Mitochondrial DNA Diversity and Variability in the Adélie Penguin of Antarctica

Gillian Claire Gibb

A thesis presented in partial fulfilment of the requirements for a degree of Master of Science in Genetics at Massey University, Palmerston North, New Zealand

2003

Cape Bird, Antarctica 2002

Abstract

In Antarctica, there are two distinct lineages of Adélie penguin (*Pygoscelis adeliae*) characterised by 8.3% divergence in mitochondrial DNA hypervariable region I (mt DNA HVR I). These two lineages are known as the *Antarctic* and *Ross Sea* lineages (*A* and *RS* respectively). This study aims to characterise aspects of mutation and variation as seen in HVR I of the Adélie penguin, by sequencing the DNA of individuals from different locations around Antarctica.

The geographic distribution of the two lineages was examined in greater detail. A dramatic decrease in the RS lineage was discovered on the edge of the Ross Sea region of Antarctica. Because the two lineages have different geographic distributions, and are separated by 8.3% sequence divergence, this study also investigated the possibility that these two lineages were in fact cryptic species. Sequencing of mt DNA and microsatellite genotyping proved that individuals of the two lineages mate randomly and produce offspring.

Recently, a rate of evolution based on serially preserved DNA from Adélie penguins was estimated at 0.96 substitutions/site/Million years. (0.53-1.43 s/s/Myr). This rate is four to seven times higher than previous avian control region evolution rates estimated by phylogenetic methods, and is more akin to rates of mutation determined by pedigree studies in other species such as humans. In the light of this higher direct estimate of the rate of evolution in Adélie penguins, this study also begins to determine a rate of mutation in Adélie penguins based on pedigree analysis. No new mutations were found, however three cases of inherited single point heteroplasmy were detected. The inclusion of heteroplasmy in mutation rate calculation is also addressed.

One of the arguments as to why pedigree studies find a higher rate of mutation than phylogenetic studies is that pedigree studies preferentially find mutations at 'hot spots' in the DNA sequence. This study also seeks to characterise the distribution of variable sites in hypervariable region I in relation to the two mt DNA lineages, and

also to geographic location. While the exact sites of variation differ between the two lineages, it was seen that the regions where variation was high or low is very similar in both lineages. This could be due to underlying physical constraints on DNA sequence variation.

Looking towards future work in Adélie penguin mt DNA and an expansion of the studies undertaken here, the complete mitochondrial genome of the Adélie penguin was determined. This now provides the opportunity to estimate rates of change in the entire Adélie penguin mitochondrial genome, using ancient DNA from the extremely well preserved sub-fossil bones in Antarctica.

Acknowledgements

Firstly, and most importantly I wish to thank my supervisor, Professor Dave Lambert for the opportunity to work on this project, and to visit Antarctica. You see the bigger picture, and I appreciate being able to be part of your vision. I have enjoyed working on this project very much.

This work was made possible by a Marsden Fund of New Zealand grant to Dave Lambert. I would like to acknowledge Massey University for financial support from a Molecular Genetics Research Scholarship.

I am indebted to Pete Ritchie for all the work he has done previously on Adélie penguins. You have been very patient with me, and had so much advice every time I had a problem. This project would not exist without the work you had done before!

Cheers to Craig Millar for help collecting samples down in Antarctica, and advice while writing up. Thanks also to Greg Arnold and Tom Parsons for statistical advice.

To everyone in the ME lab, past and present, for making a wonderful working environment, and for helping me get through this with some form of sanity. Thanks to Pete Ritchie, Hillary Miller, Lara Shepherd, Jennie Hay, Jenn Anderson, Gwilym Haynes, Leon (sigh) Huynen, Olly Berry, Jo Chapman, Quannah Hudson, Amy Roeder and Niccy Aitken. You've all been great.

A big thank you to everyone in David Penny's lab for help and advice in whole genome sequencing. Especially of course to Trish McLenachan for her enthusiasm, excitement and so much advice.

To Craig, Kerry, BJ and Malcolm for good friendship and an amazing time on the ice. Go penguins!

Last but certainly not least, a huge thank you to all my family and friends for a life outside this thesis! Thanks for your support and understanding. I could not have done this without you.

Preface

Many people were involved in collecting and analysing the Adélie penguin blood samples used in this thesis. Their contributions are listed below.

Blood samples from Cape Adare, Balleny Islands and Port Martin, Antarctica were collected in the austral summer of 2000/2001 by David Lambert (Massey University), John MacDonald and Peter Metcalf (Auckland University).

Adult Adélie penguin blood samples were collected from Cape Bird, Antarctica by Craig Millar (Auckland University), Peter Ritchie, Lara Shepherd (Massey University) and Bruce Thomas (Landcare Research, Nelson) in November 2001.

Adult and chick Adélie penguin blood samples were collected from Cape Bird by Craig Millar and myself in December 2001 and January 2002.

25 additional blood samples from adult and chick Adélie penguins from Cape Bird were kindly donated by Fiona Hunter (Plant and Animal Sciences, University of Sheffield).

7 DNA samples from Mawson and Davis were donated by Knowles Kerry (Australian Antarctic Division, Kingston, Tasmania). 16 DNA samples from the Antarctic Peninsula were donated by Carol Vleck (Iowa State University).

