European phylogeography and genetic structure of wood and yellow-necked mice *Apodemus sylvaticus* and *Apodemus flavicollis* based on whole-genome, high-density genotyping by restriction-site-associated DNA sequencing (RAD-seq)

Author:

Maria Luisa MARTIN CEREZO

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree Doctor of Philosophy.
The University of Huddersfield

in the

School of Applied Sciences

Supervisor:

Dr. Jarek BRYK

18th February, 2019

Declaration of Authorship

I, Maria Luisa MARTIN CEREZO, declare that this thesis titled, "European phylogeography and genetic structure of wood and yellow-necked mice *Apodemus sylvaticus* and *Apodemus flavicollis* based on whole-genome, high-density genotyping by restriction-site-associated DNA sequencing (RAD-seq)" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Mª Luisa Martín Cerezo

Date: 20th June 2019

UNIVERSITY OF HUDDERSFIELD

Abstract

Department of Biological and Geographical Sciences
School of Applied Sciences

Doctor of Philosophy

European phylogeography and genetic structure of wood and yellow-necked mice *Apodemus sylvaticus* and *Apodemus flavicollis* based on whole-genome, high-density genotyping by restriction-site-associated DNA sequencing (RAD-seq)

by Maria Luisa MARTIN CEREZO

Rodents of the genus *Apodemus* are one the most common mammals in the Palaearctic region. They play an important role in ecosystems by participating in seed dispersal and being a part of the diet of many carnivores. They contribute to the spread of human diseases such as Lyme disease and tick–borne encephalitis, are a reservoir of hantaviruses that cause hemorrhagic fever with renal syndrome (HFRS) and exhibit interesting karyotype feature – B chromosomes. They are, however, very underdeveloped in terms of genetic and genomic resources available for their study.

Apodemus flavicollis and *Apodemus sylvaticus* live in sympatry in the European Plain. They are phylogenetically related and exhibit similar behaviour and morphology. They have long been studied for elucidation of post-glaciation migration patterns where previous studies using microsatellite and mtDNA markers revealed glacial refugia in southern Europe and suggested the possibility of the existence of a northern refugium. Here, we employ double digestion restriction site-associated DNA sequencing (ddRAD-seq) to study *Apodemus* phylogeography in Europe.

I first established the feasibility of this approach in a pilot study with 82 samples from both species (72 *Apodemus flavicollis* and 10 *Apodemus sylvaticus*) from four locations spanning 500 km in north-eastern Poland. My results shown that despite presumed relatively low mobility of the species, *A. flavicollis* in the north-eastern Poland effectively constitutes a single population with neligible structure and moderate heterozygosity. Based on 21377 common loci, I was able to estimate the average genetic divergence between the two species at 1.51% and an evolutionary rate of 0.0019 substitutions per site per million of years. I also generated a catalogue of 632063 loci to enable clear genetic differentiation of the two species, and successfully verified its performance on 20 unrelated samples from Europe and Tunisia.

Based on the pilot project experience, I developed a new library preparation protocol that incorporated longer barcodes and degenerate base regions to allow detection of both PCR duplicates and chimeric sequences. After testing the efficiency of the new protocol on a set of samples with variable DNA quality, I applied it to a large scale pan-European study of *Apodemus* in one of the first application of the RAD-seq in mammals. My results show, for the first time, the existance of post-glacial northern groups, not only on *A. sylvaticus* but also on *A. flavicollis*, as well as long distance movements on *A. sylvaticus* but not *A. flavicollis*.

This thesis constitutes the first application of a whole-genome approach to study these organisms. It has allowed us to generate sequences from thousands of loci for both species, identify tens of thousands of SNPs markers and perform continental-scale analysis of the relationships between multiple populations, contributing to the development of *Apodemus* as a model organism.

Acknowledgements

First of all, I would like to express my gratitude to my advisor Dr Jarek Bryk, for giving me the chance to be part of his new group and trust me at all times. Without your support, advice, motivation, and patience, my time in Huddersfield would have been very different. I couldn't have imagined having a better supervisor than you. I will always miss being part of Bryklab.

