapters) to each of the 26bp barcoded adapters were also designed and ordered, along with a truncated FLX Titanium 'B' adapter (17bp) and a

12bp reverse complement of the 3' end. Oligos were obtained from Sigma-Aldrich and purified by desalting. Two full FLX Titanium A and B adapters, HPLC purified without PTO bonds, were ordered as well to amplify the libraries after ligation.

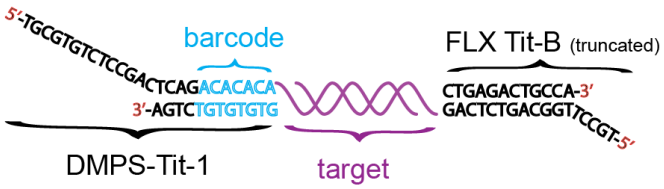


Figure III.3: Example of a tagged target sequence prior to the adapter fill-in step.

DMPS-Tit-1 is a truncated barcoded FLX Titanium A adapter which is first hybridized to a shorter reverse complement of itself (DMPS-Tit-1-comp) before ligation to the target. The FLX Tit-B truncated adapter is not barcoded and is also hybridized to a reverse complement of itself prior to ligation.

In the first place multiplex PCR products originating from the same sample and replicate were pooled and concentrated to 15μl in 0.1x TE Buffer using a Qiagen Minelute purification kit following the manufacturer's protocol. The adapters were dissolved in 1xTE buffer to a final concentration of 500μM then prepared by hybridizing tag-A and rev-comp-tag-A together with T4 Ligase Buffer (1x) (Fermentas) or hybridization buffer (1x of the 10x which has 5ml 5M NaCl and 45ml of 1xTE) and incubated in a thermal cycler to a concentration of 200μM. Ready to use adapters were diluted to 100μM in 1x hybridization buffer and stored at -20°C. Following this a blunt-end repair reaction was carried out for the multiplex products with dNTPs (100μM each, Invitrogen), Buffer Tango (1x), ATP (1mM), T4 Polynucleotide Kinase (0.5 U/μl) and T4 Polymerase (0.1 U/μl, all Fermentas). The mix was added to the 15μl of multiplex product and incubated for 15 minutes at 12°C then 15 minutes at 25°C on a thermocycler. The products were purified as described previously and eluted in 15μl 0.1xTE. Adapters were blunt-ligated to the products from replicate one in a reaction containing 1x T4 Ligase Buffer, 5% PEG-4000, and 0.125 U/μl T4 Ligase (all Fermentas), as well as 0.5μl of adapter A and B (see Figure III.4). Products from replicate two were blunt-ligated to the adapters in a reaction containing 1x T4 Ligase Buffer, 0.125U/μl T4 Ligase (both NEB) and 5% PEG-4000 (Fermentas). Reactions were incubated for 20 minutes at 22°C. The reactions were once again purified using SPRI beads (Agencourt) and eluted in 15μl 0.1xTE.

Finally, an adapter fill-in reaction was performed with 1x Thermopol buffer (New England Biolabs), 250μM dNTPs (Invitrogen), 0.27 U/μl Bst Polymerase (New England Biolabs), and then purified by Ethanol/SPRI (Agencourt) beads and eluted in 20μl 0.1xTE. 1μl of each barcoded sample was amplified using the make-454 TIT-A and B HPLC purified primers to finalize the library preparation process. Amplification reactions were carried out in 50μl containing 0.2mM dNTPs (Invitrogen), 2.5mM MgCl2, 0.5units of Amplitaq Gold, 1x reaction buffer (all Applied Biosystems), and 200nM of each primer. Cycling conditions were as follows: initial enzyme activation stage of 10 minutes at 95°C, followed by 9 cycles of 95°C for 30s, 60°C for 30s, 72°C for 30s, and a final extension of 72°C for 10 minutes.

Firstly a quantification standard was created from modern Adélie penguin DNA samples by amplifying a 124bp fragment in a 25μl reaction containing 0.2mM dNTPs, 5μl of diluted Adélie modern DNA extract, 2.5μM $MgCl_2$, 0.5u of Amplitaq Gold DNA polymerase (Applied Biosystems) and 1x of its associated reaction buffer. The primers used were one pair of the set of nuclear primers used for the multiplex protocol (MPPF and MPPRint1), at a concentration of 0.4μM. The reaction was carried out on an iCycler thermocylcer (Thermofisher) by an initial denaturation at 95°C for 9 minutes, followed by 30 cycles of 30 seconds denaturation (95°C), annealing (58°C) and extension (72°C), culminating in a final extension of 7 minutes at 72°C. The PCR products generated were pooled and purified through a Qiagen Minelute purification column following the manufacturer's protocol. The eluted product was quantified on a NanoDrop 2000c spectrophotometer, and approximately 300ng of the product was used for the downstream ligation protocol, done in duplicate for the Fermentas and NEB T4 ligase reactions to compare ligation efficiency of the different reagents. DMPS-Tit-1 was used as the A adapter for the ligations. Ligations were performed as described above with one modification; the adapter concentration was 200μM instead of 100μM. Once ligations were performed, 8μl of the standard library was visualized on a 2.5% agarose gel to verify the ligation efficiency of the two reagent sets used. No difference in ligation success or yield was noted between the Fermentas and NEB ligases.