DNA from Davis, Mawson and the Antarctic Peninsula was sequenced by Peter Ritchie. I extracted, sequenced, sexed and genotyped DNA from all other blood samples used in this study.

All DNA sequencing was performed by Lorraine Berry (Massey University Sequencing Facility), as was the genotyping of T series samples. I genotyped the 25 samples donated by Fiona Hunter, with the kind assistance of Danielle Hubbard at the Equine Blood Typing Facility, Massey University.

I performed all post-sequencing analysis on samples used in this study.

Contents Page

List of Figures

List of Tables

List of Abbreviations

A		adenine
A	Ala	Alanine
A		*Antarctic* (mt DNA lineage)
aa		amino acid
ATPase	6, 8	Adenine triphosphate synthase subunit 6, 8
BAL		Balleny Islands
bp		base pairs
C		cytosine
C	Cys	Cysteine
°C		degrees Celsius
CA		Cape Adare
CB		Cape Bird
CCD		central conserved domain
CI		confidence interval
CO	I, II, III	cytochrome oxidase subunit I, II, III
CR		control region
Cyt*b*		Cytochrome *b*
D	Asp	Aspartic
D-Loop		displacement loop
DMP		Davis, Mawson and Antarctic Peninsula
DNA		deoxyribonucleic acid
E	Glu	Glutamic
F	Phe	Phenylalanine
G		guanine
G	Gly	Glycine
H		heavy (strand)
H	His	Histidine
HSP		heavy strand promoter
HVR	I, II	Hypervariable region I, II
hX		hypoxanthine
I	Ile	Isoleucine
K	Lys	Lysine
k_a		rate of evolution
kb		kilobase
km		kilometer
k_s		rate of nucleotide substitution
kya		thousand years ago
L		light (strand)
L	Leu	Leucine
LGM		last glacial maximum
LR		long repeat
LSP		light strand promoter
μ		rate of mutation

M	Met	Methionine
mt		mitochondrial
Myr		million years
n		sample size
N	Asn	Asparagine
NADH	1-6	Nicotinamide adenine dinucleotide dehydrogenase subunits 1-6
nt		nucleotide
O_H		origin of heavy strand replication
O_L		origin of light strand replication
P	Pro	Proline
PCR		polymerase chain reaction
PM		Port Martin
Q	Gln	Glutamine
R	Arg	Arginine
RNA		ribonucleic acid
rRNA		ribosomal RNA
RS		*Ross Sea* (mt DNA lineage)
S	Ser	Serine
s/s/Myr		substitutions per site per million years
SSR		simple short repeat
T		thymine
T	Thr	Threonine
TAS		termination associated sequences
tRNA		transfer RNA
U		uracil
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
yr BP		years before present
12S		12S rRNA subunit
16S		16S rRNA subunit

Chapter One

A Study of Mitochondrial DNA Diversity and Variability Using Adélie Penguins

1.1 The Mitochondrial Genome

This thesis is a study of mutation rates and processes in the mitochondrial control region of Adélie penguins (*Pygoscelis adeliae*), together with haplotype variation in the mitochondrial genome. The mitochondrial (mt) genome, and especially its control region has become a popular tool for resolving questions in evolutionary biology (Baker and Marshall 1997; Brown 1985). This is because of its specific properties that include maternal inheritance, a probable lack of recombination and a high rate of mutation (Stoneking 2000). These properties mean mitochondrial DNA has a much higher nucleotide diversity than the nuclear genome. In addition, its small size and high copy number makes the mt genome much easier to use.

The Adélie penguin is an extensively researched species. Many aspects of lifestyle, breeding and population biology, in addition to its genetics have been studied (Ainley 2002). The control region of the mitochondrial genome, and many nuclear microsatellite markers have been described (Lambert et al. 2002; Ritchie and Lambert 2000; Roeder et al. 2001). It is relatively straightforward to collect DNA samples from Adélie penguins in order to answer questions of population genetics and evolutionary studies. In addition, the cold dry climate of Antarctica, coupled with the

breeding biology of the Adélie penguin has allowed the extraction of ancient DNA from serially preserved bones found beneath living and abandoned Adélie penguin colonies (Lambert et al. 2002; Ritchie 2001). This allows the exciting addition of another dimension to genetic studies.

In this chapter a summary of the sources of mt DNA variability will be discussed. This is followed by overviews of the current literature on rates of mitochondrial genome evolution, the role of rate heterogeneity and variation in calculating these estimates and the uses of a highly variable mitochondrial control region in population genetics. The Adélie penguin as a model for studying mt DNA variability is also examined. The aims of this thesis will then be discussed in greater detail.