Furthermore, I would like to thank Dr. Frank Chan and Dr. Marek Kucka, for accepting me on their lab and teaching me everything I know about library preparation and sequencing. I also thank all our collaborators: Dr Johan Michaux (University of Liège, Liège, Belgium), Dr Jerry Herman (National Museums of Scotland, Edinburgh, Scotland), Dr Joana Paupério (CIBIO-InBIO, University of Porto, Porto, Portugal), Dr Vladimir Jovanović (Institute for Biological Research "Siniša Stanković", Belgrade, Serbia), Dr Douglas J. Clarke (University of Huddersfield, Huddersfield, United Kingdom), Dr Karol Zub (Mammal Research Institute, Białowieża, Poland) and Dr Barbara Tschirren (University of Exeter, Devon, United Kingdom). Without all your field work and generosity sharing your appreciated samples, this project would have not been possible. I will also like to thank Dr. Allan McDevitt (University of Salford, Manchester, United Kingdom) for his inestimable help during this project and his patience with the DIYABC analysis.

I am also very grateful that I have such an amazing office group. You made Huddersfield feel like home. A special thanks to Marina Soares da Silva, with whom I spent an “amazing” time organising the 4th PEB conference (thanks again Jarek for supporting us on this unexpected but grateful job), to Rohan Raval, for all his help and friendship since his arrival to Bryklab and to Juan José Ginés, for his affection and support.

A special mention deserves my family, for their unconditional support, comprehension, and help and for the long hours spent on Skype to let me see how the new generation is growing up.

And... last but not least, a special thanks to Bernardo de Haro Reyes, for your unconditional support during my whole career. Without you by my side, I would have never considered moving abroad. You have been the best office mate that I could have ever imagined.

Contents

List of Figures

List of Tables

List of Abbreviations

A. flavicollis	*Apodemus flavicollis*
A. sylvaticus	*Apodemus sylvaticus*
AT	Austria
B	Billion
BE	Belgium
Bial	Białowieża
Bory	Bory Tucholskie
bp	Base pair
COX1	Cytochrome c oxidase subunit I
CytB	Cytochrome b
DAPC	Discriminant Analysis of Principal Components
ddRAD-seq	Double Digestion Restriction site Associated DNA sequencing
DE	Germany
DK	Denmark
DNA	Deoxyribonucleic acid
EISC	Eurasian Ice Sheet Complex
EN	England
$\mathbf{F_{IS}}$	Inbreeding coefficient
FR	France
IT	Italy
GBS	Genotyping By Sequencing
GR	Greece
Hack	Haćki
HFRS	Hemorrhagic Fever with Renal Syndrome
HGP	Human Genome Project
$\mathbf{H_e}$	Expected heterozygosity

H_o	Observed heterozygosity
HWE	Hardy-Weinberg equilibrium
IE	Ireland
IS	iceland
IRBP	Interstitial Retinol Binding Protein
IT	Italy
IUPAC	International Union of Pure and Applied Chemistry
Ka	Thousand years ago
Kadz	Kadzidło
kbp	Kilo base pair
Km	Kilometers
Kyr	Thousand years
LGM	Last Glacial Maximum
LT	Lithuania
Ma	Million years ago
maf	Minor Allele Frequencies
MK	North Macedonia
mtDNA	Mitochondrial DNA
MRI	Mammal Research Institute of the Polish Academy of Sciences
NAD1	NADH dehydrogenase subunit I
ng	Nanogram
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
NO	Norwey
nt	Nucleotide
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PL	Poland
PT	Portugal
RAD-seq	Restriction site Associated DNA sequencing
RAG1	Recombination Activating Gene 1
RNA	Ribonucleic acid
RO	Romania
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	Ribosomal RNA
RS	Serbia
RU	Russia

SC	Scotland
SK	Slovakia
Sl	Slovenia
SNP	Single-Nucleotide Polymorphism
SP	Spain
SE	Sweden
tRNA	Transfer RNA
TN	Tunisia
WL	Wales
µl	Microliters

Chapter 1

Introduction

1.1 European Phylogeography

1.1.1 The science of phylogeography

Different geological processes, such as climate changes, volcanism and/or orogenesis, also known as mountain formation, can dramatically shape the distribution of species and their intraspecific diversity. Phylogeography studies the principles and processes responsible for the geographic distribution of organisms' genealogical lineages using tools and approaches from molecular genetics, population genetics, phylogenetics, palaeontology, geology and historical geography data, among others (Avise, 2000). Despite this concept being defined for the first time in 1987 (Avise *et al.*, 1987), the interest on this topic is still growing, achieving over 1000 publications per year since 2012 (Web of Science, 2018).

