Generation of a Species-Specific DNA Sequence Library of British Mammals

Jon H. Wetton, Gemma L. Braidley, Carol S.F. Tsang, Chris A. Roney, Sarah L. Powell & Adrian C. Spriggs

A study by the
The Forensic Science Service

for
The Joint Nature Conservation Committee
The Environment and Heritage Service, Northern Ireland

Forensic Science Service
Trident Court
Birmingham Business Park
Solihull
B37 7YN

Tel: (0121) 329 5428
Fax: (0121) 622 2051
Email: jwe02@fss.org.uk

Contents

Executive Summary

The forensic identification and population monitoring of animal species would benefit greatly from a reliable DNA test. To this end the Forensic Science Service has broadened the range of UK terrestrial mammals within the reference library for which species-specific cytochrome *b* DNA sequences have been verified. This gene is ideal for species identification as it shows limited variability within and much greater variation between species. By identifying regions of the gene sequence unique to each species a positive identification of any material containing DNA can be made. DNA can be used to determine the source of shed hairs and faeces amongst other items that might indicate the presence of an organism within the environment. The sequence data together with authenticated tissue and DNA samples will provide an invaluable resource in exploring the feasibility of DNA testing as a tool for the surveillance of British mammals. This study also provides the necessary groundwork for developing rapid and cost-effective hair identification tests in the future that may be capable of identifying species from a single shed hair.

Cytochrome *b* gene DNA sequences were derived from 28 mammal species using specimens supplied by the Joint Nature Conservation Committee (JNCC). Where possible samples were collected from diverse geographic locations and their sequences compared with published data from the GenBank internet database. Sequence data was derived by replicating overlapping fragments of the DNA code from tissue and hair samples from individual animals. DNA sequence variation between individuals was negligible in some species and extreme in others, often revealing variability unreported in the published literature. Although most variation appeared to result from authentic differences between individuals, several species showed evidence of nuclear copies of the cytochrome *b* gene. These alternative sequences located within the cell nucleus rather than the mitochondria have been unwittingly reported in previous studies and are a potential major source of confusion in the development and use of DNA tests for species identification.

1. Introduction

1.1 Monitoring mammal populations

The Joint Nature Conservation Committee on behalf of the Country Agencies (Countryside Council for Wales, English Nature, Scottish Natural Heritage and the Environment and Heritage Service, Northern Ireland) has a responsibility, under the Convention on Biological Diversity, to assess and report on the state of UK biodiversity. One project being undertaken by JNCC to achieve this goal is the setting-up of a national mammal surveillance network to cover all species of terrestrial mammals, both native and introduced, that occur in Britain.

In the past surveys have been carried out on a number of species individually to assess distribution and abundance. Some surveys have been repeated at intervals and the comparison of data from the surveys used to give an idea of population trends for the species in question. However, there have been no national multi-species schemes to monitor population trends for terrestrial mammals to match those in existence for birds, butterflies and, over the last 5 years, for bats.

One of the major reasons has been a lack of adequate methods to carry out surveillance on a very difficult group of species. Surveys using visual sighting are only appropriate for a small number of species because most mammals are nocturnal. Transect surveys, using sign information, such as droppings, foot prints, hairs and nest sites can be useful for a number of species but identifying some species using these methods can be extremely difficult. The collection of hair in hair tubes is a potential method for a number of species but has not been fully investigated as a method because of the problems associated with identifying hairs to the species level. However, the advent of DNA sequencing technology in the species testing field is now providing an opportunity for reliable species identification from single hairs.

DNA technology is becoming more widely used in conservation research and is likely to increase in importance in the future. The Forensic Science Service has been investigating the identification of species listed on Appendix I of CITES (Convention on International Trade in Endangered Species) using DNA sequencing techniques and has developed a test for tiger bone (Wetton *et al*. 2002). Furthermore, wildlife surveillance methods are becoming more technologically based and species identification using DNA testing is one area that is likely to develop in the future. Although the current approach of DNA sequencing is expensive, there is potential for the development of rapid and inexpensive identification tests which would make large scale monitoring of terrestrial mammals viable. The DNA sequence library developed in this study will provide the essential knowledge base upon which any further work would be built.

1.2 DNA Technology

Each organism's genetic blueprint is encoded in long chains of DNA, (deoxyribonucleic acid) built up of four different types of chemical link, usually referred to as the bases A, C, G & T. Like the dots and dashes of a Morse code message, the sequence in which the four bases are arranged along the DNA chain encodes genetic information guiding growth, development and function. The DNA chain is double stranded in which an A on one chain is always paired with a T on the other, whilst a C is always paired with G (see Figure 1). In this way each DNA strand can act as a template to reconstruct the other using the base-pairing rule that complementary bases A&T and C&G are always matched. Hence every cell in an organism will possess copies of the DNA code which are produced by repeatedly splitting and replicating the two strands at every cell division.

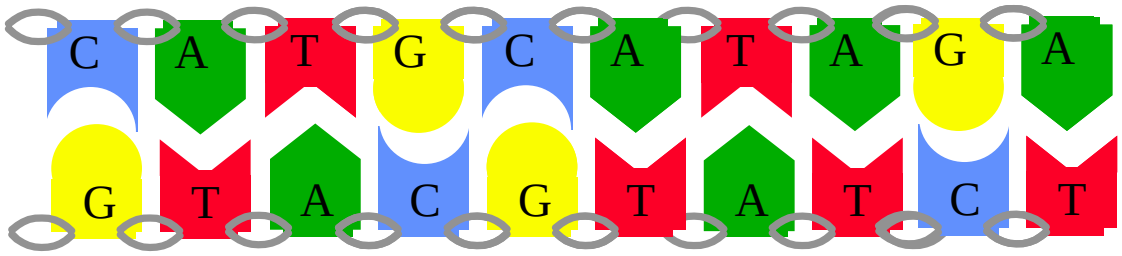

Figure 1 – Diagrammatic representation of the double stranded DNA chain showing A+T and G+C pairings.

In order to study the DNA code of an organism it is first necessary to replicate the region of interest within a test-tube to yield sufficient material for study. This is achieved using the Polymerase Chain Reaction (PCR).

1.3 The Polymerase Chain Reaction

PCR is a method that allows tiny quantities of DNA, as little as a single molecule, to be repeatedly copied within the test-tube until sufficient is available for detailed analysis. The replication process is directed towards specific areas of the DNA code through the use of synthetic single stranded DNA fragments known as primers. In a simple PCR reaction two primers are used, each is approximately 20 bases in length and is perfectly complementary to the sequence of bases flanking the region of interest. For a PCR reaction to proceed a small amount of DNA (the template) is placed in a test-tube, to this are added: primers, a solution containing the raw materials for synthesising more DNA (the four bases A, C, G & T) and an enzyme (*Taq* polymerase) which is capable of assembling these building blocks into copies of the original DNA. The template is used to determine the order of bases in the growing DNA chain.

In the first stage of the process, the denaturation step, the DNA is heated until the two strands separate, the solution is then rapidly cooled allowing the primers to bind to the complementary regions of the DNA code (annealing step). The solution is then warmed

to an intermediate temperature at which the *Taq* polymerase can replicate the DNA starting from the position at which the primers have bound (extension step). This cycle of temperature change is then repeated typically 30 times. At the end of each cycle the amount of DNA corresponding to the region flanked by the primers will have doubled, as each of the two DNA strands has acted as templates in the creation of a fragment of double stranded DNA. Consequently after 30 cycles as many as 2^{30} copies could have been created from each starting molecule; in reality the number is somewhat lower as not every template is successfully copied and the availability of bases becomes progressively reduced leading to a slowing of the replication rate in later cycles. As a result of the increased copy number the region is said to have been amplified.

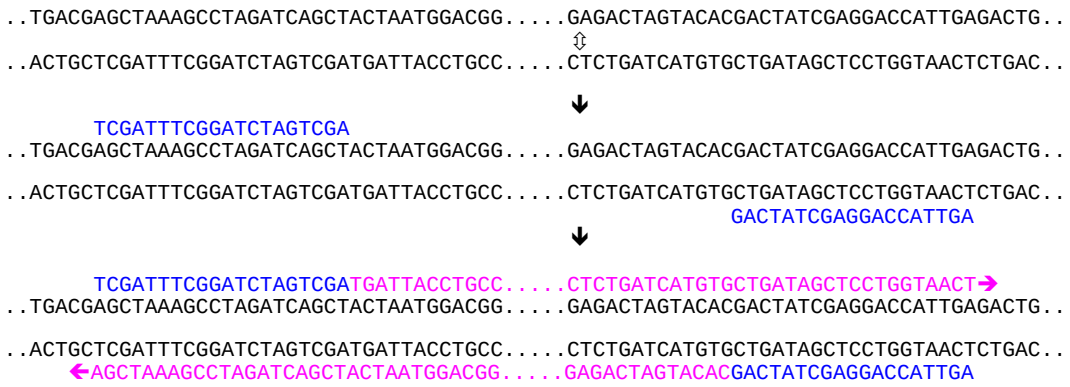

Figure 2 - Diagrammatic representation of PCR. Following separation of the two original DNA strands, synthetic primers (shown in blue) attach to complementary sequences flanking the region of interest and act as the starting point for synthesis of a new strand (shown in magenta) complementary to the template.

1.4 DNA Sequencing

Once sufficient DNA has been synthesised the species-specific order of bases within the DNA code can be determined. This is achieved by a process known as DNA sequencing which shares many characteristics with PCR. A single short synthetic primer (the sequencing primer) is used to initiate the synthesis of new DNA using the end product of the PCR reaction as a template. However in sequencing as well as the normal building blocks (the bases A, C, G & T) a much smaller number of modified bases with attached fluorescent dyes are added to the reaction mix, each of the four bases carrying a different coloured dye. Once a base with a fluorescent tag is incorporated into the growing chain the extension process is terminated. After 25 rounds of heating and cooling each of the many thousands of template molecules will have been used to produce up to 25 copies of the sequence which terminate at random positions within the code. The resultant soup of single stranded DNA fragments will comprise all possible lengths, but all fragments of the same length will terminate with the same colour base.

