Genetic diversity of British water voles (*Arvicola amphibius*) and phylogenetics of the rodent subfamily Arvicolinae.

A thesis to the University of Kent for the degree of:

MSc by Research in Genetics

2020

Corey Kirkland

School of Biosciences
Declaration:

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institutions of learning.
Acknowledgements:

I would firstly like to say a huge thanks to my supervisor, Dr Marta Farré-Belmonte, for accepting me onto this MSc-R project and for her support throughout this project. She has provided me with technical support both in the laboratory and in bioinformatics and provided me with an amazing opportunity to develop as a research scientist. These skills will be incredibly useful as I hope to continue my career in research. It has been a pleasure to be a part of her research group at the University of Kent. I would also like to acknowledge all the members of the group – Cristina, Frances, and visiting researcher Lucía – who have been wonderfully supportive and who I have really enjoyed working alongside, as well as members of the ERGO group.

Also, a big thanks to Wildwood Trust and Hazel who have allowed me to sample water voles at their wildlife park. Plus, Kent Mammal Group for providing a research grant for consumables.

Lastly, I would like to thank Matthew, my friends and my family for their love and support.
Contents:

Declaration: ................................................................................................................................... 2
Acknowledgements: ...................................................................................................................... 3
List of Figures: ................................................................................................................................ 7
List of Tables: ................................................................................................................................. 9
Abbreviations: ............................................................................................................................. 10
Abstract: ...................................................................................................................................... 11
1. Introduction: ........................................................................................................................ 12
   1.1. Population Genetics of British Water Voles: ............................................................... 12
       The European Water Vole: ............................................................................................... 13
       Population Genetics of British Water Voles: ................................................................ 14
       Genetic Sampling: ............................................................................................................ 16
   1.2. Phylogenetics of Arvicolinae: ...................................................................................... 18
   1.3. Hypothesis: .................................................................................................................. 21
   1.4. Research Aims: ............................................................................................................ 21
2. Materials and Methods: ...................................................................................................... 22
   2.1. Optimisation of DNA Extraction: ................................................................................. 22
       Sample Collection: ............................................................................................................ 22
       DNA Extraction of Tissue Samples: ................................................................................ 22
       DNA Extraction of Hair Samples: ..................................................................................... 24
       DNA Extraction of Faecal Samples: ................................................................................ 24
       DNA Extraction of Additional Sample Types: .................................................................. 25
   2.2. Amplification and Sequencing of Mitochondrial Control Region: ................................ 25
       Amplification of Marker: ................................................................................................. 25
       Sequencing: .................................................................................................................... 25
   2.3. Population Genetics of the Water Vole: ...................................................................... 26
       Multiple Sequence Alignments (MSA): .......................................................................... 26
       Haplotype Networks: ....................................................................................................... 26
Phylogenetic Tree Construction: ................................................................. 26
Population Genetics Calculations: ............................................................. 26
Population Structure: .................................................................................. 27
2.4. Phylogenetics of Arvicolinae: ............................................................... 27
Mitochondrial and Nuclear Markers: ........................................................... 27
Phylogenetic Tree Construction: ................................................................. 27
3. Results: .................................................................................................... 28
3.1. Optimisation of DNA Extraction: ............................................................ 28
DNA Extraction of Tissue Samples: ............................................................. 28
DNA Extraction of Hair Samples: ................................................................. 30
DNA Extraction of Faecal Samples: ............................................................. 31
DNA Extraction of Additional Sample Types: .............................................. 33
Comparison of Sample Types: ...................................................................... 33
3.2. Amplification and Sequencing of the Mitochondrial Control Region: ...... 34
3.3. Population Genetics of the Water Vole: .................................................... 35
Individuals from Wildwood Trust: ............................................................... 35
Comparison of Captive and Natural Populations: ........................................ 36
3.4. Arvicolinae Phylogenetics: ................................................................. 42
Mitochondrial Genomes: ........................................................................... 42
Mitochondrial Marker: ............................................................................... 44
Nuclear Markers: ....................................................................................... 45
Supermatrix Approach: ............................................................................... 47
4. Discussion: .............................................................................................. 49
4.1. DNA Extraction and Amplification: ......................................................... 49
4.2. Population Genetics of British Water Voles: .......................................... 51
4.3. Phylogenetics of Arvicolinae: ............................................................... 53
Phylogenies of the Genera of Arvicolinae: .................................................... 53
Comparison of Phylogenetic Approaches: .................................................. 56
4.4. Future Work: ....................................................................................... 57
5. Conclusion: .......................................................................................................................... 58
References: .................................................................................................................................. 60
Appendix: ..................................................................................................................................... 66
List of Figures:

Figure 1: (a) European water vole (*Arvicola amphibius*) in riparian habitat, (b) distribution in Britain (Mathews *et al.*, 2018) and (c) Wildwood Trust logo, Canterbury.............................................. 13

Figure 2: Haplotype network of British water voles showing England/Wales and Scotland haplogroups (Piertney *et al.*, 2005). Each pie chart represents a different haplotype. The size of the pie chart is proportional to the number of individuals in each haplotype. Dotted lines between haplotypes show the number of mutational steps between two haplotypes. The circle filled with dots shows the England/Wales haplogroup, whilst the circle with lines shows the Scotland haplogroup............................................................... 15

Figure 3: South East of England haplotypes (Baker *et al.*, 2020). Colours represent the 15 haplotypes. The pie charts represent the proportion of a given haplotype at each population sampled................................................................. 16

Figure 4: Simplified cladograms from (a) the proposed phylogeny of Arvicoline by Robovský, Řičánková and Zrzavý, 2008 and (b) the phylogenetic tree of Arvicolinae by Steppan and Schenk, 2017................................................................. 20

Figure 5: Gel electrophoresis of all water vole tail tissue samples. Lane ‘L’ shows the 10kb ladder, and lanes 1-4 are the corresponding DNA samples. 1% agarose, 70V, ~60 mins. ........ 30

Figure 6: Comparison of the median DNA yields (ng) for each sample type. Calculated from DNA yields from samples using the final optimised protocols.............................................. 34

Figure 7: Gel electrophoresis of the PCR samples. Samples 5, 12, and 18 were repeated........ 34

Figure 8: Analysis of Wildwood Trust water voles using the mitochondrial control region. (a) Haplotype network of Wildwood Trust samples. Each pie chart represents a haplotype and each colour represents a different sample. The dotted lines are the number of mutational steps between haplotype sequences. (b) ML phylogenetic tree of Wildwood Trust samples. Bootstrap scores are shown for each node. The tree is rooted on *Arvicola sapidus*. ................................. 36

Figure 9: Captive and wild water vole samples from the South East and East of England. Wildwood Trust sequences are individual sequences, whilst Baker 2015 sequences are haplotype sequences................................................................. 37
Figure 10: Haplotype network using DNA sequences of the mitochondrial control region for captive and natural water voles. Each pie chart represents a haplotype and the size represents the number of individuals with the haplotype. The dotted lines are the number of mutational steps between haplotypes. ................................................................. 38

Figure 11: ML phylogenetic tree containing all samples. Taxon colours represent the age or location of sample. The tree is rooted on Arvicola sapidus. Taxon “Wildwood Trust*” contains multiple Wildwood Trust samples. Only bootstrap scores greater than 50% are shown........ 40

Figure 12: Structure analysis of mtDNA for all sequences (K=2). (a) Run 3/5 for K=2. Red and green subpopulations. (b) Delta K: K=1 to K=6 with 5 runs each. K=2 selected. (c) L(K): K=1 to K=6 with 5 runs each. ......................................................................................................................... 42

Figure 13: Maximum-likelihood phylogenetic tree of Arvicolinae mitochondrial genomes. Bootstrap scores above 50% are shown. ........................................................................................................ 43

Figure 14: Maximum-likelihood trees of (a) mitogenomes and (b) mitochondrial Cytb sequences. Rooted on Mus musculus. Bootstrap scores for Cytb ML phylogenetic tree can be found in the Appendix................................................................. 44

Figure 15: Maximum-likelihood phylogenetic tree of Arvicolinae using (a) GHR and (b) IRBP nuclear markers. Both trees are rooted on Mus musculus. Bootstrap scores can be found in the Appendix........................................................................................................ 46

Figure 16: Maximum-likelihood supermatrix for Cytb, GHR, and IRBP genes. Rooted on Mus musculus. Only bootstrap scores over 50% are shown................................................................. 48
List of Tables:

Table 1: Wildwood Trust samples. ........................................................................................................ 22

Table 2: Optimisation of DNA extraction lysis buffers and protocols for water vole tissue samples. ................................................................................................................................. 29

Table 3: DNA extraction of all water vole tissue samples. Results for samples 1 and 2 from the previous table. Samples 3 and 4 using the final optimised protocol (‘Phenol-Chloroform 3’). 29

Table 4: Optimisation of DNA extraction protocols for hair samples. .............................................. 31

Table 5: DNA extraction of hair samples............................................................................................. 31

Table 6: Comparison of different DNA extraction kits for faecal samples........................................ 32

Table 7: DNA concentrations and purity of extracted faecal samples using ‘Qiagen QIAamp DNA Stool Mini Kit’ (A) and ‘Qiagen QIAamp PowerFecal DNA Kit’ (B) before and after ethanol precipitation. ...................................................................................................................... 32

Table 8: Extraction of DNA from waterbuck cell culture................................................................. 33

Table 9: Extraction of DNA from cattle buccal swabs........................................................................ 33

Table 10: European water vole (Arvicola amphibius) and bank vole (Myodes glareolus) population genetics. Number of sequences (n), alignment length in base pairs (bp), number of haplotypes (Hap. No.), haplotype diversity (Hap. Div.), variance (Var.), nucleotide diversity (π), Tajima’s D (D), P-value for a normal distribution (P-value Norm.), and P-value for a beta distribution (P-value Beta). Arvicola amphibius sequences from: (1) Baker, 2015, (2) Piertney et al., 2005, and (3) only modern samples from Brace et al., 2016. Myodes glareolus sequences were from Filipi et al., 2015 and Marková et al., 2020................................. 41
### Abbreviations:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>AMOVA</td>
<td>Analysis of Molecular Variance</td>
</tr>
<tr>
<td>BP</td>
<td>Before Present</td>
</tr>
<tr>
<td>Cytb</td>
<td>Cytochrome b</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic Acid</td>
</tr>
<tr>
<td>GHR</td>
<td>Growth Hormone Receptor</td>
</tr>
<tr>
<td>GTR</td>
<td>Generalised Time Reversible</td>
</tr>
<tr>
<td>HMW</td>
<td>High Molecular Weight</td>
</tr>
<tr>
<td>IRBP</td>
<td>Interphotoreceptor Retinoid-Binding Protein</td>
</tr>
<tr>
<td>LGM</td>
<td>Last Glacial Maximum</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum-Likelihood</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple Sequence Alignment</td>
</tr>
<tr>
<td>mtDNA</td>
<td>Mitochondrial DNA</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate-Buffered Saline</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative Polymerase Chain Reaction</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium Dodecyl Sulfate</td>
</tr>
<tr>
<td>STR</td>
<td>Short Tandem Repeat</td>
</tr>
<tr>
<td>YD</td>
<td>Younger Dryas</td>
</tr>
</tbody>
</table>
Abstract:

The European water vole (*Arvicola amphibius*) has experienced rapid decline in Britain, resulting in numbers declining by approximately 90% in the last century, making them a conservation priority. Its phylogeny within the subfamily Arvicolinae, and its relationship to other taxa, remains debated. Additionally, the impact of captive breeding programs on the genetic diversity of water voles is unknown.

We firstly optimise DNA extraction protocols for tail tissue, hair, faeces, buccal swabs, and cell culture to achieve high DNA yields and purity. We then sequence the mitochondrial control region for 17 captive water voles at Wildwood Trust using tissue, hair, and faecal samples collected at the park to assess their genetic diversity and population structure. Lastly, we use mitochondrial genomes and individual phylogenetic markers to construct the phylogeny of the subfamily Arvicolinae.

Our study provides protocols for the extraction of DNA achieving high yields and improved purity, with cell culture resulting in the highest median yield, followed by buccal swabs and tail tissue. Our results reveal considerable genetic diversity in the mitochondrial control region of captive water voles sampled at Wildwood Trust, with relatively high haplotype diversity. Similar haplotype diversity was also seen in natural populations in Britain. Captive water voles clustered closely with South East of England haplotypes, and were found within the English/Welsh clade. Additionally, we provided support for the phylogeny of several genera of Arvicolinae. Using mitogenomes provided the most resolved phylogenetic tree when compared with other genetic markers and approaches but lacked sequences for all genera to fully resolve phylogeny.

The mitochondrial genome provides a useful marker to study both the conservation and population genetics of water voles, as well as the phylogenetics of the subfamily Arvicolinae. Our study provides support for the breeding program at Wildwood Trust and provides a framework for future conservation studies.

Word Count: 299
Introduction:

1.1. Population Genetics of British Water Voles:
Population genetics is a field which examines the genetic variation within populations and is becoming increasingly useful to study declining populations and halt the decline of species. IUCN recognises the importance of conserving genetic diversity as a key conservation priority (McNeely et al., 1990). Genetic diversity can be defined as the variation in the amount of genetic information within and among individuals of a population, a species, an assemblage, or a community (defined by the United Nations in 1992). It enables populations to evolve and adapt to environmental change. It can be measured by polymorphisms such as single-nucleotide polymorphisms (SNPs), average heterozygosity, and allelic diversity.