1.1.2 Climate changes during the Cenozoic

The geographical distribution of species is largely influenced by changes in the climate (Schmitt 2007). Since the Paleocene-Eocene Thermal Maximum, which occurred around 56 Million years ago (Ma) (Figure 1.1), when temperatures were 8°C warmer than today, the climate has been cooling down, allowing the formation of the Antarctic ice sheet around 34 Ma. (DeConto and Pollard, 2003).

The ice volume in Antarctica and the northern Hemisphere further increased (Raymo and Ruddiman, 1992) during the middle Miocene (15 Ma) and late Pliocene (2 Ma). During the Pleistocene epoch of the Quaternary

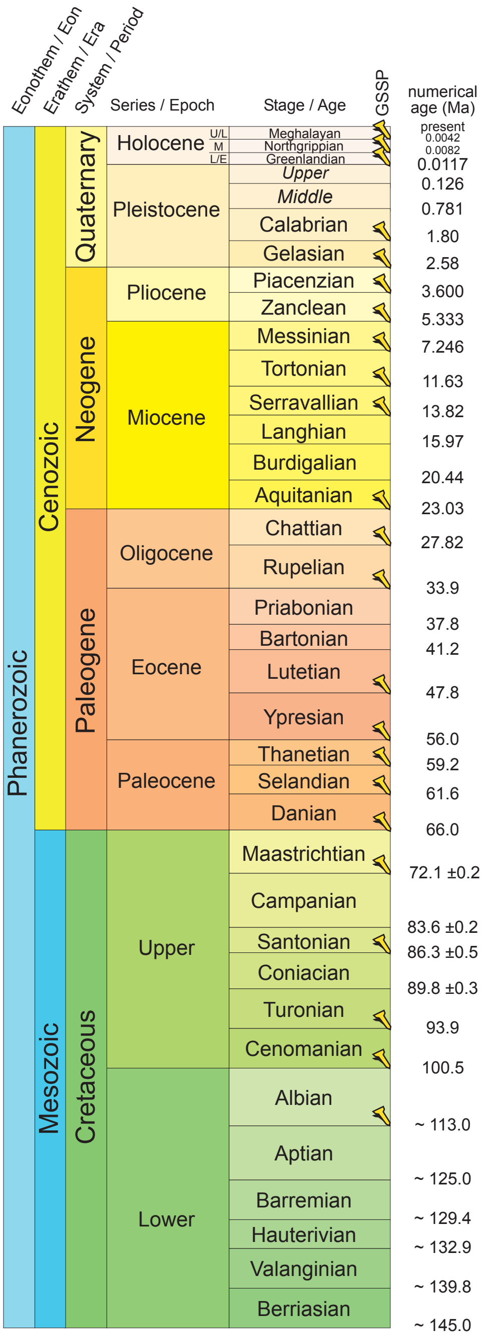


FIGURE 1.1: International Chronostratigraphic chart from the International commission on Stratigraphy. Definition of Eons, Eras, Periords, Epochs and Ages for stratigraphic successions on a global scale. Times, in Ma, are indicated for the boundaries between different Ages. Imaged modified to include only the last 145 Ma.

period, a series of cold and warm climatic periods occurred (Schmitt, 2007), starting from approximately 2.3 Ma and with a periodicity of 100 Thousand years (Kyr) (Hewitt, 1999). The Last Glacial Maximum (LGM), also known as the Weichselian glaciation, occurred in Europe between 26.5 Thousand years ago (Ka) and 19 Ka (Clark *et al.*, 2009).

During that time, northern Europe was covered by the Eurasian ice sheet complex (EISC), spanning over 4500 kilometers (Km) (Patton *et al.*, 2017). Glaciers occupied Iceland, part of the British Isles, northern Europe up to Germany and Poland and all the main mountain ranges in the continent, including the Alps, Pyrenees, Cantabrian, Carpathians and Caucasus (Figure 1.2). A large part of central Europe was occupied by tundra: polar desserts, where the growing season is extremely short and the vegetation is reduced to dwarf shrubs, lichens, mosses, sedges and grasses. In the south, tundra biomes shifted to stepped areas: semi-arid grassland plains. Forests were restricted to the southern European peninsulas, the Caucasus and the Carpathians, regions where most of extant species in Europe survived the Quaternary glaciations (Hewitt, 2000).