Before the order of bases along the code can be determined the DNA must be sorted by the process of electrophoresis, in which DNA fragments are drawn through a gel under the influence of an electric current. The gel acts as a molecular sieve ordering the fragments according to their length. During this process the ordered fragments pass a detector which determines the colour of the terminal fluorescent dye. A computer then

reconstructs the order of bases within the DNA fragment based upon the sequential passage of coloured dyes past the detector.

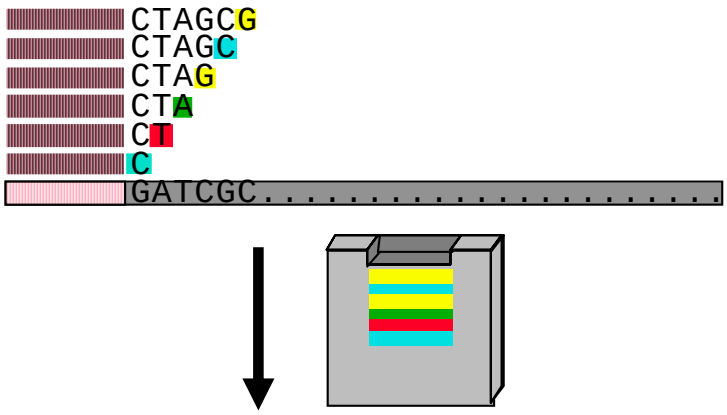

Figure 3 - Diagrammatic representation of DNA sequencing. Single stranded DNA (shown in grey) acts as the template for extension from a sequencing primer with a mixture of normal and dye tagged bases. When a dyed base is incorporated no further extension is possible. The dye-terminated fragments are then loaded onto a gel matrix and drawn through by an electric current. The sorted fragments reveal the sequence as a ladder of coloured bands.

1.5 The Cytochrome *b* gene and Species Identification

Certain critical regions of the DNA sequence have been extensively characterised by research scientists and shown to be reliable indicators of the species of origin. The cytochrome *b* gene is one such region varying little within species but showing distinct differences between species. Significantly this gene is located on a stretch of DNA which can be extracted from shed hairs. The Forensic Science Service has developed a species identification service based on the DNA sequence of the cytochrome *b* gene. By matching the gene sequence from a sample of unknown origin to a database it is possible to identify the species or, if a precise match cannot be found, draw conclusions on the likely source based on the similarity to other sequences of known origin. Using these methods to examine the entire cytochrome *b* gene we intended to identify particular base combinations that are unique to each species of UK terrestrial mammal. These base combinations could subsequently form the targets for synthetic molecules called probes that will only bind to the DNA of that species. Successful binding of the probe would be detected by interaction with fluorescent dyes that emit light only after the synthetic molecules have attached to their targets.

Identification of appropriate regions of DNA sequence ideally requires comparison of several individuals from each species. This study makes use of published information but supplements it with sequence data generated within our own laboratory from samples of known origin. Previous examinations of published cytochrome *b* sequences have revealed that some contain errors that would result in the failure of species-specific tests based upon the published data alone. By independently sequencing the DNA from each species we can confirm the validity of published data, or if there are discrepancies between the two sources of data only those regions which match will be considered as

potential targets. As some variation occurs within species as a consequence of naturally occurring mutations that have become established within populations, we are increasing the likelihood that the sequences observed are representative of the UK population in general by sequencing more than one individual from widely separated populations. Furthermore if they match a published sequence from a foreign sample then they are likely to be representative of the species as a whole.

Where possible, DNA sequences were to be obtained from two individuals from each of 30 mammal species and compared with published information held in the GenBank database. This small sample size was chosen to limit expense, however additional sequencing may prove necessary if subsequent prototype species-specific tests reveal previously unseen variation in a significant proportion of the UK population. However we would hope that the differences between species are sufficiently great that more than one species specific region will exist and this will reduce the chance of a false negative reaction where an individual lacks the characteristic traits possessed by the pair of individuals comprising the initial sample.

1.6 Genetic Variation

Variation between individuals and species arises through the progressive accumulation of mutations caused by errors in replicating the DNA strands. Mutation is a rare stochastic event, which leads to species diverging progressively over thousands of generations. Before base changes have spread fully throughout a population they will exist in both the original and mutated state, such populations are said to be polymorphic (having more than one variant). It is important to establish the extent of polymorphism within a population when choosing regions for species identification purposes. Populations which have retained a large size over very long periods may have accumulated many polymorphic variants, whilst population bottlenecks which reduce numbers to very low levels frequently result in the loss of genetic variation. Populations founded by immigration of very few individuals into a new area will often show limited, if any, variability. Populations that have been geographically isolated for many thousands of years may diverge substantially with each characterized by sequences differing at many sites. For this reason a comparison of the UK samples collected by JNCC with GenBank data from other countries may reveal significant differences.

1.7 Nuclear Copies

An alternative source of apparent variation is the occurrence of nuclear copies of the cytochrome *b* gene sequence. Cytochrome *b* is one of the 37 genes (13 protein encoding, 22 tRNAs and 2 rRNAs) within the circular mitochondrial genome which is typically 15-17,000 bases in length among mammals (Boore 1999). The mitochondria are intracellular structures that supply the energy requirements of living cells and are found in varying numbers dependant upon the metabolic activity of the cell. Despite hair shaft being a dead tissue, mitochondrial DNA is still present in low quantities.

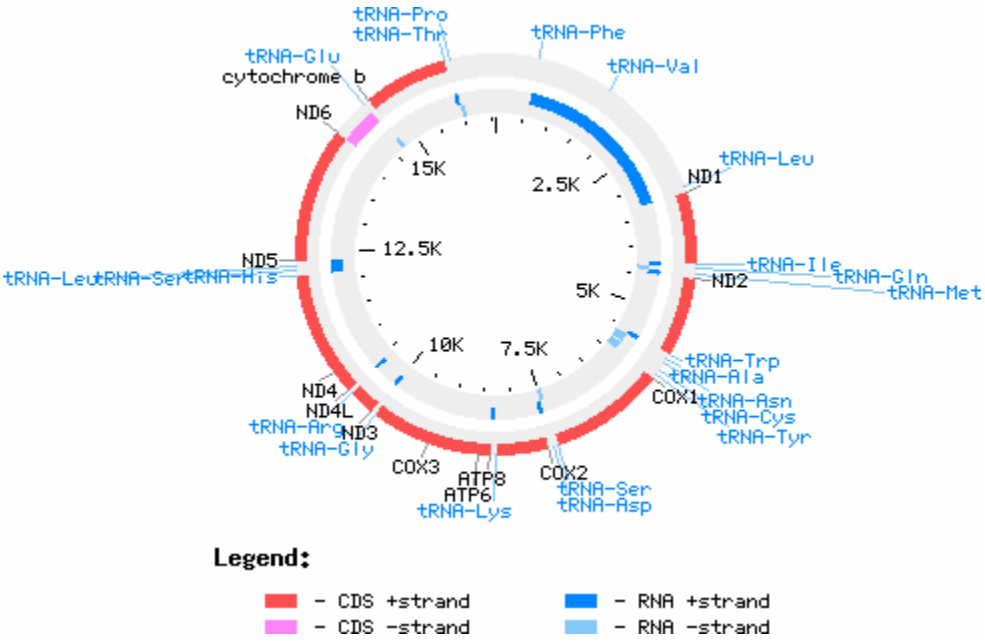

Figure 4 - The human mitochondrial genome shown above shares a standard arrangement of genes with all mammals. Protein encoding genes are labelled with black text; cytochrome *b* is flanked by the transfer RNA genes: $tRNA^{Glu}$ and $tRNA^{Thr}$..(Anderson *etal* 1981)

Cells contain many mitochondria, each possessing several copies of the mitochondrial genome consequently most cells contain hundreds or thousands of copies in total. By contrast the vast majority of genetic information is encoded by the billions of bases found in the cell nucleus. DNA in the cell nucleus exists as pairs of chromosomes, one inherited from each parent, and thus the majority of nuclear genes are present in just two copies in each cell. Many recent studies have shown that fragments of mitochondrial genetic material have become incorporated into nuclear chromosomes and in some cases have been subsequently replicated within the nucleus.

The mitochondrial and nuclear versions of cytochrome *b* gene fragments are differentially affected by natural selection and the biochemical environments within which they exist. Mitochondrial DNA is exposed to a harsh environment containing DNA damaging compounds that are the by-products of energy production. As a consequence the mutation rate is high. However, natural selection purges many of the changes induced in the DNA sequence as the alterations result in lowered efficiency of the cytochrome *b* gene product. Mutations only accumulate rapidly in regions of the gene which are not vital to its correct functioning. In contrast, the nucleus is a considerably more stable environment with highly efficient DNA repair mechanisms ensuring the integrity of the DNA code. Nuclear mutations occur at a much lower rate but are subsequently more likely to spread through populations as the nuclear copy is not constrained by the need to encode a functional cytochrome *b* protein and natural selection does not act upon the variants. Therefore, a random array of nuclear copy mutations are accumulated over many generations but at a rate at least ten times lower than the

functionally neutral changes accumulated in the authentic mitochondrial version of the cytochrome *b* gene.

When an array of PCR primers differing in sequence are used to amplify the DNA from genetically identical samples, some primer combinations might amplify a nuclear copy preferentially whilst the majority will replicate the authentic mitochondrial sequence. Consequently we might expect to see some discrepancies between overlapping PCR fragments and between sequences derived by other research groups as a result of the nuclear copy phenomena. Typically nuclear copies differ from the true mitochondrial sequence in an individual because of the selectively neutral changes that accumulate in the mitochondria. Comparison of two related species can therefore reveal a shared nuclear sequence that is similar to that present in the common ancestor of the two species.

1.8 Aims

The thirty target species to be sequenced were to be drawn from the following list of thirty-seven, dependant on the availability of adequate specimens from the JNCC. No samples were obtained from the eight species in grey text, and therefore efforts were concentrated on the twenty-nine remaining species.