Loss of genetic diversity is associated with small or declining populations and can lead to inbreeding, reduction in reproductive fitness, and inbreeding depression. Inbreeding depression reduces heterozygosity and increases homozygosity, increasing the number of alleles with deleterious effects in the population. Threatened species often have lower heterozygosity and it has been proven that genetic factors affect species before extinction does (Spielman, Brook and Frankham, 2004).

This process is known as the extinction vortex. Small populations undergo inbreeding and random genetic drift, which leads to a loss of genetic diversity. This reduces an individual’s fitness and a populations ability to adapt to changing environmental conditions, leading to lower rates of reproduction and higher mortality. The process continues with populations becoming smaller until they then become extinct.

Captive breeding programs and reintroductions (ex-situ conservation) can be a useful management tool to halt the decline of individuals in small populations, but only when the
maintenance of genetic diversity is considered. It is vital to reduce inbreeding during these programs, such as using pedigrees that minimise mean kinship. Outbreeding depression and genetic swamping can also occur in captive breeding programs and reintroductions, as local distinct genetic features and genetic variability can be lost due to the integration of alien gene pools, decreasing evolutionary potential (Laikre et al., 2010; Canu et al., 2013).

Genetic markers, such as nuclear and mitochondrial DNA (mtDNA) are extensively used to determine the genetic diversity of natural and captive animal populations. Mitochondria are found in abundance throughout cells, resulting in thousands of copies of mitochondrial DNA in each cell. In mammals, the mitochondrial genome contains 37 genes. It is inherited maternally, thus offspring inherit a single mtDNA genotype from their mother. The mitogenome is haploid, unlike the nuclear genome, and there is generally no recombination of DNA.

The European Water Vole:

The European or northern water vole (Arvicola amphibius) is a species of rodent in the subfamily Arvicolinae, a group which contains voles, lemmings, and muskrats. It has a wide distribution throughout Europe and Asia, with the overall population trend being stable and the species listed as ‘Least Concern’ (Batsaikhan et al., 2016). The mammal is mostly found in riparian zones, in habitat alongside rivers or streams.

Figure 1: (a) European water vole (Arvicola amphibius) in riparian habitat, (b) distribution in Britain (Mathews et al., 2018) and (c) Wildwood Trust logo, Canterbury.
In Britain, populations of water voles are rapidly declining, faster than any other British mammal in the last century (e.g. Strachan and Jefferies, 1993; Strachan, 2004). Regionally, water voles are classified by the ‘IUCN Red List’ as ‘Endangered’ in England, ‘Near Threatened’ in Scotland, and ‘Critical’ in Wales (Mathews et al., 2018). The main cause of their decline has been increased predation by the invasive American mink (*Neovison vison*), with habitat loss and pollution of watercourses also having a negative effect on populations (Jefferies, Morris and Mulleneux, 1989; Barreto et al., 1998). Therefore, water voles are a conservation priority in Britain, with several projects seeking to increase numbers and manage suitable habitat. Wildwood Trust near Canterbury UK have a conservation program offering mitigation services, health screening, captive breeding, and reintroductions of British water voles.

*Population Genetics of British Water Voles:*

The phylogeographic structure of European water voles, as with the majority of northern hemisphere extant species, has been determined by glaciation events (Hewitt, 1996, 1999; Bernatchez and Wilson, 1998). The Younger Dryas (YD), between 12,800 to 11,500 before present (BP), was a period of retreat and re-advance of glaciers (Walker et al., 2012). Northern Europe was fully glaciated and central Europe was mostly permafrost during this period [4]. Regions in southern Europe (Iberia, Italy, and the Balkans) were ice and permafrost free, creating pocket refugia for populations, known as a refugial peninsula (Hewitt, 1999). Populations then recolonised north with the retreat of the glaciers.

The first study using molecular markers to study European water vole phylogeography used the mitochondrial cytochrome b (*Cytb*) gene (summarised in Taberlet et al., 1998). The study found that samples were from three lineages: (1) eastern and central Europe, the Balkans, and Fennoscandia; (2) Italy; and (3) France and Spain. The small number of English samples used were clustered in the first lineage.

Later studies used the mitochondrial control region to study the evolutionary history of British water voles. Piertney et al., 2005 sampled 62 individuals from 57 locations, both from extant individuals and museum samples less than 75 years old. The study found a major division between English/Welsh and Scottish water voles, with individuals from the two regions found in separate haplogroups and phylogenetic clades (Figure 2). A minimum of 16 mutational steps was found between the two haplogroups, indicating considerable mitochondrial divergence. Previous studies have debated this distinction, with some stating that Scottish water voles morphologically have darker hair than English and Welsh water voles (Miller, 1912) and differences in the size of their tail and hind feet. Whilst more recent studies have found no considerable differences morphologically between the two water voles (Corbet et al., 1970; Telfer et al., 2003).
Figure 2: Haplotype network of British water voles showing England/Wales and Scotland haplogroups (Piertney et al., 2005). Each pie chart represents a different haplotype. The size of the pie chart is proportional to the number of individuals in each haplotype. Dotted lines between haplotypes show the number of mutational steps between two haplotypes. The circle filled with dots shows the England/Wales haplogroup, whilst the circle with lines shows the Scotland haplogroup.

Another study used ancient museum specimens from Britain and mainland Europe, both before the last glacial maximum (Pleistocene period) and following the YD (Holocene period), to understand the colonisation history of British water voles (Brace et al., 2016). The study supported the hypothesis that there were two colonisation events by water voles into Britain. The first colonisation occurred before the last glacial maximum (LGM) throughout England. A second colonisation event then occurred after the Pleistocene, displacing the first colonisers, while the newer colonisers remained in England throughout the Holocene until the present day. The first colonisers were displaced north into Scotland, creating the two modern evolutionary distinct groups.

Studies have also looked at the genetic diversity of natural water vole populations, using both the mitochondrial control region and microsatellite markers. Populations within the two locations have been shown to have significant genetic structure when using mtDNA (Piertney et al., 2005). Analysis of Molecular Variance (AMOVA) revealed that haplotype frequencies between regions and between populations within regions was a significant proportion of the observed haplotype frequency. The remaining haplotype frequency was between the two major divisions. Several studies looking at the phylogeography of water voles in the South East and East of England found significant genetic diversity (Baker, 2015; Baker et al., 2020; Figure 3). Using mtDNA alone uncovered substantial genetic structure between watershed populations,
whilst finer scale structure between populations within watersheds was found when using both mtDNA and microsatellites (Baker et al., 2020). Mitochondrial DNA also showed high levels of haplotype diversity among natural populations.

Figure 3: South East of England haplotypes (Baker et al., 2020). Colours represent the 15 haplotypes. The pie charts represent the proportion of a given haplotype at each population sampled.

Genetic Sampling:

In order to study the genetic diversity of animal populations, we need to obtain DNA from several individuals. Non-invasive genetic sampling is a method of collecting DNA that is left behind by animals, without disturbing them (Taberlet, Waits and Luikart, 1999). Non-destructive sampling is more invasive and can include capturing the animal for a short period of time to obtain the sample, but less invasive than destructive sampling, where the animal is killed for the sample. Non-invasive sampling is becoming more widely used in population genetic studies, with the development of DNA extraction, amplification, and sequencing technologies. It can include sampling hair, faeces, urine, feathers, shed skin, saliva, and eggshells that are left behind by animals (Waits, Lisette and Paetkau, 2005).

One aim of genetic sampling is to obtain sufficient quantities of high-quality DNA. Using non-invasive sampling frequently results in lower quantities of DNA and DNA samples that contain contaminants causing lower polymerase chain reaction (PCR) success rates. When using nuclear loci this can result in genotyping errors such as allelic dropout, where only one of the two alleles present at a heterozygous locus is amplified (Taberlet, Waits and Luikart, 1999). This can arise when concentrations of the DNA template are below 0.05ng/10ml (Gagneux, Boesch and
Woodruff, 1997). Mitochondrial DNA is more reliable when it comes to PCR amplification, as it is haploid, so avoids allelic dropout, and is in larger quantities in non-invasive samples than nuclear DNA.

DNA extraction of non-invasive samples has improved greatly, resulting in higher DNA concentrations and improved DNA quality. There are many different approaches to extracting DNA, but all contain the same main steps of lysis with an extraction buffer, precipitation with alcohol, then resuspension in water or an elution buffer. The extraction buffer contains salts (such as NaCl and Tris-HCl) that stabilises the pH, protecting the negatively charged phosphate groups on the DNA backbone. Detergents, such as sodium dodecyl sulfate (SDS), dissolve the lipid membranes of the cell, releasing the contents of the cell and nucleus. Chelating agents (e.g., ethylenediaminetetraacetic acid (EDTA)) reduce protease or DNase activity. A proteinase (e.g., proteinase K) can also be added to the extraction buffer which enzymatically breaks down proteins, that would otherwise degrade DNA. The concentrations of each of these components differ between extraction protocols and can be optimised for different sample types.

Traditionally, phenol-chloroform was the preferred method for extracting DNA from samples. However, this method uses extremely hazardous chemicals, and it requires significant bench time compared to DNA extraction kits (Schiebelhut et al., 2017). Phenol-chloroform-isoamyl alcohol separates lipids and cellular debris in the solvent phase and DNA in the aqueous phase. Between the two phases is a layer which contains protein. The aqueous phase containing DNA can then be precipitated to increase DNA concentration and DNA purity, using ethanol or isopropanol. Phenol-chloroform DNA extraction consistently produces the highest DNA yield and purity, when compared to other extraction methods (Schiebelhut et al., 2017).

Silica membrane-based extraction kits require less harmful chemicals and quicker extraction times than the more laborious phenol-chloroform extraction (Schiebelhut et al., 2017). Spin columns contain silica beads which cause DNA to bind to them when high concentrations of chaotropic salt are passed through the column. Contaminants also pass through the column during multiple washes and centrifugations. DNA is then eluted into water or a buffer. This method is widely used in conservation and population genetics studies but produces lower DNA concentrations and lower DNA purity. Studies have found that mitochondrial DNA is lost from silica membrane-based kits at a higher rate than nuclear DNA (Guo et al., 2009).

Other extraction protocols have been developed, such as an extraction method for mouse tails using ethanol instead of phenol-chloroform (Wang and Storm, 2006) or using isopropanol instead of ethanol. These methods provide fewer tube changes, increasing speed of extraction and are more cost effective than buying commercial kits.
1.2. Phylogenetics of Arvicolinae:

Phylogenetics is the study of evolutionary relationships in biology, from individual genes to populations, species, and groups of species. Traditionally, morphological characteristics were used to construct phylogenies of species. More recently molecular markers and whole genomes are being used to further resolve phylogenies and offer a greater understanding of evolutionary processes. Phylogenetics is incredibly important in conservation management by resolving taxonomic uncertainties between species and defining evolutionary significant units (ESUs) within species that need to be protected separately. Phylogenetics is also used in various other biological disciplines, such as pathology to identify emerging pathogens, understanding the relationship to other pathogens, and the likely source of transmission.

Rodents (order Rodentia) are one of the most speciose orders within the mammalian kingdom, containing 2,552 species (513 genera) out of the 6,495 species (1,314 genera) of mammals (Burgin et al., 2018). Within Rodentia is the family Muroidea (mice, rats, voles, hamsters, etc.) and within this group the subfamily Arvicolinae (containing voles, lemmings, and muskrats). The phylogeny of genera within Arvicolinae remains debated, with the evolutionary relationships of many genera and species unresolved. One reason is the subfamily contains 150 species in 30 genera (Carleton and Musser, 2005), with very few sequenced genomes nor a wide selection of sequenced phylogenetic markers.

Several studies have looked at the phylogenetics of rodents and species within Arvicolinae, using mitochondrial and nuclear molecular markers. An early study used 1.2kb of the interphotoreceptor retinoid-binding protein (IRBP) nuclear gene with 22 rodent species (DeBry and Sagel, 2001). This validated the monophyletic group Muroidea, which contains the

<table>
<thead>
<tr>
<th>BOX 2 – Phylogenetic Terms:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monophyletic</strong> – A group of organisms all sharing a common ancestor.</td>
</tr>
<tr>
<td><strong>Paraphyletic</strong> – An artificial group of organisms sharing a common ancestor but that does not include all descendants.</td>
</tr>
<tr>
<td><strong>Clade</strong> – A monophyletic group.</td>
</tr>
<tr>
<td><strong>Node</strong> – A branch point on a phylogenetic tree.</td>
</tr>
<tr>
<td><strong>Polytomy</strong> – An internal node of a phylogenetic tree that has more than two immediate descendants.</td>
</tr>
<tr>
<td><strong>Basal Group</strong> – The earliest diverging group within a clade.</td>
</tr>
</tbody>
</table>
subfamily Arvicolinae, but did not cover sufficient arvicoline taxa. Later studies, such as Blanga-Kanfi et al., 2009, used more molecular markers to resolve the evolutionary relationships between Rodentia families, confirming the monophyletic group Arvicolinae.

Studies focusing on the phylogenetics of Arvicolinae started by sequencing and analysing the mitochondrial cytochrome b (Cytb) gene. Several studies found rapid, near simultaneous radiations when using this marker (summarised in Robovský, Řičánková and Zrzavý, 2008). Another hypothesis is that substitution saturation has occurred in this gene, reducing its value as a phylogenetic marker. Genetic saturation was found at both transitions and transversions of the Cytb gene in arvicoline species (Triant and DeWoody, 2008).