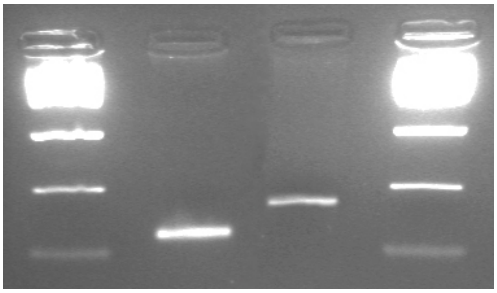


Figure III.4: 3% agarose gel showing the result of a ligation test of one adapter.

A low-mass ladder was used as size marker. 4μl of ladder was loaded. The first lane after the ladder shows the DNA control prior to ligation protocol (5μl loaded), and the second lane shows the sample control after all ligation and purification steps (10μl loaded). The size shift from the ligation of A and B adapters is clear.

1μl of each of the two standards were amplified in 8 separate reactions using the make-454 TIT-A and B HPLC purified primers to finalize the library preparation process. Amplification reactions were carried out in 50μl containing 0.2mM dNTPs (Invitrogen), 2.5mM MgCl2, 0.5units of Amplitaq Gold, 1x reaction buffer (all Applied Biosystems), and 200nM of each primer. Cycling conditions were as follows: initial enzyme activation stage of 10 minutes at 95°C, followed by 9 cycles of 95°C for 30s, 60°C for 30s, 72°C for 30s, and a final extension of 72°C for 10 minutes. Following this, standards were verified visually on a 2.5% agarose gel, pooled and purified using Qiagen Minelute columns according to the manufacturer's protocol, and eluted in 30μl 0.1xTE buffer. This pooled standard was quantified using a NanoDrop 2000c spectrophotometer, yielding 34 ng/μl of DNA, which equates to 1.5 x 10^{11} copies of 200bp fragments of DNA. A dilution series of the standard was created ranging from 1.5 x 10^{9} copies/μl to 1.5 x 10^{2} copies/μl.

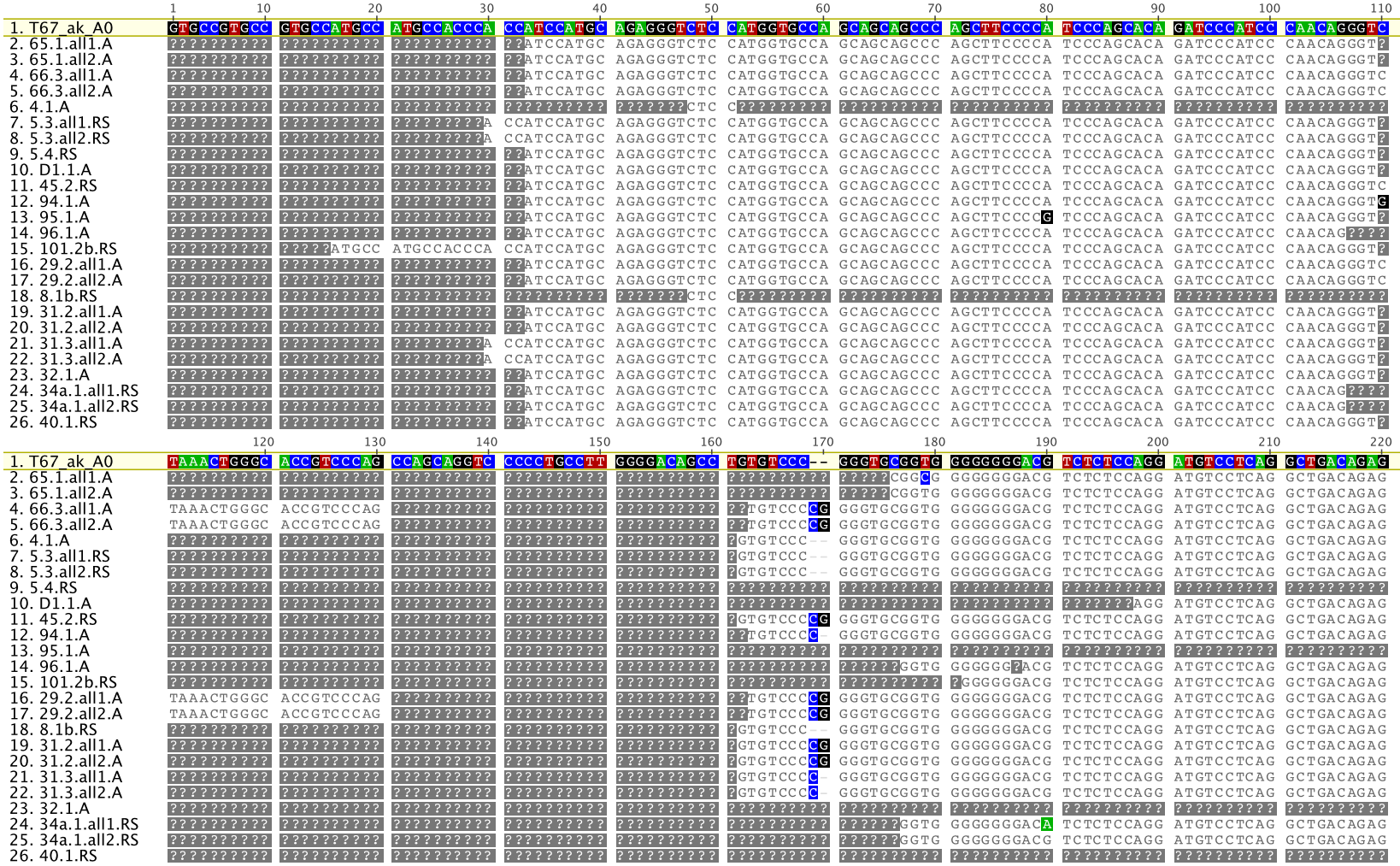
A quantitative PCR was set up on a Roche 480 lightcycler with 1μl of undiluted amplified libraries, negatives (extraction, library amplification and QPCR blanks), and standards. The standards were set up in 2 replicates, at the concentrations mentioned above. QPCRs were carried out in 25μl reactions with 1x Absolute QPCR SYBR Green Low ROX mix, 200nM of each primer (make-454-TitA and B) and 1μl of template. An initial activation step of 15 minutes at 95°C was followed by 45 cycles of denaturation at 95°C for 30 seconds, annealing at 60°C for 30 seconds, and extension at 72°C for 45 seconds. Results from the run showed that the concentrations of the libraries were higher than the range covered by the standard, so