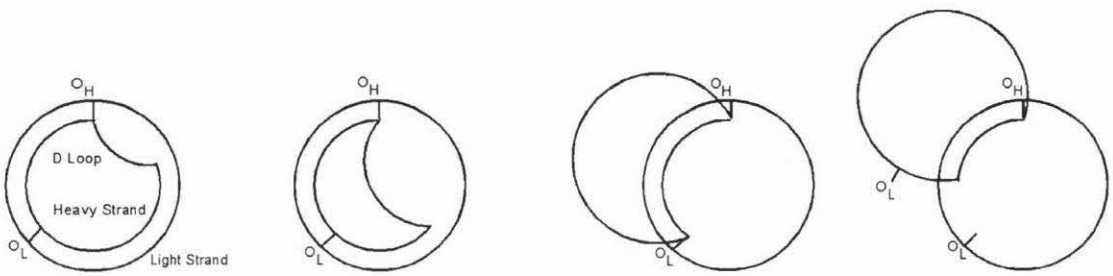
1.2 A Molecular Basis for the High Nucleotide Diversity of the Mt Genome

Mitochondrial DNA (mt DNA) has been observed to have a very high rate of sequence mutation relative to nuclear DNA (Brown, Jr. and Wilson 1979). The high rate of mt DNA mutation is thought to be partially a byproduct of the biochemical processes that govern DNA repair and replication of the mitochondrial genome (Avise 1991; Brown 1985). The proximity of the mt genome to the source of oxidative phosphorylation which takes place in the inner mitochondrial cell membrane, coupled with the fact that mt DNA lacks histones is part of the reason for such a high rate of mutation in the mitochondrial genome.

Another part of the reason for such a high rate of mt DNA mutation lies in its method of replication. The mitochondrial genome utilises a method called asymmetric replication (Figure 1.1a). In replication, synthesis of the daughter H strand begins in the displacement loop (D-Loop) in the control region by displacing the parental H strand. This parental H strand is now single stranded, and remains so until two thirds of the daughter H strand is synthesised. At this point the daughter L strand begins replication in the opposite direction. The consequence of this is that the parental L strand is never single stranded. However with replication taking up to two hours, cytochrome *b* on the parental H strand can be single stranded for approximately 80 minutes (Clayton 1982). During this single stranded phase, spontaneous deamination

of cytosine (C) to uracil (U) and adenine to hypoxanthine (hX) through hydrolytic attack can occur. If it is not repaired, when the daughter L strand is synthesised, U will pair with adenine (A) and hX with C, causing guanine (G) to be replaced by A and thymine (T) by C (Figure 1.1b). Hence, there is a high level of transitions in the H strand.

a)



b)

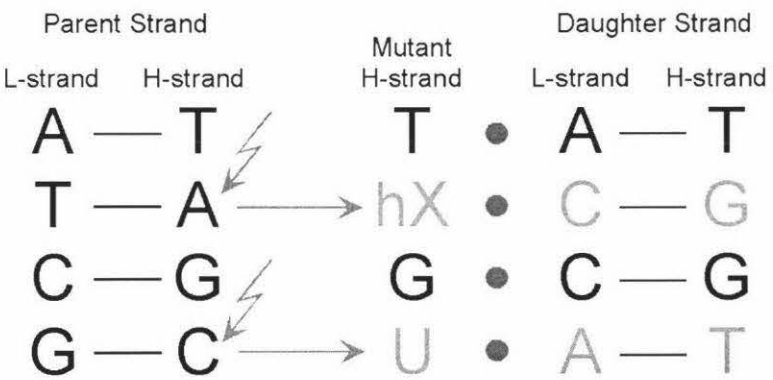


Figure 1.1
a.) Asymmetric replication of the mitochondrial genome. The black lines represent the parent molecule, grey lines the daughter molecule. O_L is the origin of light strand replication, O_H the origin of heavy strand replication.
b.) Deamination by hydrolytic attack of A and C to hX and U respectively causes high levels of A-G and C-T transitions on the H strand

1.2.1 The Mitochondrial Control Region

The control region (CR) is a noncoding segment of the mt genome that contains the promoters for transcription (light and heavy strand promoters: LSP and HSP respectively), the origin of heavy strand replication (O_H) and the displacement loop (D Loop) in vertebrates, as well as termination associated sequences (TAS), F, E, D and

C boxes and conserved sequence block 1 (CSB1) (Chang and Clayton 1985; Clayton 1982; Randi and Lucchini 1998). In most avian species including Adélie penguins the CR spans the region between the genes for tRNA-Glu and tRNA-Phe. Based on the distribution of variable sites and differing nucleotide frequencies, the control region is divided into three regions (Brown et al. 1986; Saccone, Pesole and Sbisa 1997). The 5' peripheral domain is known as hypervariable region I (HVR I). This is followed by a central conserved domain (CCD), then a second hypervariable region (HVR II). These three regions are sometimes also known as domains I, II and III respectively (Baker and Marshall 1997). The control region is often the most variable region of the mt genome, although in some species cytochrome *b* is more variable (Ruokonen and Kvist 2002). Most of the variability is in the two hypervariable regions, which show greater length variation and nucleotide variability than the central conserved domain (Brown et al. 1986; Ruokonen and Kvist 2002). The Adélie penguin CR is 1758 bp long, with HVR I being 560 bp, the CCD 200 bp and HVR II 998 bp long (Ritchie and Lambert 2000). The length of HVR II is only slightly smaller than the complete CR of many avian species (Baker and Marshall 1997).