Multiple genetic markers have been used in subsequent studies. For example, one study used mitochondrial Cytb and nuclear growth hormone receptor (GHR) genes, as well as morphological characters (Robovský, Řičánková and Zrzavý, 2008). They found the basal arvicoline in the proposed phylogeny tree (Figure 4a) to be Ellobius, Prometheomys, Hyperacrius, Eolagus and Lagurus. The next to branch was the clade ‘Dicrostonychini’ containing genera Dicrostonyx, Phenacomys, and Arborimus. The next group to diverge was not fully resolved in this study, with Dinaromys, Neofiber and Ondatra, and the clade ‘Lemmini’ (containing Synaptomys, Lemmus, and Myopus) forming a polytomy at this node. The clade ‘Clethrionomyini’ (Eothenomys, Myodes, and Alticola) was well-supported in both this study and other studies. The final clade ‘Arvicolini’ contains genera Arvicola, Lemmiscus, Stenocranius, Chionomys, and Microtus. Other genera were originally found within this group, but this led to the paraphyly of Microtus. The genera Neodon, Alexandromys, Mynomes, Lasiopodomys, Proedromys, and Terricola were reclassified in this study to Microtus. The relationships between taxa in this clade is poorly resolved due to polytomies.

A more recent tree of Arvicolinae can be found in a later study which sampled 900 Muroidea species, with substantial numbers of arvicolines included, using six molecular markers (Steppan and Schenk, 2017; Figure 4b). Although not all genera were accounted for, there is better resolution at some nodes. The basal arvicolines in this tree were Prometheomys, followed by a clade containing Ondatra and Neofiber. This was followed by a monophyletic group containing Dicrostonyx, Arborimus, and Phenacomys in one clade and another clade containing Lemmus, Myopus, and Synaptomys. This supports the previously described study with the clades ‘Dicrostonychini’ and ‘Lemmini’. The well supported clade containing Eothenomys, Myodes, and Alticola diverged next. Dinaromys and Lagurus and Eolagus diverged much later in this tree, followed by Ellobius. This is significantly different to their proposed phylogeny in Robovský et al., 2008. Arvicola grouped together with Lemmiscus, as the basal taxa within ‘Arvicolini’, and followed by the branching of Chionomys. There is then a polytomy between a group containing...
Volemys and Proedromys bedfordi, a branch containing Proedromys liangshanensis, and a clade containing Microtus, Lasiopodomys, Neodon, and Blanfordimys species. Microtus is also paraphyletic in this study. This demonstrates even with extensive molecular markers the phylogeny of all Arvicolinae genera is still unresolved due to lack of support at some nodes.

Figure 4: Simplified cladograms from (a) the proposed phylogeny of Arvicoline by Robovský, Řičánková and Zrzavý, 2008 and (b) the phylogenetic tree of Arvicolinae by Steppan and Schenk, 2017.

Recently, the mitochondrial genome has been sequenced for a number of arvicoline species (e.g. Folkertsma et al., 2018; Bondareva and Abramson, 2019; Zhu et al., 2019; Alqahtani et al., 2020). These studies focus on genomic sequencing and mapping of mitochondrial genes, with brief phylogenetic analyses using available mitogenomes. However, the analyses are not sufficient to fully resolve the phylogenetic relationships between genera and species within Arvicolinae, due to small sample sizes and lack of extensive phylogenetic analysis.
1.3. Hypothesis:
We hypothesise that firstly sufficient quantities of DNA will be extracted from both tissue and non-invasive samples to sequence the mitochondrial DNA control region. Secondly, the genetic diversity of water vole populations in Britain will be low and considerably lower in the captive population, due to the impact that captive breeding programs can have on genetic diversity. Thirdly, using the mitochondrial genome will improve and further resolve the phylogeny of taxa within the rodent subfamily Arvicolinae.

1.4. Research Aims:
This research project is divided into three sections with the following aims:

1. To optimise DNA extraction protocols from various sample types, prioritising non-invasive genetic sampling.
2. To assess the genetic diversity of British water voles (*Arvicola amphibius*) at Wildwood Trust through the random sequencing of individuals in captivity, comparing the genetic diversity between the captive population and natural populations in the South East of England, Britain, and the rest of Europe.
3. To improve the phylogeny of the rodent subfamily Arvicolinae using available molecular markers and various phylogenetic approaches.
2. Materials and Methods:

2.1. Optimisation of DNA Extraction:

Sample Collection:
Water vole samples were collected from Wildwood Trust on several occasions totalling 20
individuals (Table 1). Four samples were collected from tail tissue of deceased water voles and
stored at -20°C. Six samples were collected from hair tufts collected in 2019 and stored in paper
envelopes at room temperature. Another 10 samples were collected in 2019 from faecal pellets
found in water bowls within enclosures and stored at -20°C. All individuals were randomly
chosen, and non-invasive sampling was prioritised. Faecal samples were only collected from
enclosures with single voles or those containing mother and offspring.

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>Sample Type</th>
<th>Enclosure No.</th>
<th>Local ID</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tissue</td>
<td>TB31</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Tissue</td>
<td>WW46</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Tissue</td>
<td>WW0304/34</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>Tissue</td>
<td>WW34/39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Hair</td>
<td>Q88</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>6</td>
<td>Hair</td>
<td>Q100</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>7</td>
<td>Hair</td>
<td>R95</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>8</td>
<td>Hair</td>
<td>R12</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>9</td>
<td>Hair</td>
<td>R28</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>10</td>
<td>Hair</td>
<td>Q100</td>
<td>-</td>
<td>Male</td>
</tr>
<tr>
<td>11</td>
<td>Faecal</td>
<td>R2</td>
<td>2228</td>
<td>Male</td>
</tr>
<tr>
<td>12</td>
<td>Faecal</td>
<td>Q52</td>
<td>2245</td>
<td>Female</td>
</tr>
<tr>
<td>13</td>
<td>Faecal</td>
<td>Q42</td>
<td>2218</td>
<td>Female</td>
</tr>
<tr>
<td>14</td>
<td>Faecal</td>
<td>Q7</td>
<td>2264</td>
<td>Female</td>
</tr>
<tr>
<td>15</td>
<td>Faecal</td>
<td>Q75a</td>
<td>2326</td>
<td>Female</td>
</tr>
<tr>
<td>16</td>
<td>Faecal</td>
<td>R50</td>
<td>2232</td>
<td>Male</td>
</tr>
<tr>
<td>17</td>
<td>Faecal</td>
<td>R51</td>
<td>2225</td>
<td>Male</td>
</tr>
<tr>
<td>18</td>
<td>Faecal</td>
<td>Q58</td>
<td>2314</td>
<td>Male</td>
</tr>
<tr>
<td>19</td>
<td>Faecal</td>
<td>Q100</td>
<td>2185</td>
<td>Female</td>
</tr>
<tr>
<td>20</td>
<td>Faecal</td>
<td>R27</td>
<td>2445</td>
<td>Female</td>
</tr>
</tbody>
</table>

DNA Extraction of Tissue Samples:
Tail tissue was obtained from frozen deceased water voles due to its accessibility and ease when
dissecting. Approximately 1-2 cm of tail tissue was used and fragmented into smaller pieces.
Water vole samples 1 and 2 were used to optimise DNA extraction of tissue samples. DNA was
then extracted from samples 3 and 4 using the optimised protocol. Based on preliminary studies the ‘Qiagen DNeasy Blood and Tissue Kit’ was omitted due to very low concentrations of DNA and poor purity ratios. All DNA concentrations and ratios were measured using NanoDrop and is applicable to all sample types.

Extraction buffers were assessed using phenol-chloroform DNA extraction. Three buffers were selected: buffer 1 (Jain et al., 2017) containing 10 mM Tris-Cl (pH 8.0), 25 mM EDTA (pH 8.0), 100 mM NaCl, and 0.5% SDS; buffer 2 (Green, M. R., Hughes, H., Sambrook, J. and MacCallum, 2012) containing 20 mM Tris-HCl (pH 8.0), 5 mM EDTA (pH 8.0), 400 mM NaCl, and 1% SDS; and buffer 3 (Wang and Storm, 2006) containing 100 mM Tris-HCl (pH 8.0), 5 mM EDTA (pH 8.0), 200 mM NaCl, and 0.2% SDS. For the lysis step, 500 µl of the selected buffer and 20 µl of proteinase K was added to the sample and incubated overnight at 55°C on a shaking platform. Phenol-chloroform extraction followed protocol by Green and Sambrook, 2012.

For the phenol-chloroform extraction, an equal volume of phenol:chloroform:isomyl alcohol was added to tissue samples and placed on a rocking platform for 30 mins. Samples were then centrifuged at 15,000x g for five minutes (at room temperature) and the upper aqueous phase was transferred to a clean Eppendorf tube. DNA was precipitated by adding an equal volume of isopropanol and centrifuged for 15 mins at 13,000x g (at 4°C). Isopropanol was removed, the pellet was rinsed with 70% ethanol, and dried at room temperature. The pellet was dissolved overnight in 100 µl of ddH$_2$O at 4°C.

Ethanol DNA extraction followed (Wang and Storm, 2006). A total of 300 µl of extraction buffer and 6 µl of proteinase K were added to the tissue sample, then incubated overnight at 55°C on a rocking platform. Next, 1 mL of 100% ethanol was added and centrifuged at 15,000x g for 30 mins. Ethanol was poured out and DNA pellets were washed with 70% ethanol. Samples were centrifuged at 15,000x g for 20 mins, ethanol poured out, 300 µl of ddH$_2$O added, and then incubated at 55°C for two hours with lids open. An overnight incubation at 4°C allowed the DNA to fully dissolve.

For the isopropanol protocol 485 µl of extraction buffer and 15 µl of proteinase K were added to the tissue sample, and then incubated overnight at 55°C on a rocking platform. Samples were then centrifuged for 20 mins at 3000 rpm (at room temperature) and the supernatant transferred to a new Eppendorf tube with 500 µl of isopropanol, followed by incubation for 30 mins at -80°C and further centrifugation for 20 mins at 3000 rpm (at 4°C). The supernatant was discarded, 70% ethanol was added, and tubes were centrifuged for 20 mins at 3000 rpm (4°C). The supernatant was discarded again, and the pellet was dried at 37°C. A total of 100 µl of ddH$_2$O was added and incubated overnight at 4°C.
The phenol-chloroform extraction protocol was then optimised. ‘Optimised Phenol-Chloroform (1)’ was identical to the previous protocol described above and used as a control. ‘Optimised Phenol-Chloroform (2)’ was identical to the control apart from an ethanol precipitation rather than an isopropanol precipitation. ‘Optimised Phenol-Chloroform (3) was different in that samples were centrifuged with isopropanol for 30 mins rather than 15 mins. Protocol 3 was chosen for the remaining tissue samples. A 200 µl ddH$_2$O elution volume was used in all three phenol-chloroform optimisation samples.

**DNA Extraction of Hair Samples:**

Prior to DNA extraction, tufts of hair were collected from water voles and stored in paper envelopes. Each tuft contained multiple hairs and follicles. The ‘Qiagen DNeasy Blood and Tissue Kit’ (following the manufactures protocol) was compared with the optimised phenol-chloroform extraction for tissue samples (described previously). Multiple hairs were placed into each Eppendorf tube. Hairs from both protocols were incubated overnight at 56°C with the protocols stated volume of extraction buffer and proteinase K. DNA was eluted into 100 µl of AE (‘Qiagen’) or 100 µl ddH$_2$O (phenol-chloroform).

The buffer was modified to improve the lysis process. The new extraction buffer (Pfeiffer et al., 2004) contained 100 mM Tris HCl (pH 8.0), 100 mM NaCl, 3 mM CaCl$_2$, 2% SDS, and 40 mM dithiothreitol (DTT). A total of 1360 µl of extraction buffer and 80 µl of lysis buffer (proportional to the number of repeats) was added to a falcon tube containing the tuft of hair and incubated overnight at 56°C with agitation. The optimised phenol-chloroform protocol for the tissue samples was used, but with an addition of 5 minutes of centrifugation with isopropanol at maximum speed. For this experiment elution volumes differed (100 µl, 75 µl, and 50 µl).

The remaining hair samples were extracted using the optimised buffer and protocol. Each hair tuft was placed into a falcon tube with 680 µl of extraction buffer and 80 µl of proteinase K, which allowed for a repeat and a change in concentration. DNA was finally eluted in 50 µl of ddH$_2$O.

**DNA Extraction of Faecal Samples:**

DNA from the faecal samples was extracted using the ‘Qiagen QIAamp DNA Stool Mini Kit’ and/or the ‘Qiagen QIAamp PowerFecal DNA Kit’ following the manufactures protocols. The former method eluted 100 µl of Qiagen buffer ATE, whilst the latter eluted 75 µl of ATE. DNA was precipitated using 3M Na-Acetate (pH 5.2) and 100% ethanol, with an incubation at -20°C, followed by two washes with 70% ethanol. DNA was resuspended in 50 µl of ddH$_2$O.
**DNA Extraction of Additional Sample Types:**

Because we could not establish a cell culture for water vole, due to lack of sample availability, DNA was extracted from waterbuck (*Kobus ellipsiprymnus*) cell culture using a phenol-chloroform extraction protocol adapted from (Green, M. R., Hughes, H., Sambrook, J. and MacCallum, 2012). Cells were washed with phosphate-buffered saline (PBS) before DNA extraction. Proteinase K was added to a final concentration of extraction buffer of 400 µg/ml containing 20 mM Tris-Cl (pH 8.0), 5 mM EDTA (pH 8.0), 400 mM NaCl, and 1% SDS. An appropriate volume of buffer was added to cells and incubated overnight at 56°C with agitation. DNA pellets were resuspended in TE buffer.