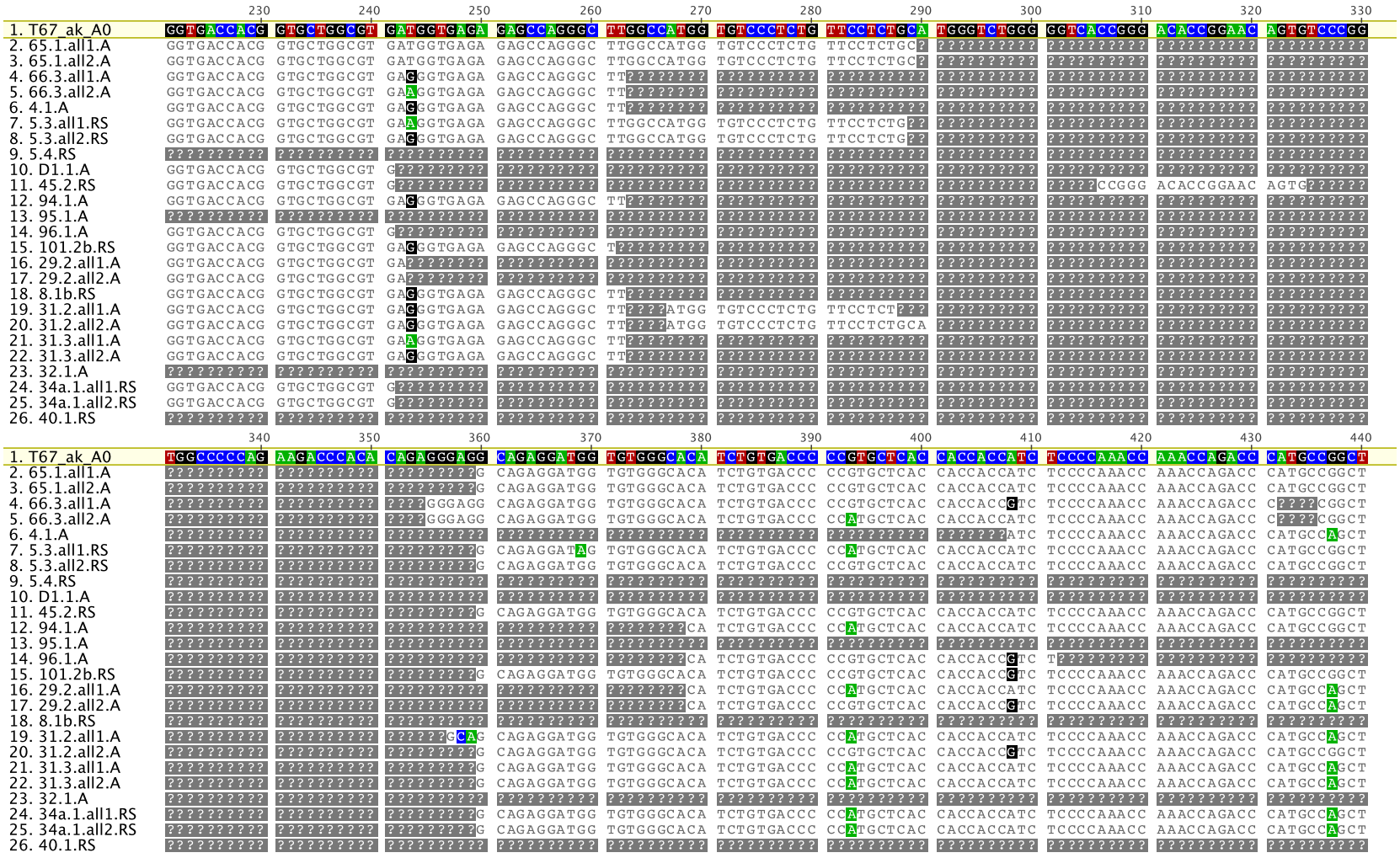
libraries were diluted 1000 fold and a second QPCR was performed with the same conditions as above.