Common name	Scientific name	*	Common name	Scientific name	*
Hedgehog	*Erinaceus europaeus*	I	Wild Cat	*Felis sylvestris*	C
Water Shrew	*Neomys fodiens*	I	Otter	*Lutra lutra*	C
Common Shrew	*Sorex araneus*	I	Pine Marten	*Martes martes*	C
Pigmy Shrew	*Sorex minutus*	I	Badger	*Meles meles*	C
Mole	*Talpa europaea*	I	Stoat	*Mustela erminea*	C
			Weasel	*Mustela nivalis*	C
Yellow-necked Mouse	*Apodemus flavicollis*	R	Polecat/Ferret	*Mustela putorius*	C
Wood Mouse	*Apodemus sylvaticus*	R	American Mink	*Mustela vison*	C
Water Vole	*Arvicola terrestris*	R	Fox	*Vulpes vulpes*	C
Bank Vole	*Clethrionomys glareolus*	R			
Fat Dormouse	*Glis glis*	R	Roe Deer	*Capreolus capreolus*	D
Harvest Mouse	*Micromys minutus*	R	Red Deer	*Cervus elaphus*	D
Field Vole	*Microtus agrestis*	R	Sika Deer	*Cervus nippon*	D
Orkney Vole	*Microtus arvalis*	R	Fallow Deer	*Dama dama*	D
House Mouse	*Mus musculus*	R	Muntjac	*Muntiaacus reevesi*	D
Common Dormouse	*Muscardinus avellanarius*	R			
Coypu	*Myocastor coypus*	R	Brown Hare	*Lepus capensis*	L
Brown/Common Rat	*Rattus norvegicus*	R	Mountain Hare	*Lepus timidus*	L
Black Rat	*Rattus rattus*	R	Rabbit	*Oryctolagus cuniculus*	L
Grey Squirrel	*Sciurus carolinensis*	R			
Red Squirrel	*Sciurus vulgaris*	R			

Table 1. * I=insectivore, R=rodent, C=carnivore, D=deer, L=lagomorph

For each species DNA was to be extracted from samples provided by the JNCC from two geographically isolated populations where possible. The cytochrome *b* gene sequence would then be replicated by PCR and its sequence determined. Comparisons could then be made between the JNCC samples and data already submitted by the wider scientific

community to the GenBank sequence database (accessible at **www.ncbi.nlm.nih.gov)**. This would highlight both invariant and variable DNA regions within and between species and invariant sequences within species that consistently differ in sequence from all other species could be targeted in future species identification assays.

1.9 Future Species Identification Tests

Currently the most promising approach to species identification tests would involve the design of synthetic DNA molecules with complementary sequences targeted against sequence regions that are unique to particular species. A synthetic sequence which was exposed to DNA from a biological trace sample such as shed hairs would be capable of binding if a match was present and this would result in some detectable change such as release of fluorescent light, if a matching DNA sequence was not present in the hair sample then no light would be released. By attaching fluorescent tags which emit at different wavelengths to each species-specific synthetic DNA sequence the species composition can be determined from the wavelengths of light emitted. The principle is demonstrated below.

Three species, **X**, **Y** & **Z** each possess a different cytochrome *b* sequence, a small section of which is shown with the differences underlined. Five synthetic DNA sequences are also shown; Sequence **1** would bind to species **X** as the sequence is perfectly complementary (G paired with C and A paired with T), Sequence **2** is perfectly complementary to both species **Y** and **Z** and therefore would bind to both as this region of their cytochrome *b* gene is conserved (invariant). By contrast the second half of the region differs between all species and Sequence **3** will only bind to **X**, Sequence **4** to **Y** and Sequence **5** to **Z**. If DNA was extracted from a hair sample left by species **X**, only sequence **1** and **3** would bind and the fluorescent light emitted would be just red and green, if Species **Z** had also contributed to the hair sample, the colours detected would be red, green, yellow and blue. As species **Y** never contributed hairs to the examined samples Sequence **4** cannot bind and purple light would not be detected.

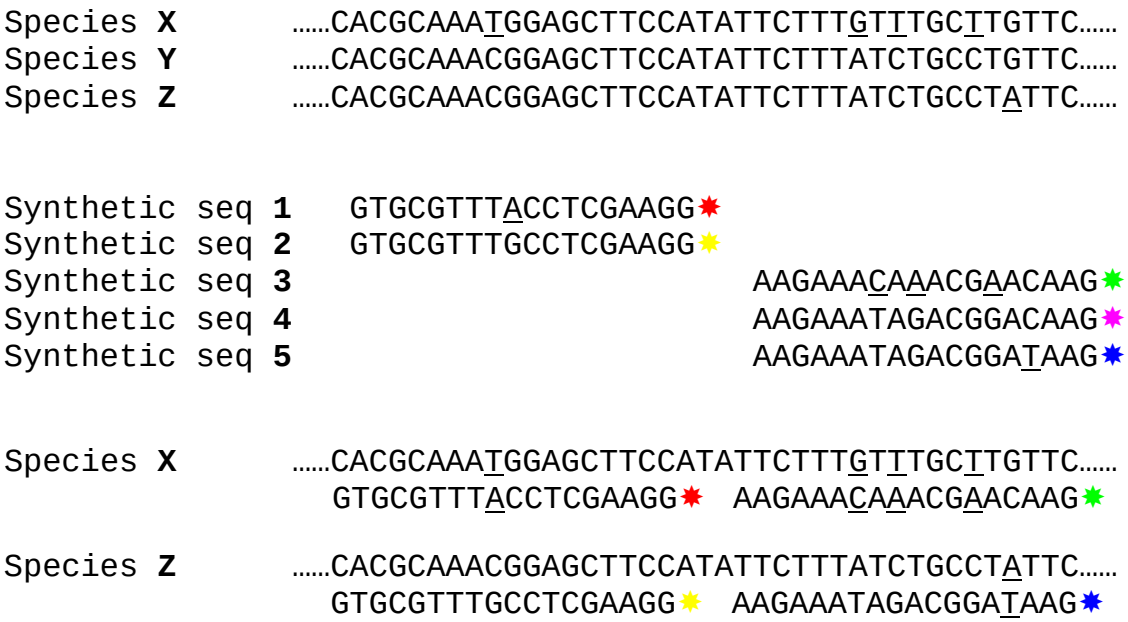

Figure 5 – Diagrammatic representation of a species identification test based on fluorescently tagged synthetic sequences (probes).

2. Methods

2.1 Primer Design

Primers were designed based upon alignments of published GenBank sequences from the species listed in Table 1. All sequences available in February 2001 were downloaded from the Internet, the sequences were aligned so that the bases at a particular point within the gene could be compared across sequences. Differences were detected by creating an ambiguity sequence; a string of characters in which invariant bases across sequences are represented by a '-' and variable bases by a '*'. A consensus sequence was also constructed in which invariant bases were shown (A, C, G or T) and variable bases were encoded depending on the combination of different bases that occur at that specific point across the sequences (see Table 2). Computer printouts of ambiguity and consensus sequences were used to rapidly identify conserved (unchanging) and polymorphic (variable) sites between sequences (see Figure 7 for example).

A = Adenosine	**R** = A or G (puRine)
C = Cytidine	**Y** = C or T (pYrimidine)
G = Guanosine	**K** = G or T (Keto)
T = Thymidine	**M** = A or C
B = C, G, or T	**S** = G or C (Strong 3H bonds)
D = A, G, or T	**W** = A or T (Weak 2H bonds)
H = A, C, or T	**N** = aNy base
V = A, C, or G	