Cattle (*Bos taurus*) buccal swabs were collected from three individuals and samples were incubated with agitation overnight at 56°C in an extraction buffer (Ghatak, Muthukumaran and Nachimuthu, 2013) containing 10 mM Tris-HCl (pH 8.0), 10 mM EDTA (pH 8.0), 2% SDS, and 20 mg/ml proteinase K. The previous optimised tissue and hair phenol-chloroform extraction protocol was used. DNA was eluted into 50 µl of ddH2O.

### 2.2. Amplification and Sequencing of Mitochondrial Control Region:

**Amplification of Marker:**

Forward and reverse primers for the mitochondrial DNA control region were selected from a previous publication (5’-TTAATCTACCATCCTCCGTGAAACC-3’ and 5’-TKGACACTGGTCTAGGGATATTTGC-3’, respectively; Piertney et al., 2005). All 20 samples were amplified using a PCR reaction mix containing 1x PCR buffer, 200 µM of each dNTP, 0.5 µM forward primer, 0.5 µM reverse primer, and 2.5 units/reaction ‘Qiagen HotStarTaq DNA Polymerase’. Template DNA was then added at a separate workstation (between 9-47 ng/µl). A 15-minute denaturation stage was required at 95°C, followed by 35 cycles of annealing and elongation (94°C for 1 min, 50°C for 1 min, and 72°C for 1 min), ending with 10 mins at 72°C. A negative control and a separate PCR workstation was used for the preparation of the PCR reaction mix to prevent contamination. Amplification was viewed with gel electrophoresis (1% agarose gel, 90 V, ~60 mins, and viewed with Syngene Gel Doc. Samples were purified using ‘Qiagen QIAquick PCR Purification Kit’ following the manufactures protocol. DNA concentrations and purity were measured following PCR clean-up using NanoDrop.

**Sequencing:**

The amplified DNA was sent for sequencing at DBS Genomics, Durham, UK. Both the forward and reverse strands for each individual water vole were sequenced. We used the package ‘Geneious Prime 2020.1 (https://www.geneious.com)’ to create a consensus sequence from
both strands of DNA. This included reverse complementing the reverse strand and subsequently aligning both strands using the global alignment tool, with free gaps and 93% similarity. Consensus sequences were exported as FASTA files.

2.3. Population Genetics of the Water Vole:

Multiple Sequence Alignments (MSA):
We aligned DNA sequences from Wildwood Trust and DNA sequences from selected papers (Piertney et al., 2005; Baker, 2015; Brace et al., 2016) using the ‘R’ package ‘Ape v5.3’ (Paradis and Schliep, 2018) and the program ‘Clustal W v2.0’ (Larkin et al., 2007). MSAs were carried out using the ‘clustal()’ command in ‘Ape v5.3’, with default parameters. Alignments were trimmed based on gaps on the borders of the alignment.

Haplotypes were computed from the MSA results using the ‘R’ package ‘Pegas v0.13’ (Paradis, 2010) with the ‘haplotype()’ command and haplotype networks were computed using the ‘haploNet()’ command, both with default parameters.

Phylogenetic Tree Construction:
Neighbour Joining (NJ) and Maximum-Likelihood (ML) trees were constructed using the MSA data in ‘R’ packages ‘Phangorn v2.5.5’ (Schliep et al., 2017), ‘Ape v5.3’ (Paradis and Schliep, 2018) and ‘ggtree v2.0.2’ (Yu, 2020). Both tree types were rooted on the outgroup. For ML trees, the best nucleotide evolution model was chosen using ‘modelTest()’ and then ‘bootstrap.pml()’ was used to perform the bootstrap analysis (1000 replicates with optimised topology). Bayesian phylogenetic trees were constructed in the program ‘MrBayes v3.2.7’ (Huelsenbeck and Ronquist, 2001) and visualised in ‘FigTree v1.4.4’. Standard parameters were used except for changing the nucleotide evolution model (computed in ‘R’) and the number of Markov Chain Monte Carlo (MCMC) generations, which depended on the average standard deviation of split frequencies and whether additional generations were needed if the standard deviation was above 0.01. A consensus tree was produced using the ‘sumt’ command and rooted on the outgroup taxon.

Population Genetics Calculations:
Nucleotide diversity, haplotype diversity, and Tajima’s D were calculated in the ‘R’ package ‘Pegas 0.13’ (Paradis, 2010) using the MSA. Mitochondrial control region sequences of *Myodes glareolus* were used as a comparison to *Arvicola amphibius*. Sequences were obtained from Filipi et al., 2015 and Marková et al., 2020 and aligned, before calculating population genetics statistics.
Population Structure:
All 144 sequences and 642 loci were analysed in the program ‘STRUCTURE 2.3.4’ (Pritchard, Stephens and Donnelly, 2000). Parameters for length of burn-in period was set at 10,000 and number of MCMC generations after burn-in was set at 100,000. The admixture model was chosen, and the number of populations assumed (K) was set from K=1 to K=6, with 5 iterations each. ‘Structure Harvester’ (Earl and vonHoldt, 2012) was then used to identify the most appropriate K value using graphs of L(K) and delta K.

2.4. Phylogenetics of Arvicolinae:
Mitochondrial and Nuclear Markers:
All available mitochondrial genomes, mitochondrial cytochrome b (Cytb) genes, nuclear growth hormone receptor (GHR) genes, and nuclear interphotoreceptor retinoid-binding protein (IRBP) genes were found utilizing the ‘National Centre for Biotechnology Information (NCBI) Nucleotide’ and ‘NCBI BLAST’ databases for available Arvicolinae species. Outgroup taxa were chosen from three subfamilies of the Cricetidae family: Cricetinae (*Cricetulus griseus*), Neotominae (*Peromyscus polionotus*), and Sigmodontinae (*Sigmodon hispidus*). One additional outgroup taxon was chosen from a Muroidea family Muridae (*Mus musculus*) and was used to root the phylogenetic trees.

Phylogenetic Tree Construction:
ML and Bayesian inference phylogenetic trees were constructed using the previously described methods. Trees were constructed for each of the three markers, as well as for the available mitochondrial genomes. A supermatrix approach was used for the three markers. Individual markers were firstly aligned in ‘R’ using ‘Clustal W’ and trimmed. FASTA files of the alignments were concatenated in ‘MEGA X’. Concatenated sequences were then read into ‘R’ as a FASTA file and ML (in ‘R’) or Bayesian inference phylogenetic trees (in ‘MrBayes’) were constructed using the concatenated sequence.
3. Results:

3.1. Optimisation of DNA Extraction:

**DNA Extraction of Tissue Samples:**

The DNA extraction for tissue samples was optimised using fragmented water vole tail tissue. Firstly, the lysis buffer was optimised by comparing three buffers, containing different concentrations of NaCl, EDTA, and Tris-Cl, using the published phenol-chloroform extraction protocol. Buffer 3 had the highest DNA concentration (172.27 ng/µl) and highest DNA yield (17,227 ng), compared to buffers 1 and 2 (Table 2 ‘Optimisation of Buffer’). The two absorbance ratios, 260/280 and 260/230, were 1.51 and 1.56, respectively. The 260/280 absorbance ratios test for the presence of protein and other contaminants in the DNA sample and are expected to be ~1.8 for 260/280 and ~2.0 for 260/230. This indicates a “pure” DNA sample. Buffer 3 was chosen for the following DNA extractions of tissue samples.

Next, three extraction protocols were tested to determine the highest DNA yield. Phenol-chloroform extraction had the highest DNA yield at 17,227 ng, with ethanol extraction and isopropanol extraction having much lower DNA yields (Table 2: ‘Optimisation of Protocol’). Phenol-chloroform extraction also had a purer DNA sample after DNA extraction. The phenol-chloroform extraction protocol was optimised further by considering the precipitation of DNA, either isopropanol as the control (Phenol-Chloroform 1) or ethanol in ‘Phenol-Chloroform 2’ (Table 2: ‘Further Optimisation’). Or the centrifugation of isopropanol to pellet the DNA which was 15 mins in the control (Phenol-Chloroform 1) or 30 mins in ‘Phenol-Chloroform 3’. The added 15 mins to the isopropanol centrifugation increased DNA yield (20,666 ng) compared with the control (3,088 ng).
Table 2: Optimisation of DNA extraction lysis buffers and protocols for water vole tissue samples.

<table>
<thead>
<tr>
<th>Optimisation of Buffer:</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffer 1</td>
<td>100</td>
<td>89.11</td>
<td>8911</td>
<td>1.43</td>
<td>1.47</td>
</tr>
<tr>
<td>Buffer 2</td>
<td>100</td>
<td>102.18</td>
<td>10218</td>
<td>1.39</td>
<td>1.54</td>
</tr>
<tr>
<td>Buffer 3</td>
<td>100</td>
<td>172.27</td>
<td>17227</td>
<td>1.51</td>
<td>1.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimisation of Protocol:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenol-Chloroform</td>
</tr>
<tr>
<td>Ethanol</td>
</tr>
<tr>
<td>Isopropanol</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Further Optimisation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenol-Chloroform (1)</td>
</tr>
<tr>
<td>Phenol-Chloroform (2)</td>
</tr>
<tr>
<td>Phenol-Chloroform (3)</td>
</tr>
</tbody>
</table>

Further Optimisation:

An optimised protocol consisting of the selected buffer, phenol-chloroform extraction, and the added 15 mins of centrifugation with isopropanol was used for the DNA extraction of all tissue samples (Table 3). Gel electrophoresis was used to determine size of the DNA fragments (Figure 5). In all four tissue samples high molecular weight (HMW) bands were seen at around 10 kb. Some samples contained more fragmented DNA than others, as shown by the DNA streak from high to low molecular weight.

Table 3: DNA extraction of all water vole tissue samples. Results for samples 1 and 2 from the previous table. Samples 3 and 4 using the final optimised protocol ('Phenol-Chloroform 3').

<table>
<thead>
<tr>
<th>Sample</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>172.27</td>
<td>17227</td>
<td>1.51</td>
<td>1.56</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>103.33</td>
<td>20666</td>
<td>1.57</td>
<td>1.41</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>30.68</td>
<td>3068</td>
<td>1.59</td>
<td>1.49</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>43.00</td>
<td>4300</td>
<td>1.53</td>
<td>1.61</td>
</tr>
</tbody>
</table>
Figure 5: Gel electrophoresis of all water vole tail tissue samples. Lane ‘L’ shows the 10kb ladder, and lanes 1-4 are the corresponding DNA samples. 1% agarose, 70V, ~60 mins.

DNA Extraction of Hair Samples:

Two hair extraction protocols were compared, the ‘Qiagen DNeasy Blood and Tissue DNA Extraction Kit’ and the phenol-chloroform extraction, for DNA yield and purity of the DNA sample. Both protocols had low DNA yield and poor absorbance ratios (i.e., not similar to the expected values) (Table 4). A new buffer was used that contained CaCl$_2$ and DTT but without EDTA. This was tested with different elution volumes of ddH$_2$O. It was found that using this buffer and eluting 50 µl of ddH$_2$O increased DNA yield (3,794 ng).
Table 4: Optimisation of DNA extraction protocols for hair samples.

<table>
<thead>
<tr>
<th></th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qiagen Kit (1)</td>
<td>100</td>
<td>4.38</td>
<td>438</td>
<td>2.41</td>
<td>0.97</td>
</tr>
<tr>
<td>Qiagen Kit (2)</td>
<td>100</td>
<td>2.21</td>
<td>221</td>
<td>7.23</td>
<td>5.18</td>
</tr>
<tr>
<td>Phenol-Chloroform (1)</td>
<td>100</td>
<td>4.3</td>
<td>430</td>
<td>1.15</td>
<td>0.53</td>
</tr>
<tr>
<td>Phenol-Chloroform (2)</td>
<td>100</td>
<td>10.3</td>
<td>1030</td>
<td>1.41</td>
<td>1.23</td>
</tr>
<tr>
<td>Optimised Phenol-Chloroform (1)</td>
<td>100</td>
<td>17.88</td>
<td>1788</td>
<td>1.24</td>
<td>1.00</td>
</tr>
<tr>
<td>Optimised Phenol-Chloroform (2)</td>
<td>75</td>
<td>17.09</td>
<td>1281.75</td>
<td>1.25</td>
<td>1.19</td>
</tr>
<tr>
<td>Optimised Phenol-Chloroform (3)</td>
<td>50</td>
<td>75.88</td>
<td>3794</td>
<td>1.34</td>
<td>1.46</td>
</tr>
</tbody>
</table>

DNA was extracted from the remaining hair samples using the new lysis buffer and optimised hair extraction protocol, using the previously optimised protocol for tissue samples (Table 5).

Table 5: DNA extraction of hair samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>50</td>
<td>10.3</td>
<td>515</td>
<td>1.41</td>
<td>1.23</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>75.88</td>
<td>3794</td>
<td>1.34</td>
<td>1.46</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>87.56</td>
<td>4378</td>
<td>1.44</td>
<td>1.68</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>13.83</td>
<td>691.5</td>
<td>1.51</td>
<td>1.49</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>9.04</td>
<td>452</td>
<td>1.39</td>
<td>1.30</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>47.02</td>
<td>2351</td>
<td>1.36</td>
<td>1.61</td>
</tr>
</tbody>
</table>

**DNA Extraction of Faecal Samples:**

The ‘Qiagen QIAamp DNA Stool Mini Kit’ and ‘Qiagen QIAamp PowerFecal DNA Kit’ were tested for DNA yield and purity (Table 6). Both extraction kits produced similar DNA yields (a mean of 1,091.25 ng and 1,136.25 ng, respectively). However, values for 260/230 absorbance ratio were considerably different in the two extraction kits. ‘Qiagen QIAamp DNA Stool Mini Kit’ had values greater than 2.0 (the expected value for pure samples), whilst ‘Qiagen QIAamp PowerFecal DNA Kit’ had values considerably lower. Overall, the latter DNA extraction kit performed better for the samples tested.
Table 6: Comparison of different DNA extraction kits for faecal samples.