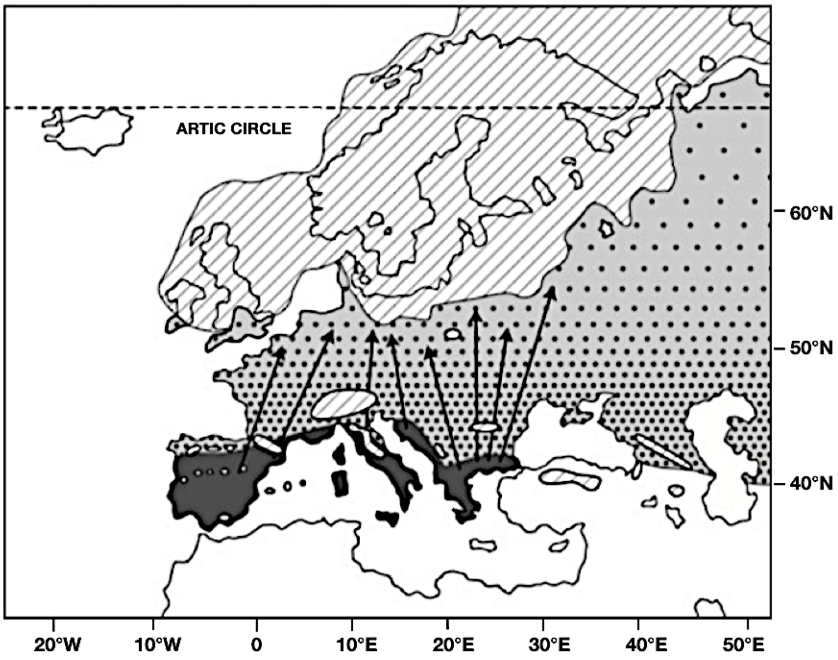


FIGURE 1.2: Last Glacial Maximum in Europe: Extension of the Ice sheet during the LGM (hatched) and permafrost-tundra (light grey). Glacial refugia in the southern European peninsulas appear in dark grey while dark grey dots on light grey background suggest the extension of a continuous gradient of northern refugia. From Bhagwat and Willis (2008)

1.1.3 European refugia during the Quaternary glaciations

The biome distribution during the LGM, and during all the Quaternary glaciations, determined the suitability of different regions for the survival of various species. Moreover, the geography of Europe affected the displacement of species in response to climate change (Hewitt, 1999). In the south, in an east-west orientation, the Mediterranean and Black Sea act as strong barriers, limiting the movement of organisms during cold periods (Taberlet *et al.*, 1998; Hewitt, 1999). In addition, mountain ranges, such as the Pyrenees and the Alps, also limit the expansion to the north during warm periods (Taberlet *et al.*, 1998). Most of the species inhabiting Europe during the Quaternary glaciation cycles experienced extinctions of their northern populations but survived in areas in the south of Europe, where the conditions were less severe, and recolonised northern Europe at the end of the LGM (Michaux *et al.*, 2004).

These areas, known as refugia, are largely Iberian, Apennine and the Balkan peninsulas (Feliner, 2011). Species that have survived the Pleistocene glaciations in these refugia are: the field vole or short-tailed vole, *Microtus agrestis* (Jaarola and Searle, 2002), the wood mouse, *Apodemus sylvaticus* (Michaux *et al.*, 2003), the Scots pine, *Pinus sylvestris* (Cheddadi *et al.*, 2006), the yellow-necked mice, *Apodemus flavicollis* (Michaux *et al.*, 2004) and the European hedgehog, *Erinaceus europaeus* (Seddon *et al.*, 2001).