Table 2 – IUB Codes

Mutation in mitochondrial sequences is strongly biased towards transitional (A↔G & T↔C) changes rather than the alternative transversional changes (A↔C, A↔T, C↔G & G↔T). The 1140 bases of the cytochrome *b* gene encode the order of the 380 amino acids that make up the cytochrome *b* protein chain, each amino acid being specified by a consecutive triplet of bases, known as a codon. Differences occur predominately at the third base within a codon; this base can usually change without altering the amino acid sequence of the protein. Some variation at first base positions is also "silent" with respect to amino acid encoded, however all changes at the second position within the codon result in a replacement of the amino acid with an alternative. Hence there is a distinctive distribution of the occurrence of changes (see Figures 6 and 7).

```
  AAs  = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = --------------------------------MMMM--------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Figure 6 – The vertebrate mitochondrial genetic code

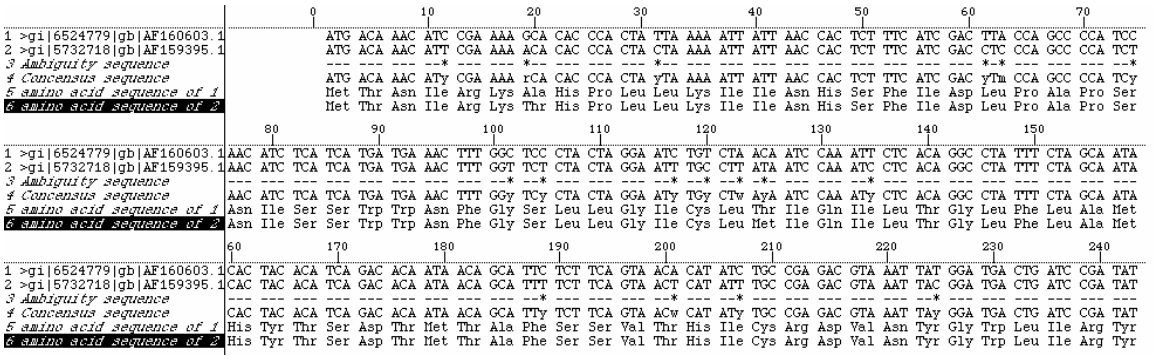

Figure 7 - GenBank entries for *Apodemus sylvaticus* aligned to show ambiguity, concensus and amino acid sequences. Note the distribution of polymorphic sites and how this relates to the protein sequence.

Certain regions of the protein sequence are vital to the efficient functioning of the cytochrome *b* protein and these regions show little variation between species (Griffiths 1997). Consequently the internal PCR primers were targeted against such regions and were frequently designed to terminate on a second base position within a codon as this critical position is the least likely to differ within or between species. The sequence alignments used to identify the primer sites can be found in Appendix 5 and the primers themselves in Appendix 4.

2.2 Sample Collection

Samples were obtained by the JNCC from a number of sources, including animals trapped during monitoring programs, roadkills and museum specimens. Samples from live animals were either plucked or cut hairs and blood samples where available. Samples of hair and/or tissue were obtained from roadkills and museum specimens. Hairs were stored at ambient temperature within clip-seal polythene bags whilst blood and tissue were stored in several volumes of ethanol at or below 4^{o}C prior to delivery to the FSS. Subsequent storage was at –20^{o}C.

Due to the onset of Foot & Mouth Disease and the subsequent restrictions on access to the countryside, collection of samples was seriously affected resulting in the availability of fewer fresh samples and greater use of old preserved material that might have a greater degree of DNA damage.

2.3 DNA Extraction

DNA was recovered from the available samples using Qiagen extraction columns (see Appendix 1 for experimental details). Excess hair, blood and tissue was retained to provide a source for future investigations.

2.4 PCR

To obtain DNA sequence data from the entire cytochrome *b* gene it was necessary to amplify a larger region. The transfer RNA genes, $tRNA^{Glu}$ & $tRNA^{Thr}$ flank the gene. Primers were designed which would bind to these genes using published data obtained from GenBank covering species from each of the taxa of interest; Insectivores, Lagomorphs, Rodents, Deer and Carnivores. The length of an amplicon spanning the region between the two tRNA primers was approximately 1,300 bases, a span unlikely to be successfully amplified from hair shaft or degraded tissue. Additional primers were designed to bind approximately one and two-thirds into the cytochrome *b* gene such that the whole region could be amplified in three overlapping fragments, none of which exceeded 600 bases in length. Not only does this increase the likelihood of amplifying intact substrate; it provides additional evidence of the authenticity of the sequences produced.

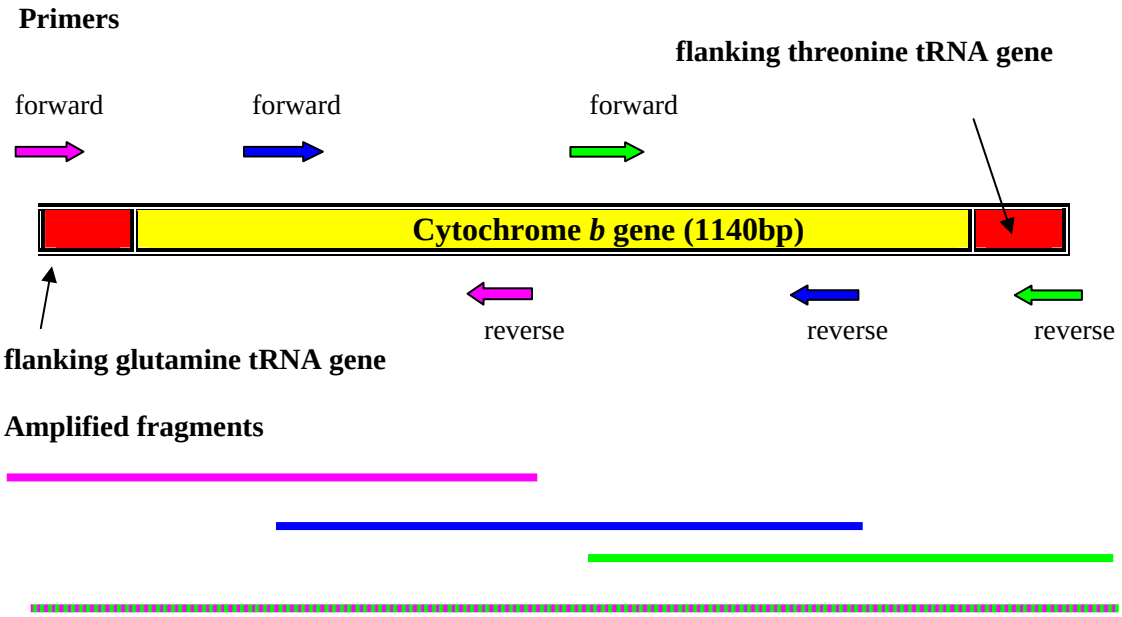

Figure 8 –Representation of the relative positions of the three pairs of internal primers in relation to the entire gene and the four overlapping fragments produced.

Nuclear copies of the cytochrome *b* gene are often present in addition to the desired mitochondrial target. In general these will be present in lower copy numbers but by virtue of the reduced evolutionary rate of nuclear sequences compared with those retained in the mitochondrial environment, they are more likely to resemble the target sites of primers designed to match an array of related species, hence nuclear copies may be amplified in preference. However the lower copy number should ensure that nuclear copies will usually be at an initial disadvantage during the replication process and thus nearly pure mitochondrial sequence or a mixture of nuclear/mitochondrial is the expected result. By comparing the regions of overlap between fragments it should be possible to

identify the presence of nuclear sequences as these will lead to incompatible regions of overlap with true mitochondrial sequence.

2.5 DNA Sequencing and Sequence Analysis

Both strands of the four fragments derived from each sample were sequenced using the amplification primers and Big-Dye Terminator sequencing chemistry. The reactions were electrophoresed on an Amersham MegaBACE 1000 96 sample capillary electrophoresis machine (as detailed in Appendix 1). The sequence data was checked by eye for quality, and trimmed to remove primer sequences. Each sequence was imported into a Sequence Navigator layout within which the fragments were orientated (forward and reverse strands) and aligned with a cytochrome *b* reference sequence from the cow (GenBank accession NC_001567). Regions of overlap were checked for compatibility and any discrepancies examined.

2.6 Polymorphism Identification

Where sequence data from more than one individual was available any differences were recorded, this was extended to a comparison with other data from the species accessed through the GenBank database. A unanimity sequence for each species showing invariant bases was created to highlight potential sites for species-specific primers or detection probes.

3. Results

3.1 PCR Amplification

DNA extracts were obtained from the majority of species for which samples were provided. The sole small clipped hair sample from the water shrew (*Neomys fodiens*) yielded insufficient DNA for analysis and that from the pigmy shrew (*Sorex minutus*) was shown to be virtually identical by sequencing to the common shrew (*S. araneus*) suggesting probable misidentification of the sample in the field. Full or partial DNA sequences were obtained from all other samples. As expected tissue samples generally yielded greater quantities of DNA than hair samples due to the much higher concentration of DNA within them and the larger amount of available material. Amplification from hair samples showed varying levels of success with excellent results from freshly plucked deer hairs whilst barely detectable quantities were recovered from some of the finer haired small mammals.

The initial intent was to derive DNA sequence data from hair wherever possible but it became apparent that the low starting quantities of hair DNA combined with the non-optimal primers which may include mismatches with the target region within the gene of some species yielded insufficient DNA to sequence some fragments. Some mismatches between primers and target regions were inevitable but was exacerbated by the paucity of published data (particularly for the flanking tRNA primers) and the need to design universal primers for groups of species to avoid the prohibitive cost of synthesising specific primers for each species. Hence tissue samples were used when available following failed amplification from hair.

For future work it may be possible to use the sequences derived in this study to design better primer sets which bind to specific groups of related species which occur in the same habitat. Such primers would target shorter regions of the order of 100bp in length, these are more likely to detect a higher proportion of intact DNA fragments within hair samples. As hair shaft is essentially a dead tissue, DNA degradation is a common problem and amplifying shorter regions is generally more successful. Furthermore, the quantities needed for sequencing exceed the amount that can be detected by alternative methodologies and thus it remains a realistic possibility that hair samples, which were unsuitable for the preliminary sequencing work, will yield sufficient DNA for optimized species-specific tests. However small numbers of hairs from the smallest and shortest haired insectivores and rodents may pose a significant challenge.

3.2 DNA Sequencing

The DNA sequences derived during this study are presented in Appendix 7. Due to the large amount of sequence data produced; four fragments per individual sequenced in both directions, only a summary sequence for each species is presented here. The full sequence data is available on request as a CD in Sequence Navigator format. Here we

discuss some interesting findings which are of relevance to species identification problems.

Several species demonstrated very low levels of polymorphism (sequence variation), both in the samples provided by the JNCC and within the GenBank database. For example the Pine Marten (*Martes martes*) showed no variation between the two JNCC samples, and just one and two differences from samples collected in North America (GenBank accession AF154975) and Russia (AB051237), suggesting that variation within the species might be limited or at least that common haplotypes have very wide geographic distributions. By contrast the Hedgehog (*Erinaceus europaeus*) sample collected by JNCC mismatched with the GenBank reference (NC_002080) for the species at 59 bases resulting in only 94.8% similarity along the full length of cytochrome b. However, the authenticity of both sequences is confirmed by a study of European Hedgehogs (Seddon *et al.* 2001), which sequenced a smaller region of the cytochrome *b* gene, and demonstrated that several highly divergent sequences exist in populations spread across Europe. The JNCC sample matched with just one base difference the E2-01 mitochondrial haplotype characteristic of UK populations over their 381 base overlap whilst the reference sequence matched haplotype E1-01 which is found at high frequency in Scandinavia and Northern Europe southwards into Italy. Since identical primers were used by Seddon *et al*. in their study we can assume that the results reflect a true but extreme polymorphism within Europe.