Table III.2: Barcoded Adapters used for DMPS FLX Titanium Sequencing

Barcoded Adapters and Oligos for Intron DMPS FLX Titanuim Sequencing			
Truncated Titanium A barcoded shotgun adapters and B adapter		**Complementary oligos (rc of TCAG-tag)**	
Name	**Sequence (5'-3')**	**Name**	**Sequence (5'-3')**
DMPS-Tit-1	TGCGTGTCTCCGACTCAGACACACAC	DMPS-Tit-1-comp	GTGTGTGTCTGA
DMPS-Tit-3	TGCGTGTCTCCGACTCAGTCTCTCTC	DMPS-Tit-3-comp	GAGAGAGACTGA
DMPS-Tit-5	TGCGTGTCTCCGACTCAGACGCGCGC	DMPS-Tit-5-comp	GCGCGCGTCTGA
DMPS-Tit-7	TGCGTGTCTCCGACTCAGTGCGCGCG	DMPS-Tit-7-comp	CGCGCGCACTGA
DMPS-Tit-9	TGCGTGTCTCCGACTCAGAGAGACAC	DMPS-Tit-9-comp	GTGTCTCTCTGA
DMPS-Tit-11	TGCGTGTCTCCGACTCAGATATATAT	DMPS-Tit-11-comp	ATATATATCTGA
DMPS-Tit-13	TGCGTGTCTCCGACTCAGTGTGTATA	DMPS-Tit-13-comp	TATACACACTGA
DMPS-Tit-15	TGCGTGTCTCCGACTCAGTGTGTCTC	DMPS-Tit-15-comp	GAGACACACTGA
DMPS-Tit-17	TGCGTGTCTCCGACTCAGTCTCTGTG	DMPS-Tit-17-comp	CACAGAGACTGA
DMPS-Tit-19	TGCGTGTCTCCGACTCAGTCTCATCA	DMPS-Tit-19-comp	TGATGAGACTGA
DMPS-Tit-21	TGCGTGTCTCCGACTCAGACACATAT	DMPS-Tit-21-comp	ATATGTGTCTGA
DMPS-Tit-23	TGCGTGTCTCCGACTCAGTACATATA	DMPS-Tit-23-comp	TATATGTACTGA
DMPS-Tit-25	TGCGTGTCTCCGACTCAGACAGTATA	DMPS-Tit-25-comp	TATACTGTCTGA
DMPS-Tit-27	TGCGTGTCTCCGACTCAGAGACTATA	DMPS-Tit-27-comp	TATAGTCTCTGA
DMPS-Tit-29	TGCGTGTCTCCGACTCAGACGAGAGT	DMPS-Tit-29-comp	ACTCTCGTCTGA
DMPS-Tit-31	TGCGTGTCTCCGACTCAGTCACACTA	DMPS-Tit-31-comp	TAGTGTGACTGA
DMPS-Tit-33	TGCGTGTCTCCGACTCAGTAGACACA	DMPS-Tit-33-comp	TGTGTCTACTGA
DMPS-Tit-35	TGCGTGTCTCCGACTCAGTCTACTCA	DMPS-Tit-35-comp	TGAGTAGACTGA
DMPS-Tit-37	TGCGTGTCTCCGACTCAGTCGAGAGA	DMPS-Tit-37-comp	TCTCTCGACTGA
DMPS-Tit-39	TGCGTGTCTCCGACTCAGTCACGAGA	DMPS-Tit-39-comp	TCTCGTGACTGA
DMPS-Tit-41	TGCGTGTCTCCGACTCAGTCTCAGAG	DMPS-Tit-41-comp	CTCTGAGACTGA
DMPS-Tit-43	TGCGTGTCTCCGACTCAGAGAGTGTG	DMPS-Tit-43-comp	CACACTCTCTGA
DMPS-Tit-45	TGCGTGTCTCCGACTCAGACGTGTGT	DMPS-Tit-45-comp	ACACACGTCTGA
DMPS-Tit-47	TGCGTGTCTCCGACTCAGAGCGCTGT	DMPS-Tit-47-comp	ACAGCGCTCTGA
DMPS-Tit-49	TGCGTGTCTCCGACTCAGACGCGACA	DMPS-Tit-49-comp	TGTCGCGTCTGA
DMPS-Tit-51	TGCGTGTCTCCGACTCAGACACGCGA	DMPS-Tit-51-comp	TCGCGTGTCTGA
DMPS-Tit-53	TGCGTGTCTCCGACTCAGAGCAGAGC	DMPS-Tit-53-comp	GCTCTGCTCTGA
DMPS-Tit-55	TGCGTGTCTCCGACTCAGAGAGCTGA	DMPS-Tit-55-comp	TCAGCTCTCTGA
DMPS-Tit-57	TGCGTGTCTCCGACTCAGAGCGATGA	DMPS-Tit-57-comp	TCATCGCTCTGA
DMPS-Tit-59	TGCGTGTCTCCGACTCAGACGCAGAT	DMPS-Tit-59-comp	ATCTGCGTCTGA
DMPS-Tit-61	TGCGTGTCTCCGACTCAGACAGCTGT	DMPS-Tit-61-comp	ACAGCTGTCTGA
DMPS-Tit-63	TGCGTGTCTCCGACTCAGACACTCTC	DMPS-Tit-63-comp	GAGAGTGTCTGA
DMPS-Tit-65	TGCGTGTCTCCGACTCAGAGAGTCTC	DMPS-Tit-65-comp	GAGACTCTCTGA