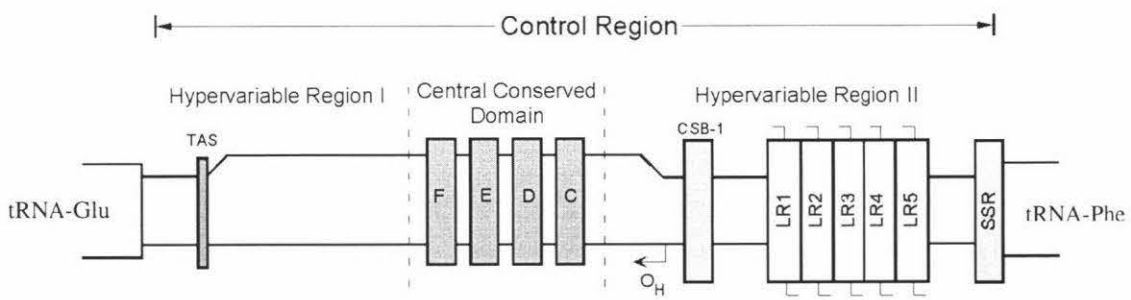


Figure 1.2. A diagram of the mitochondrial control region and flanking genes in Adélie penguins. OH, origin of heavy strand replication; LR1-5, 81bp large repeats 1-5 each containing putative sites for L and H strand promotion; SSR, a simple sequence repeat of 4 nucleotides. For additional abbreviations see text. Figure adapted from Ritchie and Lambert (2000).

1.2.2 Heteroplasmy and Bottlenecks

Another component of the diversity of mt DNA types is the presence of heteroplasmy and bottlenecks. Mutations in the mitochondrial genome occur at the level of the individual mitochondrion, not at the level of a cell. Somatic cells have between 10

and 10,000 mt DNA copies per cell, and oocytes have about 100,000 copies (Piko and Taylor 1987). When mutant forms of mt DNA arise, more than one mitochondrial type is then present in a single cell. This is known as heteroplasmy. Varying levels of heteroplasmy have been recorded in different tissues of a single individual as somatic mutations occur throughout the life of the individual (Bendall, Macaulay and Sykes 1997). When mutations occur in the oocyte, the variation can be passed on to the next generation. Transmission of only some oocyte variants of mt DNA to the next generation is known as the mitochondrial DNA bottleneck (Bergstrom and Pritchard 1998).

Mitochondrial DNA does not replicate until the 100 cell stage of the blastocyst. Because of this, the mt genomes within the oocyte are randomly sorted into each cell as they divide. This random sorting means only a small proportion of the mt DNA pool will be separated into the precursor germ cells to be transmitted to the next generation. In some species, such as humans, the mt DNA bottleneck appears to be very severe, as complete shifts from one genotype to another in a single generation have been detected (Parsons et al. 1997; Sigurðardóttir et al. 2000). Single point heteroplasmy, as well as length heteroplasmy has been observed between generations in many species (for example Parsons et al. 1997; Taylor and Breden 2002), although it does not appear to be common in humans (Sigurðardóttir et al. 2000).

Studies that have examined mitochondrial bottlenecks include Bendall et al. (1996), who investigated heteroplasmy in humans. Using DNA from twin studies, they found four heteroplasmic point mutations in 180 twin pairs. Changes in heteroplasmy within mother/offspring pairs were used to estimate developmental bottleneck size ranges. Their data suggested a bottleneck of between three and 20 segregating units, supporting the hypothesis of a tight generational bottleneck for mitochondria in humans.

Although studies in humans seem to suggest heteroplasmy is a rare observation (Bendall et al. 1996; Sigurðardóttir et al. 2000), studies in other species would perhaps indicate it is more widespread than previously thought. In other mammals such as cattle and mice the effective size of the bottleneck appears slightly larger (reviewed in Bendall, Macaulay and Sykes 1997). In guppies, individuals may

possess up to nine different sized mitochondrial haplotypes, and comparison of related individuals provided no evidence for a mt DNA bottleneck (Taylor and Breden 2002). As sequencing detection techniques improve and the cost of sequencing multiple cloned mitochondrial sequences decreases, detection of mitochondrial heteroplasmy is likely to increase.

1.3 Rates of Mutation and Evolution

1.3.1 The Neutral Theory of Molecular Evolution

In 1983, Kimura published a landmark work in the theory of neutral evolution. Briefly, neutral theory states that at the molecular level, most variability in species comprises variant mutant alleles that are selectively neutral, whose fate is determined mostly by random genetic drift (Kimura 1983). This theory describes how the rate of evolution (k_a) is related to the intergenerational rate of mutation (μ). The formula

$$k_a = N.\ \mu.\ P_{fix}$$

explains this relationship, where N is the number of haploid genomes and P_{fix} is the probability of fixation. Since any mt DNA sequence in a population could become fixed in a future generation, only a very small fraction of observable intergenerational mutations will be fixed, hence

$$P_{fix} = 1/N$$

Together, these two formulae give

$$k_a = N.\ \mu.\ 1/N$$

$$k_a = \mu$$

This means that under neutrality, the rate of evolution is equal to the rate of mutation and so one is an indirect measure of the other (Kimura 1983, summarised by Loewe, 1997).