In many cases the GenBank entry accompanying a published sequence gives little or no information on the geographic source of a sample and for some species such as the House Mouse (*Mus musculus*) and Brown Rat (*Rattus norvegicus*) the sequence is derived from lab strains or even cell lines. The two JNCC samples of House Mouse show a single base difference between them whilst one matches perfectly with six out of seven GenBank reference sequences. The remaining GenBank sequence (AB033699) shows 27 base changes from the other reference sequences. Whether this represents true mitochondrial polymorphism could not be determined directly but examination of the original paper (Suzuki *et al.* 2000) reveals the source to be Lab strain MOA of *M.m molossinus*, the Japanese House Mouse, hence the difference may reflect sub-specific variation.

Extreme polymorphism within the JNCC samples was seen only in the Water Vole (*Arvicola terrestris*); the two samples differing at 39 bases. Neither matched the two available GenBank sequences though a difference of just 8 bases was observed between one JNCC sample and AF119269. The two GenBank sequences differed by 18 bases (98.4% similarity).

Other instances of extreme sequence variation may be influenced by differential amplification of nuclear copies by authors using different primer combinations (Wetton *et al.* 2002).

Clear evidence of nuclear copies was revealed in a number of ways. Prior to sequencing all PCR products were electrophoresed through an agarose gel in the presence of ethidium bromide, a fluorescent dye that binds to DNA, to confirm that the region of

interest had been successfully amplified. When attempting to amplify the third fragment of Fox (*Vulpes vulpes*) an additional fragment some 50 bases longer than expected was also produced, sequencing demonstrated that both fragments shared similar sequences but one, presumably nuclear, contained an inserted region. Data for this fragment is incomplete as the mixed sequence was uninterpretable beyond the insertion. Previous sequencing of canid species in our laboratory has shown evidence of many different nuclear copies that vary in length. Other cases of uninterpretable mixtures of mitochondrial and nuclear sequences were observed for the third fragment in Mink (*Mustela vison*) and Rabbit (*Oryctolagus cuniculus*).

3.3 An Examination of Collection Tapes

The primary method for collecting hair samples for mammal surveillance is the use of baited tubes of appropriate diameter within which are small patches of double-sided sticky tape. It was planned to carry out a series of trials to identify a suitable tape for hair capture and estimate the likely costs of recovering material for DNA analysis. Initially a number of tape manufacturers were contacted with details of our requirements and samples of suitable tapes acquired. Each type was to be tested against the criteria listed below.

1 Sufficient adhesion to capture hairs from passing mammals
2 Ability to retain hairs in all weathers
3 Ideally the tape will be compatible with the chemicals and enzymes involved in breaking down the structure of the hair and liberating the DNA thereby avoiding the need to remove the hairs from the tape. Alternatively, the tape will release hairs following solvent treatment or simple manual manipulation.
4 Lack of chemical or physical properties which might adversely affect subsequent DNA analysis

3.3.1 Methods

The following tests were planned

1 Adhesion was to be tested by gently rolling a cylinder covered with the tapes over a standard sample (e.g. a piece of woollen material).
2 Hair would be applied by gentle pressure to samples of the tape which would then be given prolonged exposure to temperatures varying from 30^{o}C to -20^{o}C in both dry and wet conditions, to determine the effect on hair retention.
3 Tape samples would initially be tested for compatibility with the initial stages of the DNA extraction process to determine whether DNA can be liberated from hairs held *in situ*. If this simple approach failed then tape samples would be exposed to a range of laboratory solvents and the proportion of hairs released from the backing recorded. Appropriate means of removing the solvent would be examined prior to starting the DNA extraction process and ease of manual removal noted.

4 If an efficient method of removal was found the hairs would be subjected to standard Qiagen DNA extraction protocols and the yield of hair DNA and a spiked control DNA determined following PCR to check the purity of the extract compared with plucked hair controls.
5 Tapes meeting the desired criteria should then be tested in the field, and DNA extracted using the preferred protocol. Testing authentic samples will reveal problems that might arise through environmental contaminants such as soil and plant material.

3.3.2 Results & Discussion

Jason Reynolds of Durham Wildlife Trust kindly provided samples of the standard double-sided sticky tape used in hair tubes, each sample carried a number of squirrel hairs. The tapes were exposed to a range of solvents including ethanol, methanol, chloroform, xylene and water, none of these were effective in breaking down the adhesive and releasing hairs into solution. Unfortunately samples of this tape in the unused form were not received during the course of the study and without this control material we were unable to carry out the lab trials within the allocated time.

A range of tape manufacturers (including Sellotape & 3M) were contacted regarding the adhesives used and it became apparent that the solvents required to release the hairs would be both expensive and lead to health and safety concerns, including problems with disposal of large quantities of solvent waste. One manufacturer, Industrial Self Adhesives Limited (ISA, www.isatape.co.uk), produced a product 4081R with a permanent adhesive on one side and a peelable adhesive on the other. The formulation of the adhesives originally permitted washing off in warm water but this has now been made more weather resistant. However as the hairs remain on the surface of this tape they are easily available for manual picking into tubes. If a water-based approach using this tape is not successful then manual collection by field workers remains an option. ISA will also provide tapes in custom widths, lengths and formats that may be useful in the production of hairs tubes. As the tape is extremely thin it will be relatively easy to ensure the whole area can be crumpled and immersed in an extraction solution.

Summary

Using published sequences from the GenBank database 27 primers were designed that were capable of amplyfying the mitochondrial cytochrome *b* gene from insectivores, rodents, lagomorphs, carnivora and deer in three overlapping fragments (Appendix 4). These were used on DNA extracted from tissue samples from 29 UK terrestrial mammal species which have been retained as reference samples for the development of future species identification tests (Appendix 6). The cytochrome *b* sequence was determined from all but one (Water Shrew) for which the small clipped hair sample provided insufficient DNA for full sequence analysis. For 25 species comparisons were made between more than one individual from dispersed geographic locations using the samples

acquired by JNCC (Appendix 7) and further comparisons are possible with data obtained from GenBank (summarised in Appendix 5).

The JNCC sequence alignments provide all the data required to identify sequence regions of the cytochrome *b* gene which provide unique species-specific signatures, potential targets for DNA testing can then be compared with data obtained through GenBank representing a wider sampling which extends beyond the UK. Assays developed using the sequence data summarised here can then be validated using the reference tissue samples and DNA extracts.

Delays resulting from Foot & Mouth Disease restrictions and the lack of control samples of adhesive tape from hair tubes limited work on the development of hair sampling methods. Although problems were experienced in extracting sufficient DNA for sequence analysis of hairs from the smallest mammals, the greater sensitivity of alternative methods of assaying DNA variation provide hope that species identification of mammal hairs is now a realistic prospect.

References

Anderson S, Bankier AT, Barrel BG, De Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger R, Schreier PH, Smith AJH, Staden R & Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature, Lond.* **290**: 457-465.

Boore JL (1999) Animal mitochondrial genomes. *Nucl. Acids Res.* **27**: 1767-1780.

Griffiths CS (1997) Correlation of functional domains and rates of nucleotide substitution in cytochrome *b*. *Mol. Phylo. & Evol.* **7**: 352-365

Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX & Wilson AC (1989) Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc. Natl. Acad Sci. USA* **86**: 6196-6200.

Seddon JM, Santucci F, Reeve NJ & Hewitt GM (2001) DNA footprints of European hedgehogs, *Erinaceus europaeus* and *E. concolor*: Pleistocene refugia, postglacial expansion and colonization routes. *Mol. Ecol.* **10:** 2187-2198.

Suzuki H, Tsuchiya K & Takezaki N (2000) A molecular phylogenetic framework for the Ryuku endemic rodents *Tokudaia osimensis* and *Diplothrix legata*. *Mol Phylogenet. Evol.* **15 (1)**: 15-24.