DMPS-Tit-67	TGCGTGTCTCCGACTCAGATGTCTCT	DMPS-Tit-67-comp	AGAGACATCTGA
DMPS-Tit-69	TGCGTGTCTCCGACTCAGAGATGCTC	DMPS-Tit-69-comp	GAGCATCTCTGA
DMPS-Tit-71	TGCGTGTCTCCGACTCAGACTAGAGA	DMPS-Tit-71-comp	TCTCTAGTCTGA
DMPS-Tit-73	TGCGTGTCTCCGACTCAGACATAGAT	DMPS-Tit-73-comp	ATCTATGTCTGA
DMPS-Tit-75	TGCGTGTCTCCGACTCAGAGATAGTG	DMPS-Tit-75-comp	CACTATCTCTGA
DMPS-Tit-77	TGCGTGTCTCCGACTCAGTATATGTG	DMPS-Tit-77-comp	CACATATACTGA
DMPS-Tit-79	TGCGTGTCTCCGACTCAGTCTAGTGT	DMPS-Tit-79-comp	ACACTAGACTGA
DMPS-Tit-81	TGCGTGTCTCCGACTCAGACATACTC	DMPS-Tit-81-comp	GAGTATGTCTGA
DMPS-Tit-83	TGCGTGTCTCCGACTCAGACAGTCAG	DMPS-Tit-83-comp	CTGACTGTCTGA
DMPS-Tit-85	TGCGTGTCTCCGACTCAGACATGCGC	DMPS-Tit-85-comp	GCGCATGTCTGA
DMPS-Tit-87	TGCGTGTCTCCGACTCAGAGATCAGA	DMPS-Tit-87-comp	TCTGATCTCTGA
DMPS-Tit-89	TGCGTGTCTCCGACTCAGAGCTACAG	DMPS-Tit-89-comp	CTGTAGCTCTGA
DMPS-Tit-91	TGCGTGTCTCCGACTCAGACTCACAG	DMPS-Tit-91-comp	CTGTGAGTCTGA
DMPS-Tit-93	TGCGTGTCTCCGACTCAGATACACAG	DMPS-Tit-93-comp	CTGTGTATCTGA
DMPS-Tit-95	TGCGTGTCTCCGACTCAGTATCACAC	DMPS-Tit-95-comp	GTGTGATACTGA
DMPS-Tit-97	TGCGTGTCTCCGACTCAGTCATACAC	DMPS-Tit-97-comp	GTGTATGACTGA
DMPS-Tit-99	TGCGTGTCTCCGACTCAGAGTATCAC	DMPS-Tit-99-comp	GTGATACTCTGA
DMPS-Tit-101	TGCGTGTCTCCGACTCAGACTCATAC	DMPS-Tit-101-comp	GTATGAGTCTGA
DMPS-Tit-103	TGCGTGTCTCCGACTCAGATCACTAC	DMPS-Tit-103-comp	GTAGTGATCTGA
DMPS-Tit-105	TGCGTGTCTCCGACTCAGAGTCATCT	DMPS-Tit-105-comp	AGATGACTCTGA
DMPS-Tit-107	TGCGTGTCTCCGACTCAGACTCTGAG	DMPS-Tit-107-comp	CTCAGAGTCTGA
DMPS-Tit-109	TGCGTGTCTCCGACTCAGATCTGCGC	DMPS-Tit-109-comp	GCGCAGATCTGA
DMPS-Tit-111	TGCGTGTCTCCGACTCAGTACTCTCA	DMPS-Tit-111-comp	TGAGAGTACTGA
DMPS-Tit-113	TGCGTGTCTCCGACTCAGTGATCTCT	DMPS-Tit-113-comp	AGAGATCACTGA
DMPS-Tit-115	TGCGTGTCTCCGACTCAGTGACATCT	DMPS-Tit-115-comp	AGATGTCACTGA
DMPS-Tit-117	TGCGTGTCTCCGACTCAGACGTGCAT	DMPS-Tit-117-comp	ATGCACGTCTGA
DMPS-Tit-119	TGCGTGTCTCCGACTCAGAGTCTGAC	DMPS-Tit-119-comp	GTCAGACTCTGA
DMPS-Tit-121	TGCGTGTCTCCGACTCAGATCAGTGT	DMPS-Tit-121-comp	ACACTGATCTGA
DMPS-Tit-123	TGCGTGTCTCCGACTCAGTATCGTGT	DMPS-Tit-123-comp	ACACGATACTGA
DMPS-Tit-125	TGCGTGTCTCCGACTCAGTCTGCACA	DMPS-Tit-125-comp	TGTGCAGACTGA
DMPS-Tit-127	TGCGTGTCTCCGACTCAGAGTAGCTC	DMPS-Tit-127-comp	GAGCTACTCTGA
DMPS-Tit-129	TGCGTGTCTCCGACTCAGATAGACAT	DMPS-Tit-129-comp	ATGTCTATCTGA
454 Tit-B-trunc	TGCCTTGGCAGTCTCAG	DMPS-Tit-B-comp	CTGAGACTGCCA
library amplification oligos (HPLC purified) - FLX Titanium Shotgun Adapters			
make-454 TIT-A		CCATCTCATCCCTGCGTGTCTCCGACTCAG	
make-454 TIT-B		CCTATCCCCTGTGTGCCTTGGCAGTCTCAG	