Traditionally, the rate of mutation has been calculated indirectly through phylogenetic studies that use sequence variation between two related taxa calibrated to the fossil record to estimate the rate of evolution. Until recently, it has been very difficult to directly calculate the intergenerational rate of mutation. With the advent of high

throughput sequencing this has become feasible. Now it is possible to independently calculate k_a and μ and experimentally investigate Kimura's neutral theory of molecular evolution.

1.3.2 Calculating a Rate of Evolution

Also termed the 'relative branch length method' (Ruvolo 1996), phylogenetic studies estimate k_a, the rate of evolution along an ancestral lineage. Phylogenetic studies have in the past used restriction fragment length polymorphism (RFLP) and lately sequence variation between closely related species, for example humans and chimpanzees, to derive a phylogenetic tree using such methods as maximum parsimony, neighbour joining or maximum likelihood. Branch lengths are then calibrated by some external measure, such as species divergence dates from the palaeontological fossil record. This information can then be used to calculate a rate of nucleotide substitution (k_s) (Shields and Wilson 1987; Stoneking et al. 1992).

1.3.3 The Rate of Evolution in Birds

Shields and Wilson (1987) used this method to calculate a rate of substitution for the mitochondrial genome in birds. Using RFLP analysis, they calculated the extent of sequence divergence between species of the geese genera *Anser* and *Branta*. The midpoint root of the phylogenetic tree showed 9% sequence difference between the two genera, and the oldest fossil records from each genus were dated about 4-5 million years ago. By dividing the sequence difference by the age of the oldest fossils (9%/4.5 Myr), Shields and Wilson estimated a mean rate of divergence of 2%/Myr for the whole mitochondrial genome in geese. The authors argued that the similarity of their results and those from mammals was evidence for the validity of their calculations.

The results of Shields and Wilson (1987) were further extrapolated by Quinn (1992) to calculate the rate of substitution of the mt DNA control region. Quinn estimated that the control region evolves approximately 10.4 times faster than the overall mt

genome. This estimate was multiplied by the Shields and Wilson rate of 2%/Myr to give a rate of substitution for the control region of 20.8%/Myr (0.208 s/s/Myr). This result has subsequently been used as an estimate of the substitution rate of the mt DNA control region for all bird species (e.g. Baker and Marshall 1997).

However, such phylogenetic analyses do have inherent difficulties. They assume that subsequent to the point of divergence the two lineages evolved at the same rate, i.e., that the substitution rate is the same in *Anser* and *Branta*, or humans and chimpanzees (e.g., Stoneking et al. 1992). Another problem is the possibility of multiple substitutions occurring at the same site (especially in the rapidly changing mt DNA control region), and consequently these need to be corrected for (Howell, Kubacka and Mackey 1996).

Stoneking et al. (1992) attempted to correct these problems by the use of intraspecific calibration. Intraspecific calibration calculates sequence divergence *within* a species and avoids the shortcomings of comparing two species. Instead, the calibration is achieved by using monophyletic clusters of mt DNA types specific to defined geographic locations. Stoneking et al. wished to calculate the rate of mt DNA control region substitution in humans. To do this, the authors used samples from people living in Papua New Guinea, a location essentially colonised once, with a relatively well defined colonization time, and little back-migration (parameters to avoid errors in rate estimation). Using calculations of 'within group' and 'between group' mt DNA diversity, Stoneking et al. calculated a substitution rate of 11.8%/Myr (0.118 s/s/Myr) for the mt control region. This estimate is similar to that calculated by previous phylogenetic estimates in humans (Vigilant et al. 1991), and similar to the avian rate calculated in geese (Quinn 1992).

1.3.4 Calculating a Rate of Mutation.

The rate of mutation (μ) of a DNA sequence is defined as the mutation rate per individual per generation (Sigurðardóttir et al. 2000) and is then converted to substitutions per site per million years for comparison. This can be directly estimated through pedigree studies. Pedigree studies compare DNA sequences of closely

related individuals (mother-child, grandmother-grandchild and sibling-pair comparisons). The mutation rate is then estimated from the proportion of sequence changes observed.

Many key studies in the area of mutation estimation are based on human sequences, however the underlying arguments remain the same no matter which species is used. Since the onset of high throughput DNA sequencing techniques, directly estimating the rate of mutation through sequence comparison has become feasible. Use of this method in human pedigree studies has lead to estimates of mutation rate in the mt DNA control region ranging from 0 to 2.5 s/s/Myr (Parsons and Holland 1998).

Unfortunately, as can be seen from the summary in Table 1.1, this method has not lead to a consistent estimate of mutation rate. This is perhaps not surprising since these studies each use slightly different methods. Pedigree study estimates are on average much higher than the previous indirect estimates calculated by phylogenetic methods, but also cover a range that includes the phylogenetic estimates. This variation stems in part from debate about whether or not to include instances of heteroplasmy, the small size of some samples, the possibility of recombination and the importance of mutational hot spots.

Early results included those of Howell et al. (1996), who found two mutations in 81 transmission events. This gave a mutation rate of approximately 1 mutation per 40 generations (0.75 s/s/Myr). On its own, this study would appear to be an outlier, especially considering the small size of the study, and the fact that the four pedigree families chosen all contained the Leber hereditary optic neuropathy (LHON) disease. This is a mt DNA disease with pathogenic mutations that impair mitochondrial respiratory-chain function. It is entirely plausible that this abnormal mitochondrial metabolism may alter the rate of mutation in another part of the mitochondrial genome (i.e., the control region) (Jazin et al. 1998).