Wetton JH, Tsang CSF, Roney CA & Spriggs AC (2002) An extremely sensitive species-specific ARMS PCR test for the presence of tiger bone DNA. *For. Sci Int* . **126:** 137-144.

Glossary

3' end (terminus) - end of DNA chain which is extended by *Taq* polymerase.

5' end (terminus) - opposite end of DNA chain from 3' end.

Amino Acid - a building block of proteins, occurring in twenty different forms.

Amino terminal - start of a protein chain.

Amplification - process of replicating DNA by PCR (*q.v.*).

Annealing - process by which complementary single stranded DNA fragments interact.

Base - a building block of DNA, occurs in 4 forms known as A, G, C and T. The bases pair, A with T, and C with G between the two strands of the double helix.

bp - (abbreviation of base pair) indicates length of double stranded DNA fragment from the number of complementary bases.

Chromosome - very long molecules of DNA which occur in matched pairs, one inherited from each parent. Exact copies of several different pairs are found in virtually every cell.

CITES - Convention on International Trade in Endangered Species of Wild Fauna & Flora.

Codon - sequence of three bases encoding a particular amino acid.

Complementary - (of bases or sequences) sequences where A is paired with T, and C with G.

Cytochrome *b* - gene with species specific DNA sequence, located within the mitochondrial DNA (*q.v.*).

Denaturation - process by which complementary strands of a DNA fragment are separated, usually by heating.

DNA - Deoxyribonucleic acid, the double stranded helical molecule which encodes genetic information within the sequence of its constituent bases.

Electropherogram - printout obtained from an automatic DNA sequencer.

Extension - process by which complementary bases are added by *Taq* polymerase to the end of a single stranded primer annealed to a template (*q.v.*).

Genome – name of the circular DNA chain of Mitochondrial DNA (*q.v.*).

Hybridization - the process by which a strand of DNA pairs with its complementary sequence in a test sample.

Mitochondria - intracellular structures responsible for generating energy, many are present within each cell.

Mitochondrial DNA - circular DNA chain found with up to 10 copies in each mitochondria, typically 15 - 20,000 bases in length.

Nucleus - a region of the cell which contains the chromosomes.

Polymerase Chain Reaction (PCR) - process for replicating specific DNA regions within the test tube.

Polymorphic - showing variation between individuals.

Primer - short single stranded DNA chain used to initiate DNA replication in PCR or sequencing.

Probe - a DNA sequence which has been fluorescently tagged which will bind to similar sequences of DNA in a test sample.

Product - double stranded DNA made during the PCR reaction.

Sequencing - process by which the order of bases in DNA is determined.

***Taq* polymerase** - the enzyme used in PCR which sequentially adds bases as determined by the template sequence.

Transfer RNAs (tRNA) – a short nucleic acid molecule involved in the translation of genetic information from DNA to protein.

Acknowledgements

We would like to thank all who provided and verified samples for these studies including; Gail Butterill; Mick Canham; J.R. Flowerdew; Aiden Foster; M.L. Hale; Ian Hepburn; Dr. Paul Honess; Andrew Kitchener; Mark Lambert; C.P. Matcham; E. Munro; The National Museum of Scotland; F.E. Ranger; Phil Rudlin; D.A. Stroud; V. Simpson; R.C, Trout; L. Way; James Williams; Michael Woods. With special thanks to Jessa Battersby. We are also very grateful to all those within the FSS and other organisations who provided valuable information and advice.

Appendix 1 – Experimental Methods

A1.1 Sampling Method

Hair and tissue samples were provided by the JNCC for up to three individuals from five taxonomic groups comprising 28 species, see Appendix 6 Sample Table for full list.

A1.2 Extraction

Mitochondrial DNA was extracted from samples using a Qiagen DNAeasy Tissue Kit following the manufacturer's protocol for animal tissue. The number of hairs used varied from one coarse hair of 5.0-6.0cm to 30 fine hairs 0.5-1.0cm in length. Tissue was more consistent with a small piece about 8mm ^{3} being cut into smaller pieces to aid lysis.

Briefly, the protocol entailed mixing the sample with 180µl ATL Buffer (2.5-10% edetic acid, 2.5-10% sodium dodecyl sulphate), 20µl Proteinase K (10mg/ml) and 20µl DTT[1] (0.4M). The sample was then briefly vortexed before being placed in a shaking incubator set at 56°C for a minimum of 30 minutes. Occasional mixing helped to dissipate the sample thereby aiding extraction. Extraction was generally carried out following an overnight incubation to ensure complete digestion.
Once dissolved, 200µl AL buffer (aqueous solution of >25% guanidine hydrochloride) was added, the tube vortexed and then incubated in a water bath at 70°C for 10 minutes. The AL buffer contains a high concentration of guanidinium thiocyanate (GuSCN) which lysed the cells releasing the DNA into solution. Next, 210µl (200µl for tissue) of 96-100% ethanol was added and the lysate vortexed and transferred to a spin column which was centrifuged at 8,000rpm for 1 minute in a Biofuge Pico centrifuge. The GuSCN induced binding of the DNA to the silica gel in the spin column through hydrogen bonds whilst polysaccharides, proteins and other components of the lysate mixture were removed by consecutive wash steps with 500µl of AW1 and AW2 buffer. The first centrifugation step after AW1 addition was at 8,000rpm for 1 minute, after AW2 the spin was 13,000rpm for 3 minutes to remove any residual ethanol. The DNA was then eluted into a clean Eppendorf by the addition of 50µl sterile deionised water. The tube was then left for 1 minute at room temperature before a final centrifugation step at 8,000rpm for 1 minute. The eluate was stored at –20°C until PCR reactions were carried out.

1- DTT is used to break the disulphide bonds in the hair shaft and it is not required for tissue extractions.

A1.3 Primer Design

Primers (excluding flanking sequence primers B1 and *thr*tRNA) were designed based on published sequences from the GenBank database accessed via the National Centre for Biological Information (NCBI) website (www.ncbi.nih.gov).

Sequences found in GenBank were imported into Sequence Navigator v 1.0.1 (Perkin Elmer) and aligned allowing conserved regions within each of the five species groups (Rodents, Carnivores, Lagomorphs, Insectivores and Deer) to be identified. From these layouts four internal primers for each group were designed to produce three short overlapping fragments of about 500bp, the two flanking primers producing a fourth fragment encompassing the full length of the gene (1140-1143 bp).

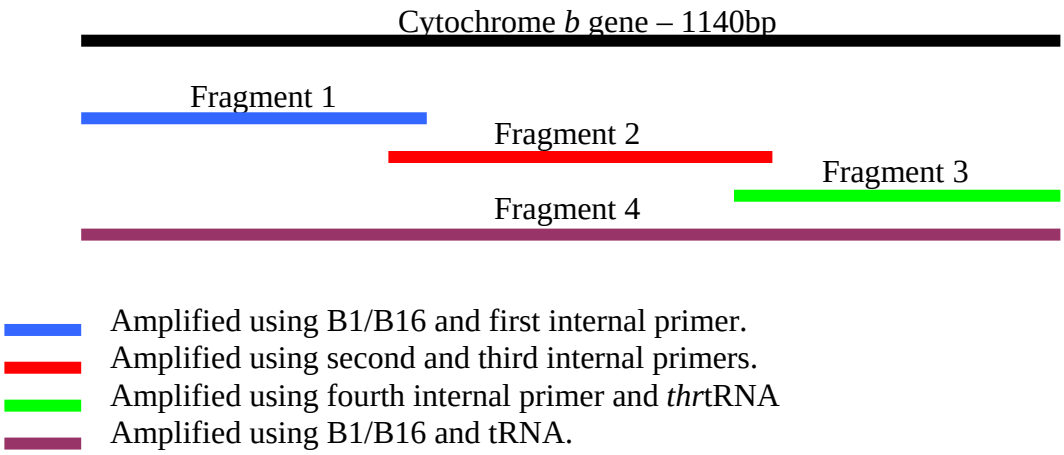

Figure 1 – Diagrammatic representation of the fragments amplified during PCR using four primer pairs.

Forward primers were designed to amplify regions starting about 400bp and 800bp from the beginning of the cytochrome *b* gene, while reverse primers initiated amplification at around 450bp and 850bp to produce around 50bp overlap between fragments. Additional primers were required for Rodents and Carnivores due to the greater variation shown within these groups.
Primer sequences were imported into Primer Express (version 1.0) which calculated the G:C ratio and T_m (melting temperature). A 50% G:C ratio spread evenly across the primer and a T_m around 60°C was aimed for. Primer Express also identified possible secondary structures within each primer and any primer dimer interactions between primer pairs that would reduce the efficiency of the PCR reaction.

A1.4 PCR (Polymerase Chain Reaction)

Each sample underwent four PCR reactions to produce three short fragments and one long fragment covering the full length of the gene (see Appendix 4 for primers). The volume of DNA added varied; 3µl for tissue extracts (made up to 10µl with sterile deionised water (SDW) while 10µl of hair extract was used to ensure adequate amounts of template was present. The template DNA was amplified in a total reaction volume of 50µl.

Each sample contained the following required for amplification:
0.25µM of forward primer and 0.25µM of reverse primer

5µl of GeneAmp 10x PCR Buffer containing 15mM $MgCl_2$ (Applied Biosystems Part Nos:N808-0006/0129).
1µl of 10mM dNTPs
0.5µl BSA (Bovine Serum Albumin)
0.25µl (0.5µl for long fragments) of AmpliTaq Gold DNA Polymerase (5U/µl)

The reaction was performed on a Tetrad Engine Thermocycler (MJ Research Inc.) using the following amplification conditions:

95°C for 20 minutes –Activation of *Taq* Gold DNA Polymerase

35 cycles of:

94°C for 1 minute - Denaturation
50°C for 1 minute - Annealing
72°C for 1 minute - Extension

then:

72°C for 10 minutes and finally holding at 4°C.

A1.5 Verification Gel

Horizontal gel electrophoresis was used to confirm the success of the PCR reactions and ensure negatives were uncontaminated.
Two 30 'node' combs were used to form wells within a 3% NuSieve agarose gel poured into a plastic tray. After a minimum of 30 minutes the combs were removed and the gel placed into a buffer tank where it was covered in 1xTBE ethidium bromide buffer until the level was about 4mm above the surface of the gel.
4µl of each PCR product was mixed with 2µl of stop mix and loaded into the gel. 4µl of Low DNA Mass Ladder was loaded to allow for an estimation of the size and concentration of each PCR product. Electrophoresis was carried out at 120V for at least 30 minutes. On completion gels were placed into an AlphaImager 1220 which used UV light to excite ethidium bromide bound to DNA causing it to fluoresce. Pictures of the gel were taken once optimum conditions for DNA visualisation had been set by adjustment of the light levels and contrast.

A1.6 Purification of PCR Products