1 10 20 30 40 50 60 70 80 90 100 110
1. T67_ak_A0 GTGCCGTGCC GTGCCATGCC ATGCCACCCA CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
2. 65.1.all1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
3. 65.1.all2.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
4. 66.3.all1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
5. 66.3.all2.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
6. 4.1.A ?????????? ?????????? ?????????? ?????????? ???????CTC C????????? ?????????? ?????????? ?????????? ?????????? ??????????
7. 5.3.all1.RS ?????????? ?????????? ?????????A CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
8. 5.3.all2.RS ?????????? ?????????? ?????????A CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
9. 5.4.RS ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
10. D1.1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
11. 45.2.RS ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
12. 94.1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTG
13. 95.1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCG TCCCAGCACA GATCCCATCC CAACAGGGT?
14. 96.1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAG????
15. 101.2b.RS ?????????? ?????ATGCC ATGCCACCCA CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
16. 29.2.all1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
17. 29.2.all2.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGTC
18. 8.1b.RS ?????????? ?????????? ?????????? ?????????? ???????CTC C????????? ?????????? ?????????? ?????????? ?????????? ??????????
19. 31.2.all1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
20. 31.2.all2.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
21. 31.3.all1.A ?????????? ?????????? ?????????A CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
22. 31.3.all2.A ?????????? ?????????? ?????????A CCATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
23. 32.1.A ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
24. 34a.1.all1.RS ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAG????
25. 34a.1.all2.RS ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAG????
26. 40.1.RS ?????????? ?????????? ?????????? ??ATCCATGC AGAGGGTCTC CATGGTGCCA GCAGCAGCCC AGCTTCCCCA TCCCAGCACA GATCCCATCC CAACAGGGT?
120 130 140 150 160 170 180 190 200 210 220
1. T67_ak_A0 TAAACTGGGC ACCGTCCCAG CCAGCAGGTC CCCCTGCCTT GGGGACAGCC TGTGTCCC-- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
2. 65.1.all1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????CGGCG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
3. 65.1.all2.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????CGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
4. 66.3.all1.A TAAACTGGGC ACCGTCCCAG ?????????? ?????????? ?????????? ??TGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
5. 66.3.all2.A TAAACTGGGC ACCGTCCCAG ?????????? ?????????? ?????????? ??TGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
6. 4.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCC-- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
7. 5.3.all1.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCC-- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
8. 5.3.all2.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCC-- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
9. 5.4.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
10. D1.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ???????AGG ATGTCCTCAG GCTGACAGAG
11. 45.2.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
12. 94.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ??TGTCCCC- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
13. 95.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
14. 96.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????GGTG GGGGGG?ACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
15. 101.2b.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?GGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
16. 29.2.all1.A TAAACTGGGC ACCGTCCCAG ?????????? ?????????? ?????????? ??TGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
17. 29.2.all2.A TAAACTGGGC ACCGTCCCAG ?????????? ?????????? ?????????? ??TGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
18. 8.1b.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCC-- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
19. 31.2.all1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
20. 31.2.all2.A ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCCCG GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
21. 31.3.all1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCCC- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
22. 31.3.all2.A ?????????? ?????????? ?????????? ?????????? ?????????? ?GTGTCCCC- GGGTGCGGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
23. 32.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
24. 34a.1.all1.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????GGTG GGGGGGGACA TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
25. 34a.1.all2.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????GGTG GGGGGGGACG TCTCTCCAGG ATGTCCTCAG GCTGACAGAG
26. 40.1.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????