<table>
<thead>
<tr>
<th>DNA Extraction Kit</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qiagen Stool Mini Kit (1)</td>
<td>100</td>
<td>9.89</td>
<td>989</td>
<td>1.53</td>
<td>2.42</td>
</tr>
<tr>
<td>Qiagen Stool Mini Kit (2)</td>
<td>100</td>
<td>11.78</td>
<td>1178</td>
<td>1.79</td>
<td>2.17</td>
</tr>
<tr>
<td>Qiagen Stool Mini Kit (3)</td>
<td>100</td>
<td>12.24</td>
<td>1224</td>
<td>1.59</td>
<td>2.73</td>
</tr>
<tr>
<td>Qiagen Stool Mini Kit (4)</td>
<td>100</td>
<td>9.74</td>
<td>974</td>
<td>1.54</td>
<td>2.02</td>
</tr>
<tr>
<td>Qiagen Power Faecal Kit (1)</td>
<td>75</td>
<td>10.81</td>
<td>810.75</td>
<td>2.43</td>
<td>0.44</td>
</tr>
<tr>
<td>Qiagen Power Faecal Kit (2)</td>
<td>75</td>
<td>19.49</td>
<td>1461.75</td>
<td>1.86</td>
<td>0.17</td>
</tr>
</tbody>
</table>

DNA was extracted from all 10 faecal samples from Wildwood Trust using one or both extraction kits (Table 7). All samples were precipitated with ethanol to increase DNA yield and improve the purity of the DNA sample. This was the case for the majority of samples, where DNA yield was higher and absorbance values were closer to the expected values following precipitation.

Table 7: DNA concentrations and purity of extracted faecal samples using ‘Qiagen QIAamp DNA Stool Mini Kit’ (A) and ‘Qiagen QIAamp PowerFecal DNA Kit’ (B) before and after ethanol precipitation.

<table>
<thead>
<tr>
<th></th>
<th>Before Precipitation</th>
<th></th>
<th>After Precipitation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E.V. (µl)</td>
<td>ng/µl</td>
<td>ng</td>
<td>260/280</td>
</tr>
<tr>
<td>Q58 (A)</td>
<td>100</td>
<td>6.16</td>
<td>616</td>
<td>2.60</td>
</tr>
<tr>
<td>Q75A.2 (A)</td>
<td>100</td>
<td>4.25</td>
<td>425</td>
<td>1.63</td>
</tr>
<tr>
<td>R2 (A)</td>
<td>100</td>
<td>12.03</td>
<td>1203</td>
<td>1.54</td>
</tr>
<tr>
<td>R51 (A)</td>
<td>100</td>
<td>5.92</td>
<td>592</td>
<td>1.52</td>
</tr>
<tr>
<td>Q42.2 (A)</td>
<td>100</td>
<td>4.05</td>
<td>405</td>
<td>1.65</td>
</tr>
<tr>
<td>R50 (A)</td>
<td>100</td>
<td>6.05</td>
<td>605</td>
<td>2.05</td>
</tr>
<tr>
<td>R27 (A)</td>
<td>100</td>
<td>9.72</td>
<td>972</td>
<td>1.78</td>
</tr>
<tr>
<td>Q52 (A)</td>
<td>100</td>
<td>5.32</td>
<td>532</td>
<td>1.94</td>
</tr>
<tr>
<td>Q100 (A)</td>
<td>100</td>
<td>10.68</td>
<td>1068</td>
<td>1.89</td>
</tr>
<tr>
<td>Q7 (A)</td>
<td>100</td>
<td>5.21</td>
<td>521</td>
<td>1.10</td>
</tr>
<tr>
<td>R51 (B)</td>
<td>75</td>
<td>10.81</td>
<td>810.75</td>
<td>2.43</td>
</tr>
<tr>
<td>R27 (B)</td>
<td>75</td>
<td>19.49</td>
<td>1461.75</td>
<td>1.86</td>
</tr>
</tbody>
</table>
**DNA Extraction of Additional Sample Types:**

DNA extraction from cell culture was used to demonstrate the benchmark method of extracting large quantities of DNA and pure DNA samples. For this, we used an available waterbuck cell culture and extracted DNA from the flask. This resulted in high DNA yield with a mean of 46,566.67 ng (**Table 8**). Sample purity was lower than expected for all samples and both absorbance ratios.

**Table 8: Extraction of DNA from waterbuck cell culture.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
<td>53.20</td>
<td>26600</td>
<td>1.39</td>
<td>1.65</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>145.00</td>
<td>72500</td>
<td>1.49</td>
<td>1.71</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>81.00</td>
<td>40500</td>
<td>1.44</td>
<td>1.69</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>46.20</td>
<td>23100</td>
<td>1.45</td>
<td>1.73</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>218.20</td>
<td>109100</td>
<td>1.55</td>
<td>1.76</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>15.20</td>
<td>7600</td>
<td>1.55</td>
<td>1.90</td>
</tr>
</tbody>
</table>

DNA was also extracted from cattle buccal swabs from three individuals. This was to demonstrate another useful way of sampling animals. We achieved a high DNA yield for all samples, with samples 2 and 3 having around three times as much DNA (**Table 9**). This may be due to the amount of saliva that was collected from each cattle. On average this resulted in 11,937.17 ng of DNA.

**Table 9: Extraction of DNA from cattle buccal swabs.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Elution Volume (µl)</th>
<th>ng/µl</th>
<th>ng</th>
<th>260/280</th>
<th>260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>106.60</td>
<td>5330</td>
<td>1.72</td>
<td>2.11</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>302.59</td>
<td>15129.5</td>
<td>1.75</td>
<td>2.03</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>307.04</td>
<td>15352</td>
<td>1.48</td>
<td>1.64</td>
</tr>
</tbody>
</table>

**Comparison of Sample Types:**

All sample types were compared for DNA yield (**Figure 6**). As expected, DNA extraction from cell culture resulted in the highest median DNA yield (33,550 ng). Tail tissue and buccal swabs both yielded the next highest DNA yield (10,763.5 ng and 15,129.5 ng, respectively). Hair and faecal samples produced the lowest DNA yield (1,521.25 ng and 1,018.25 ng, respectively). An Analysis
of Variance (ANOVA) test found a statistically significant difference in average DNA yield by sample type \( (f(4) = 7.223, p < 0.001) \).

![Boxplot showing DNA yield by sample type](image1.png)

Figure 6: Comparison of the median DNA yields (ng) for each sample type. Calculated from DNA yields from samples using the final optimised protocols.

3.2. Amplification and Sequencing of the Mitochondrial Control Region:

After optimisation of PCR protocols, the mitochondrial DNA control region for all 20 water voles was amplified. Gel electrophoresis was used to identify which samples amplified, as shown by the presence of bands between 750 bp and 1000 bp (Figure 7). Post-PCR, samples were cleaned to remove any PCR products and then sent for Sanger sequencing.

![Gel electrophoresis image](image2.png)

Figure 7: Gel electrophoresis of the amplified mitochondrial DNA control region for 20 water voles. L is the DNA ladder with markings at 750bp and 1000bp. Poorly amplified samples were repeated.
A total of 20 individuals were sequenced, but three had poor chromatogram results. Of these three samples, one was hair sample 5 and two were faecal samples 12 and 18. All three samples also proved difficult to amplify before sequencing, with numerous PCR repeats required to achieve a visible band with gel electrophoresis. These three samples were excluded from further analysis.

Chromatograms of the sequence files showed a region rich with A/T repeats, followed by weak and inhibited peaks, with the sequence unable to be accurately determined. Therefore, the forward and reverse strands were trimmed at this region and aligned to form a consensus sequence. The 17 Wildwood Trust consensus sequences were used for further analysis.

3.3. Population Genetics of the Water Vole:

*Individuals from Wildwood Trust:*

A haplotype network and ML phylogenetic tree were constructed using the 17 mitochondrial control region DNA sequences from the sampled Wildwood Trust water voles. Out of 17 individuals there were 12 haplotypes (Figure 8a) and two main haplogroups. There were six mutational steps between the two haplogroups (i.e. between haplotype 1 and 4). One haplogroup contained haplotypes 1 and 2, and the other contained the 10 remaining haplotypes. Haplotypes 1, 2 and 4 were sampled from tissue and were some of the oldest samples collected from Wildwood Trust. Three haplotypes contained more than one individual. Haplotype 3 contained individuals 3, 6, and 13, haplotype 8 contained individuals 10, 14, and 16, and haplotype 9 contained individuals 11 and 17. Within the largest haplogroup there was a maximum of two mutational steps between each haplotype.

A ML tree was constructed, and this grouped individuals 3, 6, 9, 10, 11, 13, 14, 15, 16, and 17 into the same clade with a polytomy. Therefore, only one sequence from these individuals was used in the following phylogenetic analyses to aid visualisation. The multiple individuals were grouped into a clade with individual 7 (Figure 8b) and were the last to diverge. This result conflicted with the haplotype network as these individuals are not grouped into one haplotype, but instead formed haplotypes 3, 7, 8, 9, and 10. Each of the haplotypes had multiple mutational steps between them. Individuals 1 and 2 diverged first in the Wildwood Trust samples and this was supported by haplotypes 1 and 2 forming a separate haplogroup to all other individuals and a 100% bootstrap score for this specific node.
Figure 8: **Analysis of Wildwood Trust water voles using the mitochondrial control region.** (a) Haplotype network of Wildwood Trust samples. Each pie chart represents a haplotype, and each colour represents a different sample. The dotted lines are the number of mutational steps between haplotype sequences. (b) ML phylogenetic tree of Wildwood Trust samples. Bootstrap scores are shown for each node. The tree is rooted on *Arvicola sapidus*.

**Comparison of Captive and Natural Populations:**

Captive Wildwood Trust sequences, as well as haplotype sequences from Baker, 2015 from natural water vole populations in the South East and East of England were aligned and a haplotype network was constructed. There were 26 haplotypes in total, 12 haplotypes from the Wildwood Trust samples and 15 haplotypes from Baker, 2015 (Figure 9). Haplotypes formed two haplogroups. One contained only one haplotype, the South East of England haplotype 14, which was 17 mutational steps from the other haplogroup, where the remaining samples were clustered. Captive individuals were found in separate haplotypes to natural individuals. Haplotypes 1 and 2, containing captive individuals, were closely clustered with haplotype 25, which contained natural water voles from the East of England. These haplotypes were five mutational steps from other haplotypes (i.e. between haplotype 1 and haplotype 3).
Figure 9: Captive and wild water vole samples from the South East and East of England. Wildwood Trust sequences are individual sequences, whilst Baker 2015 sequences are haplotype sequences.

The captive Wildwood Trust sequences were compared with sequences from the (Piertney et al., 2005) paper which sampled water voles across the British isle from natural populations. A haplotype network was constructed using the alignment of DNA sequences for 17 Wildwood Trust individuals and 57 sequences for natural water voles in Britain and five from mainland Europe (Figure 10).

A total of 45 haplotypes were found for 78 individuals. There were two main haplogroups, one containing individuals from Scotland and mainland Europe, and the other containing individuals from England, Wales, and mainland Europe. This latter haplogroup contained all Wildwood Trust individuals. There were 13 mutational steps between haplotype 14 from the Scotland haplogroup and haplotype 45 from the England/Wales haplogroup. All Wildwood
Trust individuals formed separate haplotypes to the natural individuals and formed the same haplotypes as the previous haplotype network (Figure 8a). In both haplogroups there were multiple haplotypes with more than one individual. Haplotypes 42-45 were individuals sampled from mainland Europe. Mainland Europe samples were found in the Scotland haplogroup, except for haplotype 45 which contained a water vole from Finland and was found in the England/Wales haplogroup.

Figure 10: Haplotype network using DNA sequences of the mitochondrial control region for captive and natural water voles. Each pie chart represents a haplotype, and the size represents the number of individuals with the haplotype. The dotted lines are the number of mutational steps between haplotypes.
Wildwood Trust sequences were also aligned with sequences from Brace et al., 2016, a paper which used ancient DNA from museum specimens dating back to the Pleistocene (before the last glacial period). Phylogenetic trees, both ML (Figure 11) and Bayesian inference, were constructed using an alignment containing sequences from Wildwood Trust, Piertney et al., 2005, Baker, 2015, and Brace et al., 2016. Both phylogenetic trees showed comparable results. All Holocene individuals, most modern English/Welsh water voles, all South East and East of England water voles, and all Wildwood Trust individuals were grouped into one clade (shaded in dark grey). Wildwood Trust and 14 out of 15 of the South East and East of England water voles were grouped more closely, whilst South East of England 11 was grouped with modern samples from Somerset, Wales, Staffordshire, and Shropshire. The sister clade contained all Pleistocene and Scottish water voles, and three modern English samples (shaded in a lighter grey). Samples from mainland Europe were found in both clades. Water vole from Italy and Switzerland were grouped into a separate, early diverging clade, after the outgroup species (*Arvicola sapidus*). *Arvicola scherman* samples were found in both the English/Welsh clade and the Scottish clade. The phylogenetic tree had relatively high bootstrap scores at nodes separating the major three clades. Polytomies were seen at nodes within the major clades, so relationships between individuals is less clear.
Figure 11: ML phylogenetic tree containing all samples. Taxon colours represent the age or location of sample. The tree is rooted on Arvicola sapidus. Taxon “Wildwood Trust*” contains multiple Wildwood Trust samples. Only bootstrap scores greater than 50% are shown.
Population genetics calculations were computed for each of the alignments (Table 10). Wildwood Trust water voles had lower haplotype and nucleotide diversity (0.949 and 0.004, respectively) compared with natural British water voles (0.971 and 0.016, respectively). Tajima’s D was lower in Wildwood Trust individuals (-2.186 compared with -0.113). Natural English and Welsh water voles had higher haplotype and nucleotide diversity (0.982 and 0.008, respectively), and lower Tajima’s D (-2.164) than natural Scottish water voles which had a haplotype diversity of 0.945, a nucleotide diversity of 0.009, and Tajima’s D of -1.857. The natural Scottish population had a lower haplotype diversity than the captive water vole population (Wildwood Trust).