Excess primers, dNTPs (deoxyNucleosideTriPhosphates), *Taq* polymerase and salts were removed from the PCR products using QIAquick PCR Purification Kits (Qiagen) or UltraClean PCR Clean-up Kits (MoBio) following manufacturer's protocols. Both involve mixing the PCR reaction with a buffer (PB-Qiagen, SpinBind -MoBio) in a 1 to 5 ratio, transferring the mixture to a spin column and centrifuging at 13,000rpm for 1 minute to bind the DNA to the column. A wash step (PE Buffer-Qiagen, SpinClean-MoBio) and two spins (13,000rpm, 1 minute) removed impurities and any residual ethanol before elution of the DNA in 50µl of SDW.

A1.7 Quantification of PCR Product

To confirm the DNA products had not been lost during purification a second verification gel was run as described earlier. The second pictures were used to determine the approximate concentration of each product to ensure the correct volume of template was used in the sequencing reaction.

A1.8 Big Dye Terminator Sequencing

For every PCR product two sequencing reactions were required, one forward and one reverse. Samples were diluted with SDW to approximately 50ng in 11µl. Each reaction required 1µl of a forward or reverse primer (3.3µM) and 8µl of BigDye Terminator Ready Reaction Mix. This contained everything needed for the reaction including AmpliTaq DNA polymerase, dye terminators and dNTPs. Every sequencing batch included a positive control; 2µl of the kit positive pGEM-sZf(+) (0.2µg/µl) was added to 1µl of it's own primer -21 M13.

The reaction was performed on a Tetrad Engine Thermocycler with a programme that involved 25 cycles of:

96°C for 10 seconds - Denaturation
50°C for 5 seconds - Annealing
60°C for 4 minutes - Extension

Once complete: 4°C forever.

A1.9 Sequencing on the MegaBACE (high through-put)

A1.9.1 Purification of Sequencing Products

Purification of sequencing products was carried out using a 96 well plate format with DyeEx 96 plates (Qiagen). Plates were spun down prior to product addition removing excess hydration buffer to produce a solid gel matrix. Spin plates were placed onto 96 well plates (Sorenson) to collect the purified product which could then be loaded straight onto the MegaBACE. Sequencing products (20µl) were applied to the centre of the gel and centrifuged at 2,500rpm for 3 minutes (Sigma 4-15C). Purified products were stored at 4°C prior to electrophoresis.

A1.9.2 Capillary Electrophoresis

The MegaBACE 1000 was used as the main sequencer. Run parameters were set before the start of each run, typically;

Injection Voltage - 2 or 3kV, Injection Time - 75 seconds,
Run Voltage – 5kV Run Time- 300 minutes.

Run Chemistry –TSII-Version2 Terminator

Before any samples were loaded a tip rinse with deionised water was performed followed by an emptying of the capillaries and injection of 0.7ml fresh Long Read LPA (Linear Polyacrylamide) Matrix. Two 96 well plates (Sorenson) one containing the purified sequencing products (as previously mentioned) and the other containing 200µl LPA Buffer per well were loaded sequentially into the MegaBACE. Each capillary took up DNA from the sample plate by electrokinetic injection followed by the buffer.

A1.10 Sequencing on the 377 (low through-put)

A1.10.1 Purification of Sequencing Products

For small batches purification was carried out using individual DyeEx spin columns. Columns were vortexed to re-suspend the resin and placed into clean Eppendorfs. Sequencing products (20µl) were applied to the centre of the column and centrifuged at 3,000rpm for 3 minutes. Purified products were stored at 4°C prior to electrophoresis.

A1.10.2 Polyacrylamide Gel Electrophoresis

Due to problems with the MegaBACE including dye blobs, compression at the start due to poorly optimised software and poor resolution at the end of the sequence, a few samples were sequenced and run on an automated ABI Prism 377 DNA Sequencer. Sequencing products were ethanol precipitated by addition of 35µl of ice cold 95% ethanol and stored at -20°C for 1 hour, then spun at 13,000rpm for 30 minutes. The supernatant was removed and the tubes left open, but covered, for 30 minutes to air-dry. Once all ethanol had evaporated 3µl Dextran Blue/formamide was added to resuspend the DNA pellet. Samples were stored at 4°C until required.

A pre-prepared 5% Long Ranger gel was cleaned with SDW to remove residual acrylamide from the outside faces of the glass plates, especially from the laser read region, and loaded into a cassette. A 48-lane paper comb was inserted and the plate placed into the sequencer. A clean plate check confirmed that the laser read region was free from fluorescence caused by acrylamide. A pre-run was performed allowing the gel to warm to 51°C while the samples were denatured by heating to 95°C for 2 minutes on a PE 9600 Thermocycler then snap cooled on ice. 0.8µl of each sample was loaded onto the gel and the single speed run (7 hours) was started.

A1.11 Data Analysis

Using the MegaBACE:
Raw data files were analysed using Sequence Analyzer software which used Cimarron 1.53 Slim base caller. The files were converted to ABD (Applied Biosystems Division)

format and imported into ABI Typer (an auto typer created using MakeAutoTyper Version 5.3) which converted the files to a format compatible with Sequence Navigator.

Using 377 Automated Sequencer:
Data was collected using ABI Prism 377 Collection software Version 1.1 (Perkin Elmer Corps.) and analysed using ABI Prism Sequencing Analysis software Version 3.2 (Perkin Elmer Corps.)
In both cases sequences were imported into and aligned using Sequence Navigator Version 1.0.1 layouts.

Equipment

ABI Prism 377 DNA Sequencer - Applied Biosystems, USA
AlphaImager 1220 - Flowgen, USA
Biofuge Pico Centrifuge- Heraeus Instruments Ltd, Essex
Buffer Tank: Wide Mini-Sub Cell - BioRad, CA, USA
4-15C Centrifuge - Sigma Philip Harris, Poole.
Electrophoresis Unit - LKB.GPS 200/400 – Pharmacia, New Jersey, USA
MegaBACE 1000 Analyser Instrument- Molecular Dynamics and Amersham Life Sciences, Little Chalfont
MS1 Minishaker – IKA, Germany
Orbital Incubator SI50 - Stuart Scientific, Belgium
Sequence Navigator version 1.0.1 - Perkin Elmer, CA, USA
96 Well Plates: Sorenson Bioscience Inc. - Dominique, Dutscher
Thermal Paper model K65 HM - Mitsubishi, Germany
Tetrad Engine Thermocycler - MJ Research Inc., USA
9600 Thermocycler - Applied Biosystems, Warrington

Materials

Ampli*Taq* Gold DNA Polymerase - Applied Biosystems, Warrington.
BigDye Terminator Cycle Sequencing Ready Reaction Kit - Applied Biosystems, Warrington.
Bovine Serum Albumin - Life Technologies, Paisley.
DNeasy Tissue Kit - Qiagen Ltd, Crawley.
dNTPs - Roche Diagnostics Ltd. Lewes.
DTT – Sigma, Dorset.
Dye Ex Spin Columns – Qiagen Ltd. Crawley.
Ethanol - Fisher Scientific, Loughbourgh..
LPA Matrix and Buffer - Amersham Pharmacia Biotech, Little Chalfont.
Low DNA Mass Ladder - Invitrogen, Life Technologies, Paisley.
MoBio UltraClean ™ PCR Clean-up Kit - MoBio Laboratories, Inc. - Solana Beach, CA
NuSieve Agarose- Biowhittaker Molecular Applications, USA.
Primers - Oswel DNA Service, Southampton.
Proteinase K – Sigma, Dorset.

Appendix 2

REAGENT PREPARATION

All aqueous solutions should be prepared using deionised water.

FOR DNA EXTRACTION		STERILISE	STORAGE	STORAGE TIME
10%(w/v)SDS	**100ml**	Use sterile water	Room temp.	1 month
10g SDS/100ml sterile deionised water in sterile bottle				
10mg/ml Proteinase K (PK)		Use sterile water	Freezer	3 months
0.1g PK/10ml sterile deionised water.				
Aliquot under sterile conditions in 0.5-1ml volumes.				
0.4M Dithiothreitol (DTT)	**20ml**	Sterile filter	Freezer	3 months
3.08g DTT/20ml sterile deionised water				
Store in 0.5 to 1ml aliquots.				
DO NOT REFREEZE - DISCARD AFTER USE.				
FOR PCR		**STERILISE**	**STORAGE**	**STORAGE TIME**
10mM dNTPs	**1ml**	No	Freezer	6 months
100mM dATP Li-salt pH7	100µl			
100mM dCTP Li-salt pH7	100µl			
100mM dGTP Li-salt pH7	100µl			
100mM dTTP Li-salt pH7	100µl			
Sterile Water	600µl			
Prepare in Amplification set-up area using pipettes and sterile tips. Aliquot to store.				
FOR MINIGEL ELECTROPHORESIS		**STERILISE**	**STORAGE**	**STORAGE TIME**
10xTBE	**5 litres**	No	Room Temp.	3 months
Tris	540g			
Boric Acid	275g			
EDTA disodium salt	41.5g			
Dissolve Tris in approx. 3.5l distilled water. Once dissolved, add boric acid and stir until the boric acid is also dissolved. Add EDTA in small quantities and stir to dissolve. Make up to 5l with water and check pH is 8.3 ± 0.1; if the pH is outside this range, discard the batch.				
Add 250µl of 10mg/ml ethidium bromide.				
1xTBE/ethidium bromide	**5 litres**	No	Room Temp.	3 months
10xTBE	500ml			
10mg/ml ethidium bromide	250µl			
make up to 5l with water				

3% NuSieve gel mix **500ml** NuSieve agarose powder 15g 1xTBE/ethidium bromide q.s. Mix agarose powder and TBE in a 500ml Duran bottle. Loosen the cap and microwave until the mix boils gently, regularly interrupting the heating to swirl the gel mix. Once all of the agarose powder has dissolved, place the mix in an incubator at 50^{o}C; do not use until the mix has cooled to this temperature	No	50^{o}C	3 months
Stop Mix (TBE gels) **50ml** 20% Sucrose 10g 15% Ficoll 7.5g 0.2% Bromophenol blue 0.1g 10mM EDTA pH 8.0 make up to 50ml	No	Fridge	1 year
FOR ACRYLAMIDE GEL ELECTROPHORESIS	**STERILISE**	**STORAGE**	**STORAGE TIME**
10XTBE pH8.3 **1 litre** Tris 108g Boris Acid 55g EDTA disodium salt 8.3g Check pH is 8.3±0.05. If solution fails pH test, discard and make fresh solution.	No	Room temp	3 months
5% Long Ranger gel mix **400ml** Urea 144g Long Ranger stock solution 40ml (FMC Bio Products Catalogue No. 50610) deionised water 150ml Combine the above in a 500ml Duran bottle and stir (with heat if necessary) until the urea is dissolved. Add 40ml filtered 10xTBE and then adjust the volume to 400ml with deionised water		Room temp	3 weeks
formamide/dextran blue **10ml** deionised formamide 10ml dextran blue (MW 2,000,000) 0.2g	No	4^{o}C	12 months

Appendix 3

Vertebrate mitochondrial genetic code