230 240 250 260 270 280 290 300 310 320 330
1. T67_ak_A0 GGTGACCACG GTGCTGGCGT GATGGTGAGA GAGCCAGGGC TTGGCCATGG TGTCCCTCTG TTCCTCTGCA TGGGTCTCGG GGTCACCGGG ACACCGGAAC AGTGTCCCGG
2. 65.1.all1.A GGTGACCACG GTGCTGGCGT GATGGTGAGA GAGCCAGGGC TTGGCCATGG TGTCCCTCTG TTCCTCTGC? ?????????? ?????????? ?????????? ??????????
3. 65.1.all2.A GGTGACCACG GTGCTGGCGT GATGGTGAGA GAGCCAGGGC TTGGCCATGG TGTCCCTCTG TTCCTCTGC? ?????????? ?????????? ?????????? ??????????
4. 66.3.all1.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
5. 66.3.all2.A GGTGACCACG GTGCTGGCGT GAAGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
6. 4.1.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
7. 5.3.all1.RS GGTGACCACG GTGCTGGCGT GAAGGTGAGA GAGCCAGGGC TTGGCCATGG TGTCCCTCTG TTCCTCTG?? ?????????? ?????????? ?????????? ??????????
8. 5.3.all2.RS GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TTGGCCATGG TGTCCCTCTG TTCCTCTG?? ?????????? ?????????? ?????????? ??????????
9. 5.4.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
10. D1.1.A GGTGACCACG GTGCTGGCGT G????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
11. 45.2.RS GGTGACCACG GTGCTGGCGT G????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????CCGGG ACACCGGAAC AGTG??????
12. 94.1.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
13. 95.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
14. 96.1.A GGTGACCACG GTGCTGGCGT G????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
15. 101.2b.RS GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC T????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
16. 29.2.all1.A GGTGACCACG GTGCTGGCGT GA????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
17. 29.2.all2.A GGTGACCACG GTGCTGGCGT GA????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
18. 8.1b.RS GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
19. 31.2.all1.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT????ATGG TGTCCCTCTG TTCCTCT??? ?????????? ?????????? ?????????? ??????????
20. 31.2.all2.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT????ATGG TGTCCCTCTG TTCCTCTGCA ?????????? ?????????? ?????????? ??????????
21. 31.3.all1.A GGTGACCACG GTGCTGGCGT GAAGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
22. 31.3.all2.A GGTGACCACG GTGCTGGCGT GAGGGTGAGA GAGCCAGGGC TT???????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
23. 32.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
24. 34a.1.all1.RS GGTGACCACG GTGCTGGCGT G????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
25. 34a.1.all2.RS GGTGACCACG GTGCTGGCGT G????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
26. 40.1.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
340 350 360 370 380 390 400 410 420 430 440
1. T67_ak_A0 TGGCCCCCAG AAGACCCACA CAGAGGGAGG CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
2. 65.1.all1.A ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
3. 65.1.all2.A ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
4. 66.3.all1.A ?????????? ?????????? ????GGGAGG CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCGTC TCCCCAAACC AAACCAGACC C????CGGCT
5. 66.3.all2.A ?????????? ?????????? ????GGGAGG CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC C????CGGCT
6. 4.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
7. 5.3.all1.RS ?????????? ?????????? ?????????G CAGAGGATAG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
8. 5.3.all2.RS ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
9. 5.4.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
10. D1.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
11. 45.2.RS ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
12. 94.1.A ?????????? ?????????? ?????????? ?????????? ????????CA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCGGCT
13. 95.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
14. 96.1.A ?????????? ?????????? ?????????? ?????????? ????????CA TCTGTGACCC CCGTGCTCAC CACCACCGTC T????????? ?????????? ??????????
15. 101.2b.RS ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCGTC TCCCCAAACC AAACCAGACC CATGCCGGCT
16. 29.2.all1.A ?????????? ?????????? ?????????? ?????????? ????????CA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
17. 29.2.all2.A ?????????? ?????????? ?????????? ?????????? ????????CA TCTGTGACCC CCGTGCTCAC CACCACCGTC TCCCCAAACC AAACCAGACC CATGCCAGCT
18. 8.1b.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
19. 31.2.all1.A ?????????? ?????????? ??????GCAG CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
20. 31.2.all2.A ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCGTGCTCAC CACCACCGTC TCCCCAAACC AAACCAGACC CATGCCGGCT
21. 31.3.all1.A ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
22. 31.3.all2.A ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
23. 32.1.A ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
24. 34a.1.all1.RS ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
25. 34a.1.all2.RS ?????????? ?????????? ?????????G CAGAGGATGG TGTGGGCACA TCTGTGACCC CCATGCTCAC CACCACCATC TCCCCAAACC AAACCAGACC CATGCCAGCT
26. 40.1.RS ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ?????????? ??????????