Table 1.1 Summary of studies into the rate of mutation

Study	Region Sequenced	Point muts. observed	No. of transmissions	Rate (s/s/Myr)
Howell et al. (1996)	HVR I & II	2	81	0.75
Bendall et al. (1996)	HVR I	4 heteroplasmies	180 twin pairs	0.06-1.35 (0.564 midpoint)
Parsons et al. (1997)	HVR I & II	10	327	2.5 (95% CI 1.2-4.0)
Jazin et al. (1998)	HVR II	0	208	<0.46 (99% CI 0.0-1.52)
Jazin et al. (1998)	HVR II pooled rate	7	804	1.17 (99% CI 0.15-2.2)
Parsons et al. (1998)	HVR I & II pooled rate	17	1065	1.35 (95% CI 0.72-1.98)
Sigurðardóttir et al. (2000)	HVR I & II	3 subs 3 hets	705	0.32 (95% CI 0.065-0.97)
Sigurðardóttir et al. (2000)	HVR I & II pooled rate	14	1221	0.852 (95% CI 0.46-1.42)
Heyer et al. (2001)	HVR I & II	4	508	0.39 (95% CI 0.113-0.917)
Heyer et al. (2001)	HVR I & II pooled rate	18	1729	0.517 (95% CI 0.34-0.737)
Howell et al. (2003)	HVR I & II pooled Howell et al. studies	3	263	1.02 (99.5% CI 0.09-3.97)
Howell et al. (2003)	HVR I & II pooled rate	28	2633	0.95 (99.5% CI 0.53-1.57)

However, soon after the high mutation rate report of Howell et al., a series of other reports were published. One of the larger studies undertaken was that of Parsons et al. (1997). They studied pedigrees from four different sources (the Armed Forces DNA Identification Laboratory (AFDIL), Oxford British families, CEPH pedigree cell lines and Old Order Amish pedigree lines); a large number of mitochondrial lineages totalling 357 individuals. Sequences from these individuals were used in the direct measurement of the intergenerational substitution rate of the human mitochondrial DNA control region. A total of 327 generational events were measured, resulting in the detection of 10 instances of substitution. This is also a high mutation rate, about one in 33 generations, or 2.5 s/s/Myr. Again, this is roughly 20 times higher than that predicted by phylogenetic analysis (Stoneking et al. 1992).

Jazin et al. (1998) argued that the high rate found by Parsons et al (1997) may be the result of mutational hot spots and a higher mutation rate in certain family lines (because of the disease state). Their own study, and that of Soodyall et al (1997) found no homoplasmic mutation events in 288 and 108 meiotic events respectively. Jazin et al. suggest pooling the results for HVR II to provide a more accurate result (see Jazin et al. and references therein). This yielded a mutation rate of 7/804 events, or 1.17 s/s/Myr (99%CI=0.15-2.2), a range that includes the phylogenetic rate estimate (~0.15) at its lower end, suggesting the two methods of estimating k_a and μ do not produce significantly different results.

Parsons et al. (1998) responded that Jazin et al.'s (1998) conclusions were due to the use of selective data and to inappropriate methods of analysis. They argue that by restricting their pooled results to only the HVR II, much published data on the control region was omitted. By pooling all published data on both the HVR I and HVR II (see Parsons et al 1998. and references therein), Parsons et al. achieved a resulting rate of 1.35 s/s/Myr (95% CI=0.72-1.98). Furthermore, they argue this result is conservative as it assumes in studies of only one hypervariable region that there were no additional mutations present in the other hypervariable region.

1.4 Mutational Hot Spots, Rate Heterogeneity

It has been shown that there is great mutation rate heterogeneity along the control region. In humans, it has been noted that roughly half the positions in HVR I are almost invariable (Meyer, Weiss and von Haeseler 1999). Although thousands of control region sequences have been determined in different species, knowledge of substitution patterns in the control region is far from complete. Control region evolution is very complex, among other things because base composition is not uniform, transitions occur more frequently than transversions, and substitution rates vary among sites. Meyer et al. (1999) have begun to unravel this complexity by looking at the pattern of nucleotide substitution in the human control region. They argued that accurate modelling of substitution rate variation is needed especially in a population genetics context. They suggested that inappropriate modelling may lead to a misinterpretation of data, for example mutational hot spots could mimic population expansion.

The question has been asked if mutational hot spots can account for some of the discrepancy between phylogenetic and pedigree analyses of mutation rate (e.g.Jazin et al. 1998; Stoneking 2000). Meyer et al. (1999) state their estimates of nucleotide substitution rates for each site in the two hypervariable regions will be useful in interpreting the conflict in mutation rate estimates for HVR I and II. In addition, there are benefits to sequence analysis in general as it allows the refinement of phylogenetic models and more precise interpretation of population sequence data.