To put British water vole genetic diversity into perspective, we compared population genetics estimates with a sister species, the bank vole (*Myodes glareolus*) which had stable numbers in Britain. Available mitochondrial control region sequences for the bank vole were aligned and population genetics calculations were computed. Focusing on British populations, bank voles had lower haplotype and nucleotide diversity (0.967 and 0.006, respectively) than natural British populations (0.971 and 0.016, respectively). British bank voles had lower haplotype and nucleotide diversity than all European bank vole samples.

Table 10: European water vole (*Arvicola amphibius*) and bank vole (*Myodes glareolus*) population genetics. Number of sequences (n), alignment length in base pairs (bp), number of haplotypes (Hap. No.), haplotype diversity (Hap. Div.), variance (Var.), nucleotide diversity (π), Tajima’s D (D), P-value for a normal distribution (P-value Norm.), and P-value for a beta distribution (P-value Beta). *Arvicola amphibius* sequences from: (1) Baker, 2015, (2) Piertney et al., 2005, and (3) only modern samples from Brace et al., 2016. *Myodes glareolus* sequences were from Filipi et al., 2015 and Marková et al., 2020.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Captive Wildwood Trust</td>
<td>17</td>
<td>706</td>
<td>12</td>
<td>0.949</td>
<td>0.001</td>
<td>0.004</td>
<td>0.000</td>
<td>-2.186</td>
<td>0.029</td>
<td>0.007</td>
</tr>
<tr>
<td>Natural South East / East of England Haplotypes (1)</td>
<td>15</td>
<td>731</td>
<td>14</td>
<td>N/A</td>
<td>N/A</td>
<td>0.007</td>
<td>0.000</td>
<td>-2.378</td>
<td>0.017</td>
<td>0.000</td>
</tr>
<tr>
<td>Natural English and Welsh (2) &amp; (3)</td>
<td>32</td>
<td>644</td>
<td>32</td>
<td>0.982</td>
<td>0.000</td>
<td>0.008</td>
<td>0.000</td>
<td>-2.164</td>
<td>0.030</td>
<td>0.011</td>
</tr>
<tr>
<td>Natural Scottish (2) &amp; (3)</td>
<td>25</td>
<td>644</td>
<td>25</td>
<td>0.945</td>
<td>0.000</td>
<td>0.009</td>
<td>0.000</td>
<td>-1.857</td>
<td>0.063</td>
<td>0.041</td>
</tr>
<tr>
<td>Natural British (2) &amp; (3)</td>
<td>67</td>
<td>639</td>
<td>39</td>
<td>0.971</td>
<td>0.000</td>
<td>0.016</td>
<td>0.000</td>
<td>-0.113</td>
<td>0.910</td>
<td>0.950</td>
</tr>
<tr>
<td>Natural Mainland European (2) &amp; (3)</td>
<td>20</td>
<td>639</td>
<td>19</td>
<td>0.995</td>
<td>0.000</td>
<td>0.024</td>
<td>0.000</td>
<td>-0.753</td>
<td>0.452</td>
<td>0.491</td>
</tr>
<tr>
<td><em>Myodes glareolus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>British</td>
<td>24</td>
<td>940</td>
<td>17</td>
<td>0.967</td>
<td>0.000</td>
<td>0.006</td>
<td>0.000</td>
<td>-1.115</td>
<td>0.265</td>
<td>0.281</td>
</tr>
<tr>
<td>All</td>
<td>118</td>
<td>940</td>
<td>97</td>
<td>0.996</td>
<td>0.000</td>
<td>0.009</td>
<td>0.000</td>
<td>-1.775</td>
<td>0.076</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Lastly, we analysed the population structure of all water vole mitochondrial DNA control region sequences (Figure 12) to distinguish the groupings of individuals into populations. The
sequences were grouped into two populations (K=2), based on the highest delta K value. All modern English/Welsh and Holocene samples were grouped into one population (red) and all modern Scottish and Pleistocene samples were grouped into another population (green). This is comparable with the ML phylogenetic tree, with two distinct clades for British water voles. Some admixture was seen between the two populations, but only in samples from other European countries.

![Figure 12: Structure analysis of mtDNA for all sequences (K=2).](image)

(a) Run 3/5 for K=2. Red and green subpopulations. (b) Delta K: K=1 to K=6 with 5 runs each. K=2 selected. (c) L(K): K=1 to K=6 with 5 runs each.

3.4. Arvicolinae Phylogenetics:

Using available mitochondrial and nuclear sequences from ‘NCBI’ for Arvicolinae taxa, we constructed ML and Bayesian inference phylogenetic trees. Several approaches were taken using different genetic markers. (1) whole mitochondrial genomes, (2) individual molecular markers (Cytb, GHR, and IRBP), and (3) concatenated alignment of all three markers.

**Mitochondrial Genomes:**

All available mitochondrial genomes from ‘NCBI’ were obtained and aligned, producing a 16,689bp multiple sequence alignment with 7,678 sites with at least one substitution and 34 taxa (30 Arvicolinae taxa). ML and Bayesian inference trees were constructed using the Generalised Time Reversible (GTR) + G + I substitution model, based on the model with the lowest Akaike Information Criterion (AIC) value. Both approaches resulted in comparable topologies. The maximum-likelihood phylogenetic tree had high bootstrap scores for most
nodes, with a few nodes having bootstrap scores below 50% (Figure 13). Four outgroup species were used as controls and all diverged before Arvicolinae taxa, with the tree rooted on *Mus musculus*. *Ondatra zibethicus* and *Dicrostonyx* species diverged first within the available Arvicolinae taxa and formed a monophyletic group. *Eothenomys* and *Myodes* species diverged next, forming a separate clade. *Arvicola amphibius* diverged after supported by 100% bootstrap score, followed by *Prodromys liangshanensis* with a bootstrap score below 50%. *Microtus fortis* and *Microtus kikuchii* formed a monophyletic group with *Lasiopodomys* and *Neodan* species, whilst the remaining *Microtus* taxa formed a clade with *Terricola subterraneus*. The Bayesian inference tree had posterior probabilities for all nodes of 1.00/100% (Appendix A1).

Figure 13: Maximum-likelihood phylogenetic tree of Arvicolinae mitochondrial genomes. Bootstrap scores above 50% are shown.
**Mitochondrial Marker:**

A phylogenetic analysis of the mitochondrial *Cytb* was applied as a comparison to the mitogenome analysis. Substitution model GTR + G + I was used to construct a ML ([Figures 14 and A2](#)) and Bayesian inference ([Figure A3](#)) phylogenetic tree was constructed. Substantial differences were seen between the two mitochondrial analyses ([Figure 14](#)). The position of *Arvicola amphibius* changed when using only the *Cytb* marker and formed a monophyletic group with *Dicrostonyx* species. *Lasipodomys* spp., *Proedromys liangshanensis*, and *Microtus kikuchii* formed a monophyletic group. *Microtus agrestis* formed a monophyletic group with *Neodon* species. *Microtus fortis* was grouped with other *Microtus* species, not *Microtus kikuchii*. Similarities were the grouping of *Eothenomys* spp. and *Myodes* spp., *Ondatra zibethicus* being the first taxa to diverge, and *Terricola subterraneous* forming a monophyletic group with selected *Microtus* species.

![Figure 14: Maximum-likelihood trees of (a) mitogenomes and (b) mitochondrial *Cytb* sequences. Rooted on *Mus musculus*. Bootstrap scores for *Cytb* ML phylogenetic tree can be found in the Appendix.](#)
All available Cytb sequences were also used to construct phylogenetic trees. This allowed for a more extensive taxon sampling, with more species for the genus *Microtus* and more genera included. The Cytb alignment contained 147 sequences with 1,143 sites. The GTR + G + I substitution model was used to construct ML and Bayesian inference phylogenetic trees (Figures A4 and A5). A significant finding was the position of *Cricetulus migratorius*, an outgroup taxon, which was found in a clade with *Eolagurus* and *Lagurus*, suggesting the unreliability of this tree. *Dicrostonyx* diverged first within Arvicolinae, but with a bootstrap score of less than 50%. This was followed by *Ondatra* and *Neafiber*. Other notable findings that were not seen in the previous Cytb tree, with only selected taxa that had mitogenomes sequenced, was *Arvicola* grouped in a clade with *Dinaromys* and *Lemmiscus*, as well as the paraphyly of *Volemys*. Overall, there was very low bootstrap scores (less than 50%) for many of the major nodes, therefore this showed a lack support for using the Cytb gene alone to determine the phylogeny of this subfamily, even with extensive taxon sampling.

**Nuclear Markers:**

Nuclear sequences for the GHR and IRBP genes were obtained and sequences for each were aligned. The GHR alignment contained 74 sequences with 725 sites. The GTR + G + I substitution model was used to construct ML (Figures 15a and A6) and Bayesian inference phylogenetic trees (Figure A7). The IRBP alignment contained 84 sequences with 1,198 sites (482 of which had at least one substitution). The HKY + G + I substitution model was used, based on the lowest AIC value, and a ML (Figures 15b and A8) and a Bayesian inference (Figure A9) phylogenetic tree was constructed.

There was little support for the tree produced using the GHR marker because there were high levels of polytomy and low bootstrap scores at many nodes. Low bootstrap scores (below 50%) were also seen for many nodes in the IRBP phylogenetic tree, but less polytomy than when using the GHR gene. Similarities included the grouping of the genera *Alticola*, *Eothenomys*, and *Myodes*, the grouping of *Arborimus*, *Phenacomys*, and *Dicrostonyx*, *Arvicola* branching separately, and the paraphyly of *Microtus*. Differences were *Prometheomys schaposchnikovi* diverged first out of the Arvicolinae taxa in the GHR tree, but support for this node was below 50%, whereas in the IRBP tree *Phenacomys*, *Arborinus*, and *Dicrostonyx* genera were the first to diverge. *Microtus* was paraphyletic with genera *Terricola* and *Blanfordimys* in the GHR tree, but in the IRBP tree *Microtus* was paraphyletic with genera *Terricola*, *Lasiopodomys*, *Proedromys*, and *Alexandromys*. 
Figure 15: Maximum-likelihood phylogenetic tree of Arvicolinae using (a) GHR and (b) IRBP nuclear markers. Both trees are rooted on Mus musculus. Bootstrap scores can be found in the Appendix.
Supermatrix Approach:

Taxa which had available sequences for mitochondrial Cytb, nuclear GHR, and nuclear IRBP genes were selected, and each individual marker was aligned. The three alignments were then concatenated. The concatenate contained 3,101 sites (1,268 sites with at least one substitution) and 57 sequences (including four outgroup taxa). This alignment was then used to construct the supermatrix tree. ML (Figure 16) and Bayesian inference (Figure A10) phylogenetic trees were constructed, based on the substitution model GTR + G + I. Using a supermatrix approach improved the support for the topology, with no polytomies and a greater number of nodes with bootstrap scores higher than 50%. *Neofiber alleni* diverged first in the Arvicolinae species, with a bootstrap score of 65.6%. *Lemmus*, *Synaptomys*, *Arborimus*, *Phenacomys*, and *Prometheomys* were grouped into a monophyletic group and were the next to diverge. *Alticola*, *Myodes*, and *Eothenomys* formed a monophyletic group. *Microtus* species were grouped into one monophyletic clade with *Terricola daghestanicus* and *Alexandromys middendorffi*. *Neodon spp.*, *Lasiopodomys spp.*, and *Volemys milicens* were grouped together into a monophyletic group, whilst *Volemys musseri* diverged separately.
Figure 16: Maximum likelihood supermatrix for Cytb, GHR, and IRBP genes. Rooted on *Mus musculus*. Only bootstrap scores over 50% are shown.
4. Discussion:

4.1. DNA Extraction and Amplification:

All sample types used in this project resulted in variable amounts of extracted DNA, with a statistically significant difference between average DNA yield and sample type. Cell culture, as to be expected, produced the highest median quantity of DNA (33,550 ng; Figure 6). This demonstrates the usefulness of this sample type. It is widely used in genomic sequencing studies for obtaining high molecular weight DNA and high concentrations of DNA required for next generation sequencing. The starting DNA requirements for 3rd generation sequencing platforms, such as PacBio and Illumina, require 50-1,000 ng of HMW DNA (Quail et al., 2012). Our study found buccal swabs from cattle and tail tissue from water voles, to be an alternative to cell culture as they produced substantial quantities of DNA (a median of 15,129.5 ng and 10,763.5 ng, respectively). All three sample types produced high molecular weight DNA when measured using gel electrophoresis (results only shown for tail tissue in Figure 5).