During the last decade, different studies have shown the suitability of northern regions, regarding to the Pyrenees, to support the survival of small populations of animals and plants during the Quaternary glaciations. Some of the organisms that have survived in these northern regions, also known as cryptic northern refugia (Stewart and Lister, 2001), typically near the Caucasus, the Carpathians or the Caspian Sea (Hewitt, 1999; Hewitt, 2000; Hewitt, 2011) are: the Eurasian pygmy shrew, *Sorex minutus* and the common shrew, *Sorex araneus* (Bilton *et al.*, 1998), the bank vole, *Myodes glareolus* (Kotlík *et al.*, 2006; Wójcik *et al.*, 2010) and the short-tailed vole, *Microtus agrestis* (Jaarola and Searle, 2002).

Turkey and the Near East could have also acted as refugia for some European species, such as the wood mouse, *Apodemus sylvaticus* (Michaux *et al.*, 2003), and the southern white-breasted hedgehog, *Erinaceus concolor* (Santucci *et al.*, 1998).

It is quite unlikely that these peninsulas acted as a single continuous refugium during the Pleistocene glaciations. Different populations, with different degrees of genetic structure, could have existed inside those refugia, making the interpretation of phylogeographic signals harder. Indeed, Gomez and Lunt (2007) have shown that the Iberian refugium was itself composed of multiple separate glacial refugia, suggesting the term "refugia within refugia" to describe these regions. The "refugia within refugia" can, mistakenly, appear to support the existence of northern refugia in cases of inadequate sampling from the southern regions.

1.1.4 Europe after the Last Ice Age

The geography of Europe at the beginning of the deglaciation period was markedly different from today. During the glaciations, the water that had accumulated in glaciers caused the oceans to retreat and former marine environments to emerge, increasing the amount of land in Europe by 40% (Harff *et al.*, 2015). Doggerland (Figure 1.3), a mass of land in the North Sea connecting the British Isles with the northern Europe, facilitated the movement of species into areas that are, nowadays, less accessible, such as the British Isles (Montgomery *et al.*, 2014).

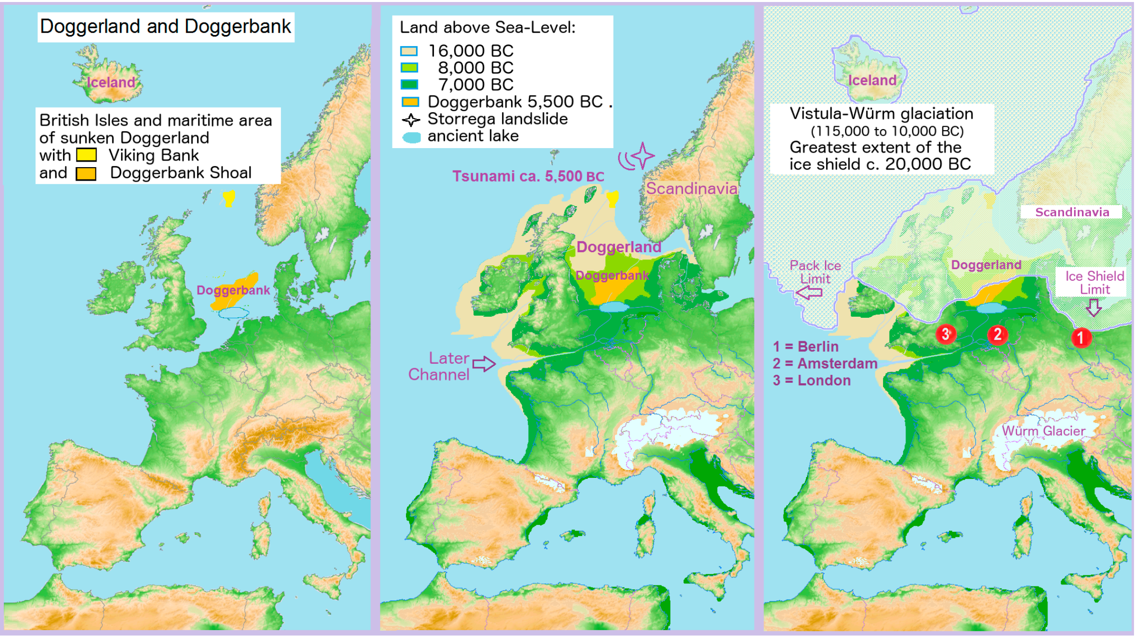


FIGURE 1.3: Changes in landmass around Doggerland from the Weichselian glaciation to the present day. By Francis Lima [CC BY-SA 4.0]

The ice retreat accelerated between 15-13 Ka, when the Eurasian ice sheet is estimated to have been loosing 750 cubic kilometres of ice per annum, with maximum of 3000 cubic kilometres of ice per year (Patton *et al.*, 2017). Doggerland flowded between 12 Ka and 6 Ka (Coles, 2000), isolating the British Isles from the rest of the continent.