Other authors have investigated the possibility that highly polymorphic sites represent mutational hot spots rather than old sites rooted early in the phylogenetic tree, and the degree to which these can account for the higher rates of substitution seen in pedigree studies. Gurven (2000) analysed linkage disequilibrium patterns in mt DNA, on the expectation that hot spots would show little or no disequilibrium as they can be interpreted as having randomly expressed patterns. Gurven's results suggested many polymorphic sites were in fact hot spots. Stoneking (2000) compared rates of mt DNA mutations occurring in germ line cells, heteroplasmic cells and somatic cancer cells (using HVR I and II nt site rates estimated by Meyer et al. 1999), and found that all three types of mutation occurred preferentially at hypervariable sites. This

supports the view that these represent mutational hot spots, not ancient sites rested early in the phylogenetic tree.

The analysis by Stoneking (2000) deserves further discussion. Not only did Stoneking show that mutations occur preferentially at hypervariable sites, but also that the rates of evolution are very similar for sites where germline, somatic and heteroplasmic mutations were observed. Stoneking suggested that this could be interpreted in two ways. Either the processes resulting in mutation are the same in germ line and somatic cells, or some percentage of the supposed germline mutations may actually be somatic mutations occurring early in prenatal development. The early somatic mutations would then drift to homoplasmy, with heteroplasmy reflecting an intermediate stage. Therefore supposed germline mutations are in fact somatic, hence the same observed rate. This second hypothesis could also be part of the answer as to why pedigree studies find a much higher rate of mutation than phylogenetic studies. Further work needs to be carried out to find out how mutations are transmitted to subsequent generations. It is however interesting to note that Sigurðardóttir et al.'s (2000) study of Icelandic populations was designed to rule out somatic mutations, and has one of the lower estimates of mutation rate (but still higher than estimates of evolutionary rate using phylogenetic methods).

Sigurðardóttir et al. (2000) agreed that there appears to be substantial heterogeneity of mutation rate across the control region, but argued that it is actually irrelevant to the estimation of mutation rate. The authors state that directly counting mutations by pedigree studies will give an unbiased estimate of the average mutation rate for the whole region, no matter how much rate heterogeneity there is within that region. This could be contested only if some sites mutate so quickly that back mutations occur within a pedigree, which seems unlikely. Howell et al. (1996) noted that rate heterogeneity is a problem for phylogenetic studies, which may consistently underestimate the average mutation rate unless they are able to accurately correct for multiple mutations at the same site. An accurate mutation rate estimation using phylogenetic methods requires a good understanding of rate heterogeneity.

Heyer et al. (2001) agree that there is rate heterogeneity in the control region, and propose a model where sites vary at one of three rates: slow, moderate or fast. They

argue that sites seen to mutate in pedigree studies are generally those that mutate at a faster rate. This is because the limited number of maternal transmissions studied means the time window is short, favouring those sites that mutate quickly. Slow and medium sites would predominate in phylogenetic rate estimates. Heyer et al. suggest this is because back mutations are more likely to occur at fast sites, masking their contribution to phylogenetic rate estimates.

1.5 Mt DNA Variability as a Tool in Population Studies

In addition to using mt DNA variation to determine rates of mutation and evolution, these data can also be applied to a diverse range of evolutionary problems. For example, mitochondrial DNA variation has been used to study theories of population migration and makeup through space and time. Because of its high polymorphism, mt DNA is an excellent marker for population studies on a smaller time scale than nuclear DNA.

Variation in mt DNA has been used many times to track population movement, especially migrations of modern humans. Matisoo-Smith et al. (1998) presented an alternative approach to questions about Polynesian settlement and mobility by looking at mt DNA control region variation in the Polynesian rat (*Rattus exulans*). The Polynesian rat was transported throughout Remote Oceania in the colonising canoes of the ancestral Polynesians. DNA phylogenies derived from mt DNA control region sequences of diverse Polynesian rats were used to test hypotheses on the degree of interaction within Polynesia. The results were more consistent with a general pattern of multiple contacts rather than isolation between Polynesian populations in the pacific.

Mitochondrial DNA can also be used to study population bottlenecks, such as those due to population isolation and fragmentation during a glacial maximum. A study by Wenink et al. (1994) of mt DNA HVR I looked at genetic variability in two shorebirds, the turnstone (*Arenaria interpres*) and the dunlin (*Calidris alpina*). By looking at population genetic structuring, Wennink et al. could hypothesize on their respective phylogeographic histories. Dunlins were shown to have considerable

variation in their sequences with strong genetic structuring subdividing the population. The variation could be explained by strong natal phylopatry combined with a high mutation rate in HVR I. The genetic subdivision was postulated to have evolved through population isolation and fragmentation into refugia during the Pleistocene. In contrast, the turnstones were shown to have very little HVR I variation and no definitive population structure. Wenink et al. (1994) postulated that this is the result of an expansion from a refugial population that was bottlenecked in the recent past.

1.6 The Adélie Penguin Model

Adélie penguins represent a unique opportunity to study rates of mutation, HVR I variation and phylogeography in a species other than human. The unique situation of Adélie penguins in Antarctica has provided the opportunity for a different approach to phylogenetic analyses of evolution rate. Beneath the densely populated living and abandoned Adélie penguin rookeries in Antarctica are large deposits of sub-fossil bones that have been continuously preserved for thousands of years (Baroni and Orombelli 1994). The cold and dry climate of Antarctica has meant these bones are good sources of ancient DNA (Ritchie 2001). This makes it an ideal situation for 'serial sampling', using ancient DNA (bones) and modern DNA (living penguins) to follow mt DNA sequence through time (Lambert et al. 2002). From a comparison of ancient and living taxa it is possible to directly estimate rates of nucleotide sequence evolution (Lambert et al. 2002).