DNA extraction from hair and faecal samples achieved lower DNA yields (median of 1521.25 ng and 1018.25 ng, respectively; Figure 6), but enough to sequence the mitochondrial control region and within the range of DNA yields required for next generation sequencing. One hair sample (sample 5) and two faecal samples (samples 12 and 18) produced poor sequencing results and were omitted. All samples amplified as viewed on the gel electrophoresis, but PCR clean-up and remaining contaminants in the DNA sample may have caused sequencing inhibition because only 83% of hair and 80% of faecal samples successfully sequenced, suggesting that non-invasive sampling is less reliable than tissue samples for DNA sequencing. The purity of the sample may also limit the success of amplifying longer reads for these sample types, and the quantity of nuclear DNA to mitochondrial DNA and non-host DNA was unknown.

We successfully optimised tail tissue, hair, and faecal DNA extractions, to achieve higher DNA yields and a purer DNA sample. For tail tissue samples buffer 3 (Wang and Storm, 2006) resulted in the highest DNA yield and improved purity scores. This buffer contained a higher concentration of Tris-HCl but lower concentrations of EDTA, NaCl, and SDS than the other two buffers. Using this buffer and the phenol-chloroform extraction achieved better results than ethanol or isopropanol extractions. A longer centrifugation step during DNA precipitation may have improved the binding of DNA to the tube, resulting in less DNA lost when isopropanol was discarded. The toxic nature of phenol-chloroform makes this extraction method less favourable than others, but our study shows it remains the gold-standard for obtaining high quantities of DNA.
For hair samples, the addition of CaCl$_2$ and DTT appears to improve the lysis of hair and hair follicles. It has been previously shown that the addition of Ca$^{2+}$ to hair samples activates the enzyme proteinase K (Walsh, Metzger and Higuchi, 1991; Pfeiffer et al., 2004). DTT is used as an alternative to physical digestion, by acting as both a reducing agent and an anionic detergent (Ghatak, Muthukumaran and Nachimuthu, 2013). The quantities of DNA can vary in hair samples due to the quantity of hair follicles and hair shafts obtained during sampling, with the recommended number of hairs per sample of at least 25 (Henry and Russello, 2011). This is because the hair shaft predominantly contains large quantities of mitochondrial DNA because the nuclei and nuclear DNA in the hair shaft are degraded during keratinization (Linch, Whiting and Holland, 2001; Graffy and Foran, 2005). Due to keratinization, hair shafts therefore contain large amounts of protein that need to be broken down to prevent PCR or sequencing inhibition. Phenol-chloroform extraction performed better when compared to a silica-bead extraction kit, with regards to both DNA yield, but more significantly purity.

Nuclear DNA can be recovered from hair, but this is more successful in hair in the anagen phase, the growth period, when there is more likely to be a root or sheath cells still attached that contain nuclei (Graham, 2007). Often non-invasive sampling obtains hair in the telogen phase, a phase where nuclear DNA is broken down into small fragments. These are often less than 200bp in length and would not be useful for sequencing nuclear markers.

Faecal samples produced the lowest DNA yield, with both extraction kits achieving similar DNA yields. On the other hand, the ‘Qiagen QIAamp DNA Stool Mini Kit’ resulted in a purer DNA sample, whilst absorbance 260/230 values were extremely low when using the ‘Qiagen QIAamp PowerFecal DNA Kit’. Both kits required further DNA precipitation using ethanol in order to increase DNA yield and improve purity. In most samples, DNA yield increased, whilst some samples decreased due to loss of DNA during the process. We show that non-invasive faecal sampling is useful to obtain more samples and is less invasive than other sampling methods, with success at amplifying the mtDNA control region in 80% of faecal samples. But using faecal samples for extracting nuclear DNA, for short tandem repeat (STR) loci, has been shown to result in mismatches between faecal and blood samples of the same individual in a minority of cases (Forgacs et al., 2019). This was mainly caused by allelic dropout in faecal samples. Consequently, care should be taken when using faecal samples in nuclear DNA analyses to undertake rigorous marker selection and to extract large quantities of DNA.

Although we obtained enough DNA to perform follow-up analysis, the percentage of host DNA is still unknown. Contaminants, such as the presence of microorganisms and human DNA on the sample, may have increased the DNA yield measured. Further screening using quantitative polymerase chain reaction (qPCR) and specific host primers to quantify the amount of host DNA
in each sample (techniques used in Baker, 2015), will attain more accurate quantifications of DNA in each sample type, and further improve optimisation of DNA extraction specifically for host DNA.

4.2. Population Genetics of British Water Voles:
The 17 captive water vole samples from Wildwood Trust overall showed considerable genetic diversity (Figure 8; Table 10). A total of 12 haplotypes were found, equating to a high haplotype diversity of 0.949, but a low nucleotide diversity of 0.004. Of the two haplogroups the largest showed considerably low genetic diversity, with a maximum of two mutational steps between each haplotype and three haplotypes with multiple individuals. Whilst the other haplogroup, containing two haplotypes, differed substantially with at least six mutational steps between the haplogroups. Water voles from these two haplotypes may have been captured from different populations than the others sampled. Overall, we find that there is maintenance of genetic diversity in the captive population at Wildwood Trust.

When comparing the captive population with natural populations from the South East and East of England we found tight clustering of haplotypes into one major haplogroup (Figure 9). Only one haplotype (haplotype 14) diverged considerably by 17 mutational steps. This haplotype consisted of samples from Dartford, Kent. The reason for this divergence from all other haplotypes has been hypothesized to be due to breeding programs in the area which may have caused admixture of captive stock and release of this admixed population into the wild (Baker, 2015). The study found that this haplotype clustered with Scottish haplotypes. Admixing between England and Scotland populations will lead to the loss of local genetic heritage. The authors sampled 58 individuals from 13 populations, finding 14 unique haplotypes, with a haplotype diversity of 0.80. This is considerably lower than the captive population at Wildwood Trust. Unfortunately, we were only able to obtain haplotype sequences, so we could not calculate our own haplotype diversity, to allow for a comparison between captive and natural populations in the South East of England.

As shown by our results, the water voles from Wildwood Trust clustered closely together with South East of England haplotypes, with a maximum of three mutational steps between natural and captive individuals, excluding natural water vole haplotypes 14 and 17, which were at least 17 and five mutational steps from other haplotypes, respectively. Only two captive haplotypes (haplotypes 1 and 2) showed divergence from the rest, clustering more closely with the East of England haplotype (haplotype 25) by at least five mutational steps. This may indicate that these water voles were captured from the East of England, rather than more locally in the South East,
or that water voles from the East of England had been translocated to this area in the past. When haplotypes 1 and 2 were removed for the alignment, haplotype diversity was slightly lower at 0.933. This is still considerably high for a declining population. This again demonstrates that genetic diversity is being maintained in the sampled captive population at Wildwood Trust.

Our study finds that all captive individuals are part of the English/Welsh group (Figure 10) when compared with wild populations from England, Wales, and Scotland (Piertney et al., 2005). This is important for the maintenance of the two lineages, as the two populations should be treated as separate ESUs. The release of these captive water voles or their offspring into the South East will not cause any considerable admixture. In all our analyses that include samples from both modern and ancient British water voles, there is support for the English/Welsh and Scottish divergence, as well as two colonisation events into Britain which have shaped current phylogeographic structure (Piertney et al., 2005; Brace et al., 2016).

Constructing a phylogenetic tree of all available sequences showed all Scottish water voles to diverge first, whilst English and Welsh water voles diverge later, forming a separate monophyletic clade (Figure 11). Three outliers were present in the tree. South East of England (haplotype 14) which as mentioned above and in Baker, 2015 has probably been relocated from Scotland and released into the South East of England. Samples collected from Northumberland and Lincolnshire are again likely to be reintroductions, or in the case of Northumberland, geographically close to Scottish populations (Piertney et al., 2005). In the phylogenetic tree the Italian and Swiss samples diverged first and were separate to all other European water voles. The three Arvicola scherman samples were paraphyletic and found in both the English/Welsh and Scottish clades. Based on mitochondrial DNA A. scherman should not be classified as a separate species (Brace et al., 2016), but further analysis is needed. Further analysis is also need for the Italian and Swiss water voles, which may be classified as a separate evolutionary distinct unit with additional molecular analysis.

Baker et al., 2020 found that the diversity of South East of England water voles mtDNA is structured by watersheds. They suggested that conservation management should take place at this local level, with reintroductions ensuring local genetic heritage is maintained. They found evidence using both microsatellite data and mitochondrial haplotypes that reintroductions in the area had left a significant genetic footprint, and this is also demonstrated in our results. Our results suggest that if captive water voles are released back into the South East of England, we expect there to be little outbreeding depression and genetic footprint on local populations, because Wildwood Trust and South East populations are closely related, in terms of the clustering of haplotypes. If water voles are released back into watersheds in the local area, then care should be taken that the populations are not too closely related to the released water voles,
to avoid inbreeding and subsequently inbreeding depression, decreasing genetic diversity further. This demonstrates the importance of considering genetic and evolutionary relationships in conservation management.

4.3. Phylogenetics of Arvicolinae:

Phylogenies of the Genera of Arvicolinae:
The vast amount of genetic data available today provides taxonomists and systematists with a wealth of information for discovering how species have evolved. Our results show the importance of selecting the appropriate genetic marker in phylogenetic studies, as different markers and phylogenetic approaches can result in different phylogenies. By using several markers, both nuclear and mitochondrial, as well as mitochondrial genomes, it allows the comparison of tree topologies and the support for phylogenies that are identical in all phylogenetic constructions.

The availability of DNA sequences is vital to fully resolve phylogenies of groups of species, however the quantity of phylogenetic markers for all Arvicolinae genera is limited to date. Our mitogenome analysis used 10 genera, with up to 30 proposed, resulting in missing information for the 20 other possible genera. Individual markers provided considerably more genera, but only 18 genera were available in all three of the selected markers for the supermatrix tree. However, our analyses do reveal new information and support previous publications on the phylogenetics of Arvicolinae. We use the proposed phylogeny of Arvicolinae by Robovský, Řičánková and Zrzavý, 2008 as a basis for discussion of our results as well as other publications.

Basal Arvicolinae – The phylogeny of the basal arvicoline has long been debated, with no resolved topology. *Hyperacrius, Prometheomys, Ellobius, Eolagurus, and Lagurus* have been proposed as the basal arvicoline, but the relationships between them are unknown (Robovský, Řičánková and Zrzavý, 2008). In our study mitochondrial genomes were unavailable for all the proposed basal arvicolines, so *Ondatra* and *Dicrostonyx* diverged first forming a monophyletic clade (Figures 13 and A1). Whereas in our supermatrix trees (Figures 16 and A10) *Neofiber* diverged first within Arvicolinae, followed by *Prometheomys*, while *Lagurus* and *Ellobius* diverged much later (Figure). This is consistent with other studies which proposed *Lagurus, Eolagurus, and Ellobius* diverged later in Arvicolinae (when sampled) and *Prometheomys* was the first to diverge (Galewski et al., 2006; Abramson et al., 2009; Steppan and Schenk, 2017). *Hyperacrius* was not sampled in any of these studies, so its phylogeny remains unknown.

‘Dicrostonychini’ – The next clade to diverge has been proposed as ‘Dicrostonychini’ and contains *Dicrostonyx, Phenacomys, and Arborimus* (Robovský, Řičánková and Zrzavý, 2008).
Mitochondrial genomes were unavailable for *Phenacomys* and *Arborimus*. *Dicrostonyx* grouped with *Ondatra* (Figures 13 and A1). Our supermatrix trees supported the grouping of *Phenacomys* and *Arborimus* but sequences for *Dicrostonyx* were unavailable to fully support this clade (Figures 16 and A10). Both the GHR (Figures A6 and A7) and IRBP (Figures A8 and A9) phylogenetic trees supported the proposed grouping, with all genera accounted for. Our results are also supported by a more recent publication using multiple mitochondrial and nuclear phylogenetic markers (Steppan and Schenk, 2017).

**Dinaromys, Neofiber, Ondatra, and the clade ‘Lemmini’ containing Synaptomys, Lemmus, and Myopus** – These two clades formed a polytomy in the proposed phylogeny by Robovský, Řičánková and Zrzavý, 2008. Our mitogenome analysis proved ineffective in resolving the phylogeny of this polytomy, due to only *Ondatra* being sampled (Figures 13 and A1). In the supermatrix trees *Neofiber* diverged first in Arvicolinae, before the ‘Dicrostonychini’ clade, disagreeing with the proposed phylogeny (Figure 16 and A10). The grouping of ‘Lemmini’ is supported, with *Lemmus* and *Synaptomys* sampled and grouped together. They were a sister clade to ‘Dicrostonychini’. *Neofiber* and *Ondatra* were grouped together in both the Cytb (Figures A4 and A5) and IRBP (Figures A8 and A9) trees. *Dinaromys* sequences were only found for Cytb gene trees, placing *Dinaromys* as a sister clade to *Arvicola*. All our analyses supported the grouping of ‘Lemmini’, in agreement with Steppan and Schenk, 2017. However, the relationship between ‘Lemmini’ and the other genera (*Dinaromys, Neofiber, and Ondatra*) is unresolved and not consistent in the literature, and our results do not shed any light on this.

‘Clethrionomyini’ – This is a highly-supported clade containing genera *Eothenomys*, *Myodes*, and *Alticola*, with the clade name proposed in (Robovský, Řičánková and Zrzavý, 2008). All our analyses highly support the grouping of *Eothenomys* and *Myodes*, and *Alticola*, when sampled, and is supported by more extensive molecular studies (Steppan and Schenk, 2017).