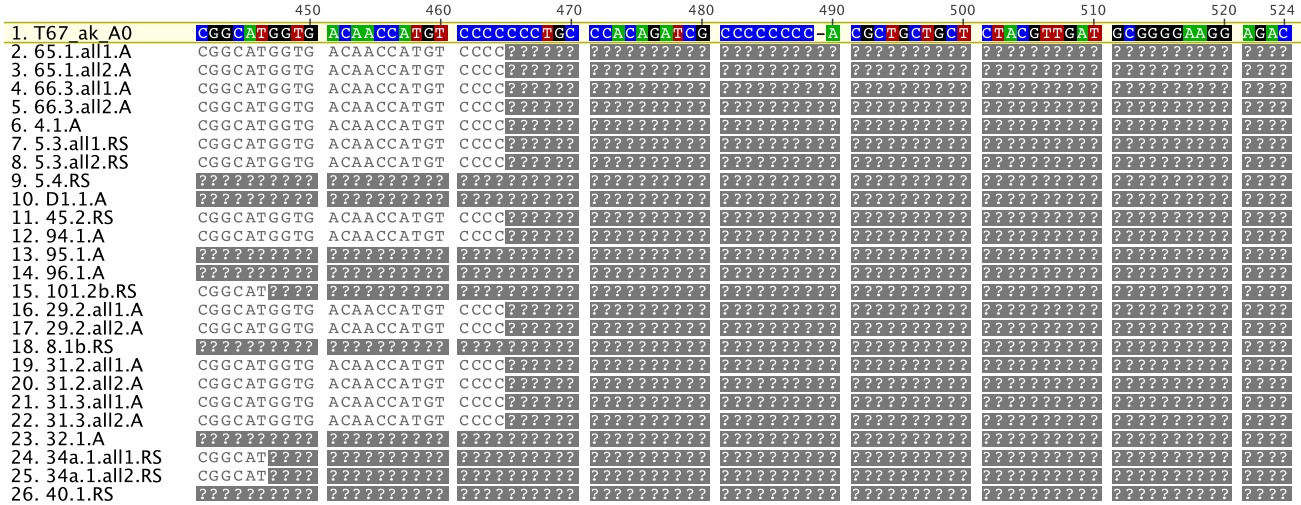


Figure III.5: Consensus ancient Adélie *AK1i5* sequences obtained from FLX sequencing.

IV APPENDIX FOUR

DRC Author contribution forms

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Gabrielle Beans-Picón

Name/Title of Principal Supervisor: Professor David Lambert

Name of Published Paper: The molecular ecology of the extinct New Zealand Huia

In which Chapter is the Published Work: Chapter Three, Appendix One

What percentage of the Published Work was contributed by the candidate: 20

Gabrielle Beans Picón — Digitally signed by Gabrielle Beans Picón

Candidate's Signature | **Date**

David Lambert — Digitally signed by David Lambert

Principal Supervisor's signature | **Date**

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Gabrielle Beans-Picón

Name/Title of Principal Supervisor: Professor David Lambert

Name of Published Paper: Is mitochondrial DNA diversity a reliable indicator of population size?

In which Chapter is the Published Work: Chapter Two

What percentage of the Published Work was contributed by the candidate: 80

Gabrielle Beans Picón

Candidate's Signature

Date

David Lambert

Principal Supervisor's signature

Date