```
  AAs  = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = ----------------------------------MMMM---------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Differences from the Standard Code:

	mtDNA		Standard	
AGA	Ter	*	Arg	R
AGG	Ter	*	Arg	R
AUA	Met	M	Ile	I
UGA	Trp	W	Ter	*

Amino acid	Three letters	One letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Appendix 4

Primers used in the study

Groups	Primer	Primer sequence	Fragment of cytochrome *b* amplified
Insectivores	I511R	5' TCGGGTGAGyGTGGCTTTGTC	1st fragment
Insectivores	I425F	5' TGAGGACAAATATCATTTTGAGG	2nd fragment
Insectivores	I937R	5' TGGGCGGAATATTATyCTTCG	2nd fragment
Insectivores	I878F	5' TAAATTTGGAGGCGTTCTAGC	3rd fragment
Rodents	R493R	5' TCTACTGAGAAGCCCCCTCA	1st fragment
Rodents	R425F	5' TGAGGACAAATATCATTCTGAGG	2nd fragment
Rodents	R934R (10)[1]	5' GGCGGAATyTTAGGCTTCGTTG	2nd fragment
Rodents	R934R (3)[2]	5' GTCGGAAAGTGAGTCCTCGTTG	2nd fragment
Rodents	R425BF[3]	5' TGAGGACAGATATCATTTTGAGG	2nd fragment
Rodents	R1024R[3]	5' AATAATAAAGGGATATTCAACTGG	2nd fragment
Rodents	R845F	5' TATTTGCCTACGCCATyCTACG	3rd fragment
Carnivores	C493R	5' GTCTACTGAGAATCCGCCTCA	1st fragment
Carnivores	C433F	5' ATATCCTTTTGAGGrGCAACCG	2nd fragment
Carnivores	C934R	5' GGCGGAATATTATGCTTCGTTG	2nd fragment
Carnivores	C922R [4]	5' TATCCCACGTTGTTTGGAGGTGTG	2nd fragment
Carnivores	C905F	5' TAGTCTTCTCCATCCTAGTCCTAGC	3rd fragment
Lagomorphs	L527R	5' AAGTGGAAAGCGAAGAATCGGG	1st fragment
Lagomorphs	L406F	5' AGGCTATGTyCTCCATGAG	2nd fragment
Lagomorphs	L950R	5' ACTTGGCTAATGGGTCGGA	2nd fragment
Lagomorphs	L836F	5' TGATATTTTCTATTTGCCTACGC	3rd fragment
Deer	D493R	5' TTTATCTACTGAAAAGCCTCCTCA	1st fragment
Deer	D409F	5' CGTAGGATACGTyCTACCATGAGGAC	2nd fragment
Deer	D964R	5' CTAGGATTCAGAATAGGCATTG	2nd fragment
Deer	D848F	5' GCATACGCAATCCTACGATC	3rd fragment
All	*thr*tRNA	5' CTTCATTTTTGGTTTACAAGACCA	4th Fragment
All	B1	5' CCAATGATATGAAAAACCATCGTT	4th Fragment
Various	B16	5′ CTAATGACATGAAAAATCATCGTTGT	4th Fragment

[1] Used with rodents with following exceptions
[2] Used with *Arvicola & Microtus sp.*
[3] Used with *S. carolensis, S. vulgaris & A. terrestris*
[4] Used with *V.vulpes*

Primer names indicate the species grouping (I, R, C, L or D), the location of the 3'base with respect to the reference cytochrome *b* sequence and the direction of amplification (F= forward, R= reverse).

Appendix 5

GenBank Sequence Alignments

The following alignments of cytochrome *b* sequences retrieved from the GenBank database are broken down into the five taxonomic groupings on which primer designs were based: Insectivores, Rodents, Carnivores, Deer and Lagomorphs.

For each species an edited consensus sequence was derived from all relevant sequences held on GenBank in February 2001. The authenticity of these sequences could not be checked at the time and may include some examples of nuclear copies and many examples (if not the vast majority) from non-UK populations. These alignments were used to design the primers shown in Appendix 4.

For each grouping

1 an ambiguity sequence is shown highlighting bases that were variable between species.
2 a consensus sequence shows the commonest base in all species within the grouping.
3 an edited consensus summarises all the bases found within and between the species within the grouping in the form of IUB codes for base mixtures
4 an ambiguity sequence highlights all the bases that are variable within the grouping

Primers were targeted primarily at sequences that varied little within the grouping as demonstrated by a region clear of *s within the ambiguity sequence.

Appendix 6

Samples Table

Group	Common name	Scientific name	Individual	Sample Type	Date Received	Origin
Insectivores	Hedgehog	*Erinaceus europaeus*	H1	Tissue	10.05.01	East Lothian
			H2	Hair & Tissue	31.08.01	Lincolnshire
Insectivores	Mole	*Talpa europaea*	M1	Tissue	10.05.01	Dumfrieshire
			M2	Hair	10.05.01	Flowerdew
			M6	Hair	10.05.01	Flowerdew
Insectivores	Pigmy shrew	*Sorex minutus*	PS	Hair	01.06.01	Somerset
Insectivores	Common shrew	*Sorex araneus*	CS	Hair	22.01.01	Staffordshire
			CS2	Tissue	10.05.01	Dumfrieshire
			CS3	Tissue	04.07.01	Lincolnshire
			CS4	Hair	01.06.01	Somerset
Insectivores	Water shrew	*Neomys fodiens*	WS	Hair	22.01.01	Staffordshire
Rodents	Dormouse/Common	*Muscardinus avellanarius*	D	Hair	01.06.01	Somerset
			CD	Hair	12.09.01	Kidderminster
			CD2	Tissue	12.09.01	Surrey
Rodents	Bank vole	*Clethrionomys glareolus*	BV	Hair	24.01.01	Staffordshire
			BV2	Tissue	01.06.01	Somerset
Rodents	Water vole	*Arvicola terrestris*	WV	Tissue	10.05.01	Aberdeenshire
			WV2	Tissue	10.05.01	Aberdeen City
			WV3	Hair	17.05.01	Cheshire
Rodents	Field/Short-tailed vole	*Microtus agrestis*	FV	Hair	24.01.01	Staffordshire
			FV2	Tissue	10.05.01	Pathead-Fife/Midlothian
Rodents	Yellow-necked Mouse	*Apodemus flavicollis*	YM1	Hair & Tissue	12.09.01	Surrey
Rodents	Wood/Long tailed Mouse	*Apodemus sylvaticus*	WM	Hair	22.01.01	Staffordshire
			WM2	Tissue	01.06.01	Somerset
			WM3	Tissue	03.07.01	?
Rodents	House Mouse	*Mus musculus*	HM	Tissue	01.02.01	Staffordshire
			HM2	Hair	01.06.01	N. Yorkshire
Rodents	Brown/Common Rat	*Rattus norvegicus*	BR	Hair & Tissue	10.05.01	Midlothian
			BR2	Hair	01.06.01	Yorshire
			BR3	Hair	03.07.01	Northamptonshire
Rodents	Red Squirrel	*Sciurus vulgaris*	RS	Hair	25.04.01	Scottish Borders
			RS2	Hair	25.04.01	Northumbria
			RS3	Tissue	10.05.01	S. Cumbria
Rodents	Grey Squirrel	*Sciurus carolinensis*	GS	Hair & Tissue	23.01.01	Staffordshire
			GS2	Tissue	25.04.01	Northumbria
			GS3	Hair	25.04.01	Northumbria
Lagomorphs	Rabbit	*Oryctolagus cuniculus*	R	Hair	12.02.01	Staffordshire
			R2	Tissue	05.07.01	Northamptonshire
Lagomorphs	Brown Hare	*Lepus capensis*	BH	Tissue	10.05.01	Dumfries & Galloway
			BH2	Tissue	10.05.01	National Museum of Scotland
Lagomorphs	Blue/Mountain Hare	*Lepus timidus*	MH	Tissue	10.05.01	Perth & Kinross
			MH2	Tissue	10.05.01	Aberdeenshire
Carnivores	Stoat	*Mustela erminea*	S	Hair & Tissue	10.05.01	Scottish Borders
			S2	Tissue	10.05.01	East Lothian
Carnivores	Weasel	*Mustela nivalis*	W	Hair & Tissue	10.05.01	Invernesshire
			W2	Tissue	10.05.01	East Lothian
Carnivores	Polecat/ferret	*Mustela putorius*	P	Tissue	10.05.01	Gwynedd

			P2	Hair	09.04.01	Oxfordshire
			FB	Tissue	10.05.01	Aberdeenshire
			FS	Tissue	10.05.01	Shetland
Carnivores	American Mink	*Mustela vison*	MK	Tissue	10.05.01	Argyll
			MK2	Hair & Tissue	13.08.01	Norfolk
Carnivores	Pine Marten	*Martes martes*	PM	Tissue	10.05.01	Highland
			PM2	Tissue	10.05.01	Scotland-ex Perth Museum
Carnivores	Otter	*Lutra lutra*	O	Tissue	10.05.01	Lancashire
			O2	Tissue	10.05.01	Argyll & Bute
Carnivores	Fox	*Vulpes vulpes*	F	Hair	17.01.01	Staffordshire
			F2	Tissue	10.05.01	City of Edingburgh
			F3	Tissue	10.05.01	Oman, Middle East
			F4	Tissue	21.12.01	Morayshire
Carnivores	Badger	*Meles meles*	B	Hair	16.01.01	Staffordshire
			B2	Tissue	10.05.01	Dumfries & Galloway
			B3	Tissue	10.05.01	Dumfries & Galloway
Deer	Muntjac	*Muntiaacus reevesi*	MJ	Hair & Tissue	12.09.01	Surrey
Deer	Fallow Deer	*Dama dama*	FD	Hair	13.02.01	Staffordshire
			FD2	Tissue	12.09.01	Worcestershire
Deer	Red Deer	*Cervus elaphus*	Red	Tissue	21.12.01	Moray
Deer	Roe Deer	*Capreolus capreolus*	RD	Hair & Tissue	12.09.01	Hampshire
			RD2	Tissue	21.12.01	Moray

Appendix 7

JNCC Sequence Alignments

This alignment summarises the sequences derived from JNCC samples. For each species a unanimity sequence is shown in which those bases which did not vary between JNCC samples are shown as A, C, G or T, whilst bases which varied between JNCC samples are indicated by a *. As with the GenBank alignments an ambiguity sequence is shown for the larger taxonomic groupings.

Individual species alignments including EPGs are available in Sequence Navigator format from FSS on CD.

NB In compiling the sample source data in Appendix 6 it became apparent that one fox sample (F3) supplied by the National Museum of Scotland originated from Oman in the Middle East, this single sample which provided data between bases 434 and 846 gave rise to polymorphism as follows;
477 T-C, 552 T-C, 582 G-A, 585 T-C, 657 G-A, 696 T-C, 723 A-G, 783 T-C813 A-G, 819 T-C where the first base is the base found in all other GenBank and JNCC sequences. The inclusion of this fragment of sequence from a non-UK individual resulted in 10 additional polymorphic sites within 432 bases.