‘Arvicolinii’ – The placement of *Arvicola, Lemmiscus, Stenocranius*, and *Chionomys* within this clade has not been fully resolved. The four genera have been proposed to diverge after ‘Clethrionomyini’ and first within the ‘Arvicolinii’ clade, before all other genera (Robovský, Řičánková and Zrzavý, 2008). *Stenocranius* was not sampled in any of our trees and is underrepresented in the literature. In our mitogenome trees, only *Arvicola* is sampled, which diverged after ‘Clethrionomyini’ taxa and before *Proedromys* (Figures 13 and A1). The supermatrix trees contain both *Arvicola* and *Chionomys*, with the former of these genera diverging first and separately. Each of our gene trees showed different topologies for these four genera (Figures 16 and A10). In the Cytb trees *Arvicola, Dinaromys*, and *Lemmiscus* formed a monophyletic group, after ‘Clethrionomyini’ taxa, whilst *Chionomys* diverged later (Figures A4 and A5). Whereas in the GHR trees *Chionomys* diverged first, followed by *Arvicola* (Figures A6
and A7). Arvicola diverged first in the IRBP tree, then Chionomys (Figures A8 and A9). We suggest, based on our results and the literature, that Arvicola (including water voles) may be the basal ‘Arvicolini’ and diverges separately from other genera.

The remaining proposed genera pertaining to ‘Arvicolini’ are Microtus, Neodon, Alexandromys, Mynomes, Lasiopodomys, and Terricola (Robovský, Řičánková and Zrzavý, 2008). It has been found that Microtus is paraphyletic in this group and therefore the nomenclature of genera in this group should be changed to Microtus to reflect monophyletic relationships, while others propose that the remaining genera should be re-classified as sub-genera within the genera Microtus (discussed in Robovský, Řičánková and Zrzavý, 2008). In all our trees Microtus was paraphyletic and Terricola was always grouped within Microtus taxa. We therefore support the reclassification of Terricola to Microtus.

In the mitogenome trees Lasiopodomys and Neodon taxa were also grouped with Microtus (Figures 13 and A1). In the supermatrix tree Alexandromys is grouped with Microtus (Figures 16 and A10). Neodon, Volemys, and Lasiopodomys formed a separate sister clade and this was also true for the GHR gene trees (Figures A6 and A7). Phylogenies of this group in individual gene trees differed due to the presence or absence of genera. In our Cytb trees (Figures A4 and A5) Microtus taxa were in a clade alongside Terricola, Blanfordimys, Alexandromys, Neodon, and Volemys taxa and in our IRBP gene trees (Figures A8 and A9) Terricola, Lasiopodomys, Proedromys, Neodon, Volemys, Proedromys, and Alexandromys taxa formed a clade with the available Microtus taxa. Our results were not consistent enough to make any further conclusions about this group.

**Other Genera – Proedromys, Volemys, and Blanfordimys.** The position of Proedromys was unresolved in our analyses, with different phylogenies for each of the different markers. The mitogenome trees placed it between Arvicola and the remaining sampled ‘Arvicolini’ genera (Figures 13 and A1). Robovský, Řičánková and Zrzavý, 2008 proposed that the genus was grouped with Lasiopodomys within ‘Arvicolini’, whereas another study (Steppan and Schenk, 2017) found that the genus groups with selected Volemys species.

For Volemys the nuclear makers showed the genus to be paraphyletic, with Volemys musseri grouping with Proedromys bedfordi, whereas Volemys millicens grouped with the remaining sampled Neodon spp. The latter grouping is also supported by the supermatrix trees (Figures 16 and A10), but Proedromys bedfordi was not sampled to support the former clade. A recent publication showed Volemys to be paraphyletic, with some Volemys species grouping with Neodon and others grouping with Proedromys (Steppan and Schenk, 2017). We support the
notion that Volemys is paraphyletic, and more support may be achieved with sequencing the mitochondrial genome.

The phylogeny of Blanfordimys within ‘Arvicolini’ is relatively supported in the literature when the genus is sampled. Our mitogenome (Figures 13 and A1) and supermatrix (Figures 16 and A10) phylogenetic trees do not provide any information regarding this genus due to a lack of sequences. More extensive sequencing of Blanfordimys is needed to resolve its phylogeny.

**Summary** – Using both the mitochondrial genome and supermatrix approach for all available genera, we support previous studies and propose new phylogenies for genera within Arvicolinae. Our results showed support for Prometheomys as a basal arvicoline, along with Neofiber. Our results disagree with the basal positioning of Ellobius, Eolagurus, and Lagurus by Robovský, Řičánková and Zrzavý, 2008. Instead we support the phylogenies of Abramson et al., 2009 and Steppan and Schenk, 2017. We fully support the grouping of Dicrostonyx, Arborimus, and Phenacomys into the clade ‘Dicrostonychini’, as well as the grouping of Eothenomys, Myodes, and Alticola into the clade ‘Cletherionomyini’. We support the positioning of Arvicola within the clade ‘Arvicolini’ and our mitogenome analyses suggest it may be the earliest to diverge within this clade, followed by Proedromys. There was not sufficient sampling to resolve the phylogenies of the remaining genera.

**Nomenclature and Classification** – Our analysis of the mitochondrial genome of Arvicolinae species, although limited by the low numbers of available taxa, provides support for the paraphyly of Microtus. We demonstrate high support, based on both high bootstrap scores and posterior probabilities, for Terricola, Neodon, and Lasiopodomys diverging with Microtus species. Therefore, based on the mitochondrial analysis alone, the nomenclature of these genera should be reflected by their evolutionary history at the genus or sub-genus level. Further support is provided for renaming of the genus Terricola, which was found to group with Microtus species in all phylogenetic trees. The supermatrix trees also support the paraphyly of Microtus with Alexandromys, but Neodon, Volemys, and Lasiopodomys group into a monophyletic clade, providing less support that the nomenclature of these genera should be change. An increased sampling of taxa is needed to resolve this.

**Comparison of Phylogenetic Approaches:**
The use of molecular markers has revolutionised the field of phylogenetics, greatly improving our understanding of the relationships between species. Approaches include using individual gene markers, using mitochondrial or nuclear genomes (phylogenomics), and approaches which combine multiple markers (i.e. concatenating sequences and constructing supermatrix trees).
However, choosing which marker or approach to take remains pivotal for achieving resolved species phylogeny.

Our results demonstrate the importance of selecting appropriate phylogenetic markers as we have considerable differences in phylogeny between markers and approaches. The main issue was the availability of genetic markers for all genera and species within Arvicoline. Sampling more taxa often greatly improves phylogenetic accuracy (Zwickl and Hillis, 2002), and provides further information. The other issue is the choice of phylogenetic marker and the use of the supermatrix approach. Differences between individual gene trees and the actual species tree can be caused by hybridisation, incomplete lineage sorting, and gene duplications (Maddison, 1997). This can often result in inaccurate phylogenies when individual genes are used.

The supermatrix approach allows for all characters to be included in a phylogenetic analysis, increasing the information available. For our concatenated sequence we only used taxa with sequences available for all three phylogenetic markers. This was to avoid missing data during the phylogenetic tree construction. However, missing data might not negatively impact the supermatrix approach (Wiens, 2006), therefore increasing the number of taxa available might have improved phylogenetic support and provided more information from our supermatrix tree.

We demonstrate that constructing phylogenies using Arvicolinae mitochondrial genomes greatly improves bootstrap and posterior probabilities, compared with using individual genes or concatenated sequences of a handful of markers. All posterior probabilities were 1.00 in the Bayesian phylogenetic tree and relatively high (>70) for most nodes in the ML phylogenetic tree. This shows the importance of sequencing the mitochondrial genome for all Arvicolinae species and may prove useful in further resolving the family’s phylogeny.

4.4. Future Work:
The future of water voles in Britain relies on the maintenance of population numbers and genetic diversity. Future work could include more extensive molecular studies, such as exploring the genetic diversity of whole mitochondrial genomes with additional samples from both captive and natural populations throughout Britain. Moreover, using nuclear DNA of captive water voles, such as microsatellites, will allow for other genetic diversity indicators to be determined, such as heterozygosity and fixation indices. This will provide a more thorough investigation into the genetic diversity of water voles in captivity. If funds were available, using population genomics to extensively assess population structure of both captive and natural water voles in Britain using techniques such as RADSeq or whole-genome sequencing. This would increase the number of markers in the study and increase confidence in the results, which would then
improve the management of the species. Using these techniques would also help to uncover the relationships between Arvicola species and their classification as ESUs. Our study provides an excellent framework for future work in this field.

To resolve the phylogeny of genera and species within Arvicolinae a more extensive sampling of taxa and genetic markers is required. The reduction in sequencing costs will allow for more species to be sequenced in the future, as well as additional genetic markers for each species. Phylogenomics is becoming more widely used with the further sequencing of non-model animals. This would increase the number of informative sites to infer phylogeny, which would prove useful to fully support relationships in this subfamily. Sequencing mitogenomes for all species would be a good start and following this whole-genome sequencing. This would also allow for comparative and functional genomics of the subfamily.

5. Conclusion:

DNA was successfully extracted from various animal sample types and DNA extraction protocols were optimised to achieve high DNA yields and pure DNA samples. We found that cell culture achieved the highest DNA yield, demonstrating its usefulness in providing high quantities of high-quality DNA. Our optimised protocol for buccal swabs and tail tissue also provided considerably high quantities of DNA. We show that DNA from all three samples types can be obtained relatively quickly, easily, and are within the quantity range needed for next-generation sequencing. Hair and faecal samples provided adequate amounts of DNA for amplifying the mitochondrial control region for water voles at Wildwood Trust. But it appears that these samples were more prone to sequencing inhibition, as they contain more contaminants. Their use in sequencing nuclear DNA was not tested in this study but is an area that needs to be explored further.

We show that using non-invasive or non-destructive sampling can provide adequate quantities of DNA for sequencing mitochondrial DNA for population genetic studies. Captive water voles at Wildwood Trust were randomly sampled to better understand their genetic diversity and population structure. We found that the captive population had considerable genetic diversity, in terms of the clustering of haplotypes and relatively high haplotype diversity. When comparing with natural populations, we show that the captive population had maintained genetic diversity, with only a slight decrease in haplotype diversity. Water vole haplotypes closely clustered with natural water voles from the South East and East of England. We proposed that the release of captive water voles back into the South East of England would not cause significant loss of local
genetic heritage. Wildwood Trust water voles were all found within the English and Welsh clade, with Scottish water voles diverging separately. Care should be taken to maintain the genetic distinctiveness of these two groups of water voles in Britain, and as others have found, a few water vole populations in England are more closely related to Scottish populations. This could lead to admixture between the England/Wales and Scotland water voles, which is concerning and should be prevented.

The complete phylogenetics of species within Arvicolinae remains unresolved. Our study provides support for the phylogeny of several genera, such as the paraphyly of Microtus and the grouping of Microtus with other arvicoline genera, such as Terricola, Lasiopodomys, and Neodon, in our mitogenome analyses. We therefore support the renaming of these genera to reflect evolutionary relationships. We also propose that the genus Arvicola, containing the European water vole, is the basal ‘Arvicolini’, with evidence from the mitogenome and supermatrix analyses. However, lack of extensive taxon sampling for all markers and mitochondrial genomes resulted in our analyses lacking crucial genera to fully resolve the phylogeny of this subfamily. Moreover, we show support for using the mitochondrial genome as a marker for Arvicolinae phylogenetics, with a fully resolved tree and high bootstrap scores and posterior probabilities that individual markers and propose that sequencing more mitogenomes will help to fully understand the evolutionary relationships of Arvicolinae.
References:


Canu, A. et al. (2013) ‘Influence of management regime and population history on genetic diversity and population structure of brown hares (Lepus europaeus) in an Italian province’,


McNeely, J. A. et al. (1990) Conserving the world’s biological diversity. International Union for conservation of nature and natural resources.


and the Subcommission on Quaternary Stratigraphy (International Commission on

Walsh, P. S., Metzger, D. A. and Higuchi, R. (1991) ‘Chelex 100 as a medium for simple extraction


(Arvicolinae, Rodentia) from China and its phylogenetic analysis’, *Mitochondrial DNA Part B:

Figure A1: Bayesian inference phylogenetic tree of Arvicolinae mitogenomes. Posterior probabilities shown as percentages.
Figure A2: ML phylogenetic tree of selected taxa for the Cytb gene. Bootstrap score shown.
Figure A3: Bayesian inference phylogenetic tree of selected Arvicolinae taxa for Cytb gene.

Posterior probabilities shown as percentages.
Figure A4: ML phylogenetic tree of all available Arvicolinae taxa for the Cytb gene. Bootstrap scores shown.
Figure A5: Bayesian inference phylogenetic tree of all available Arvicolinae taxa for Cytb gene. Posterior probabilities shown as percentages.
Figure A6: ML phylogenetic tree of Arvicolinae for GHR gene.
Figure A7: Bayesian inference phylogenetic tree of Arvicolinae for GHR gene. Posterior probabilities shown as percentages.
Figure A8: ML phylogenetic tree of Arvicolinae for IRBP gene.
Figure A9: Bayesian inference phylogenetic tree of Arvicolinae for IRBP gene. Posterior probabilities shown as percentages.
Figure A10: Bayesian inference phylogenetic tree of the concatenate for Cytb, GHR, and IRBP genes. Posterior probabilities shown as percentages.