1.1.5 Genetic consequences of postglacial expansions

The contemporary distribution of genotypes can shed light on the historical dispersal processes. Randi (2007) compiled the different phylogeographic patterns that have been observed on the recolonization of Europe, and other territories, after the Pleistocene glaciations, by different species.

The most common pattern observed is the one produced by south-north movements (Randi, 2007).

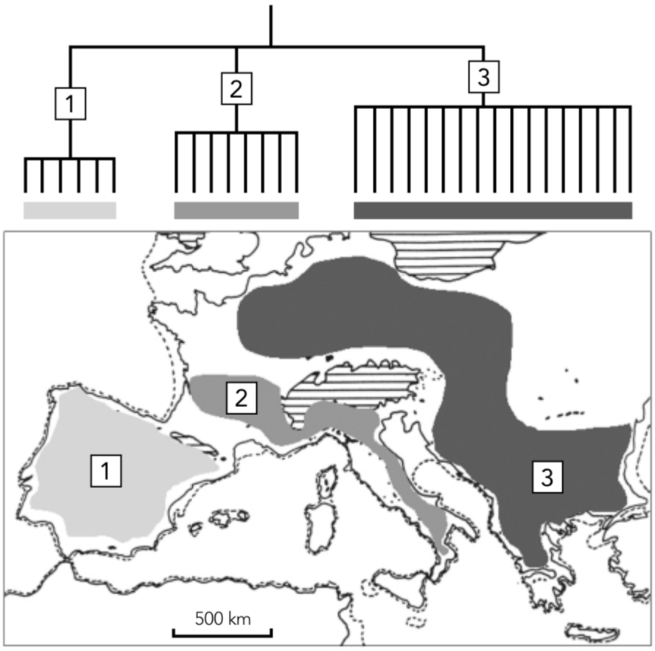


FIGURE 1.4: First recolonization scenario: Southern European peninsulas. Postglacial colonisation from populations that survived and differentiated in southern European refugia: 1-light grey: Iberian peninsula, 2-grey: Apennine peninsula and 3-dark grey: Balkan peninsula. Populations from different refugia can follow different dispersal routes and contribute in different ways to extant European populations. This model is expected to produce genetics trees with a clear phylogeographic structure. From Randi (2007)

During interglacial periods, populations from the northern limits of their distribution range would have expanded northwards, colonising new suitable territories that were previously occupied by tundra (Figure 1.4). This first colonisation probably took place by long distance dispersers, who were the first arriving in new territories and colonise them. Hewitt (2000) suggested that multiple founder effects could have occurred along the expansion route, producing a loss of alleles, increasing homozygosity and reducing the genetic diversity. New waves of founders could have arrived in regions of reduced variability. However, the genetic contribution of the newcomers to the already established population, as well as their capacity to keep colonising new environments, could have been low due to the high density of the established population.

Populations expanding their distribution range and colonising new environments were exposed to new conditions that could increase selective pressures and accelerate adaptive processes, making their genomes diverge (Hewitt, 2000). Under this scenarios, all the genetic diversity would be confined to the southern European peninsulas, were multiple lineages converged. Ancestral refugia populations will retain the highest genetic diversity (Randi, 2007), due to the concurrence of multiple lineages in the same territory. In contrast, populations from recolonised areas, which had suffered successive bottlenecks, would show lower genetic diversity.