Lambert et al. (2002) have been able to calculate a rate of evolution for the mt DNA hypervariable region 1 of Adélie penguins using the methods mentioned above in conjunction with two population based analyses: Markov Chain Monte Carlo analysis (MCMC) (Drummond et al. 2002) and a regression analysis (Drummond and Rodrigo 2000). In this way, Lambert et al. were able to calculate that the HVR I evolves at approximately 0.96 s/s/Myr (highest posterior density (HPD) interval 0.53 to 1.43). This estimate is approximately five times higher than previous indirect phylogenetic estimates (Quinn 1992), and is more akin to rate estimates estimated by pedigree analysis (e.g. Howell et al. 2003; Parsons et al. 1997).

As interesting as comparing evolutionary rates in one species to mutation rates in another might be, it would be more appropriate to compare this new rate of evolution to a rate of mutation in Adélie penguins. Unlike many bird species, Adélie penguins are well suited to calculating a rate of mutation through pedigree analysis. Pedigree analyses require two factors to calculate a rate of mutation: a large number of mt DNA transmissions (or generations) and the ability to identify siblings or the mother of an individual. Adélie penguins live most of their life amongst the pack ice off shore from Antarctica, but come ashore to breed in large colonies during the short austral summer (Ainley 2002). Both parents take turns sitting on the nest and going out to sea to fish. Because both parents are present on the nest at one time during the changeover it is possible to identify the whole family group at once. Adélie penguins raise one or two chicks, which stay at the nest site until they are about 22 days old (Ainley 2002; Spurr 1975). After this time chicks from neighbouring nests group together in crèches allowing both parents to feed at the same time (Ainley 2002). By the time the chicks are this age, it becomes difficult to distinguish family groups. Compared to many other bird species, whose nests are often widespread, hard to access and well camouflaged, it is relatively easy to collect DNA samples from Adélie penguins belonging to many families over a short period of time. This makes the calculation of a mutation rate from pedigree studies feasible. Indeed, this is a difficult task to achieve in many species other than birds as well.

Previous studies of Adélie penguin populations around the Antarctic have shown that there are two mitochondrial lineages comprising this species. These have been designated the *Antarctic* (*A*) and *Ross Sea* (*RS*) lineages (Ritchie 2001). These are so named because the *A* lineage has been recorded throughout Antarctica, while the *RS* lineage appears to be limited to the Ross Sea region. The two highly divergent lineages show 8.3% sequence divergence (Lambert et al. 2002). Using a MCMC approach, Drummond et al. (2002) have determined that the time to the most recent common ancestor of the two lineages was 75 kyr BP (95% CI 37 –122 kyr BP).

This estimate is consistent with what is known of the glacial history of Antarctica. It has been suggested that the two lineages are likely to have evolved as two isolated ancestral populations subjected to population bottlenecks during the last glaciation

(Ritchie et al. in press.). The estimated time to the most recent common ancestor of the two lineages places their divergence during the late Pleistocene, in the middle of the last glacial cycle (Ritchie et al. in press.). At this time ice covered a much greater area of Antarctica, and Adélie penguins were almost certainly greatly restricted in areas for breeding sites on the Antarctic mainland. Indeed, populations may have been mainly reduced to breeding on small offshore islands (Ainley 2002). It is under these conditions that the two lineages are expected to have evolved.

1.6.1 Looking to the Future

Rand (2001) called for "additional studies of the accumulation of mt DNA mutations in diverse organisms (e.g. Denver et al. 2000; Parsons et al. 1997). Such studies done in organisms with very different patterns of germ line differentiation will provide important comparative data on the fate of mt DNA mutations through the mitochondrial population booms and bottlenecks of extended germ lines." The work done previously on Adélie penguins, and that conducted in this thesis will help in understanding the problems put forward by Rand (2001).

The main focus of this thesis will be an analysis of the variation and diversity found in HVR I of the Adélie penguin. Looking towards the future, many of the analyses presented in the introduction and in the body of this thesis are also applicable to regions of the mt genome other than HVR I. The first step in performing analyses on part or all of the mt genome is knowledge of the complete mt genome sequence.

1.7 Thesis objectives

The objectives of this Thesis are as follows:

1) To examine the range of the *Antarctic* and *Ross Sea* lineages in the vicinity of the Ross Sea region of Antarctica (Chapter 3)

2) To determine if Adélie penguins of the *Antarctic* and *Ross Sea* lineages are or are not one species (Chapter 3)

3) To calculate a preliminary rate of mutation in the mitochondrial hypervariable region I for the Adélie penguin (Chapter 4)

4) To analyse the nucleotide diversity and variability in hypervariable region I as it relates to the two lineages and their geographic distribution (Chapter 5)

5) To sequence the entire mitochondrial genome of an Adélie penguin (Chapter 6)