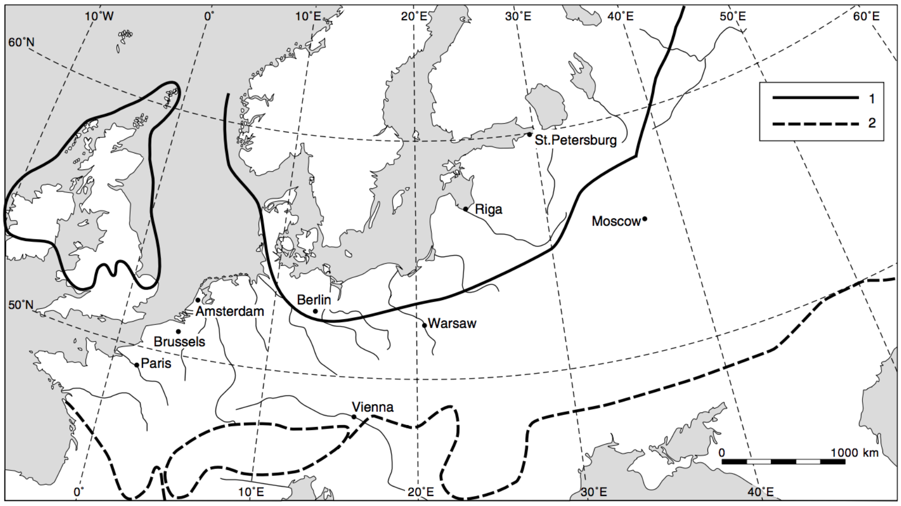


FIGURE 1.5: Mid latitude European sector during the LGM of steppe tundra communities in the European Plain. Area between the northern European ice sheets (full line) and the southern edge of the permafrots (dashed line). From Vandenberghe *et al.* (2004)

Another possibility is that multiple colonisation waves could have occurred from eastern Eurasia into western Europe during interglacial periods. In Asia, during the LGM, ice sheets covered Siberia and Tibet, leaving a corridor of steppe–tundra communities in the European Plain (Figure 1.5) that could be used to colonise Europe during interglacial periods. Individuals from the first colonisation waves, from the East, could have survived to the following glaciation in the southern European peninsulas, where they differentiated and adapted to the climate (Figure 1.6). These individuals might not have contributed to the current central European populations, due to the arrival of a second wave of colonisers from the east. Under this scenario, populations from the southern European peninsulas should be more closely related to one another than to populations from central Europe, which would have originated from a later colonisation wave (Randi, 2007).

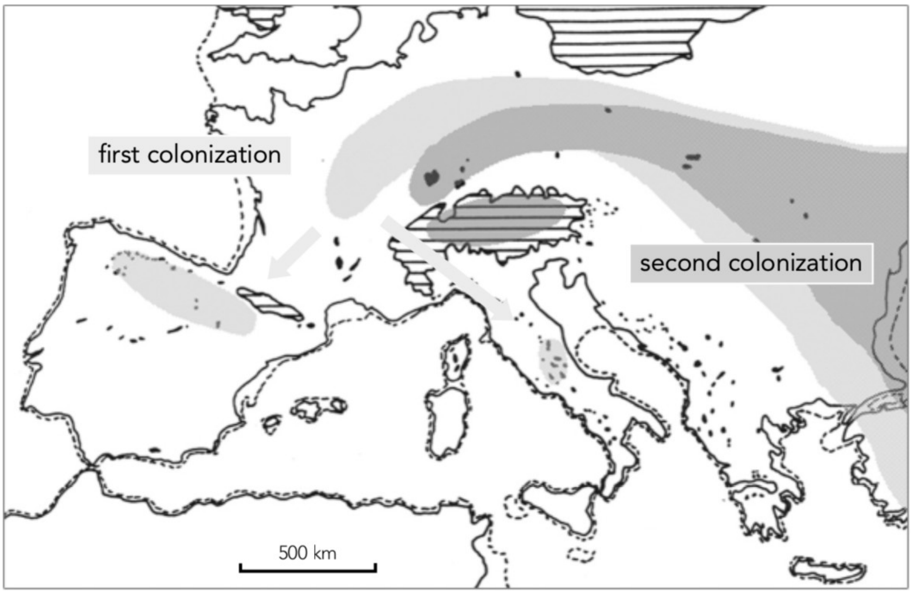


FIGURE 1.6: Second recolonization scenario: Easter recolonization. Colonization of central Europe took place from the expansion of eastern populations during interglacial periods. A first eastern colonisation wave (light grey) colonised central Europe and survived to following glaciations in the southern European peninsulas, differentiating over time. A second wave of eastern colonisers (dark grey) could have occurred in a following interglacial period, limiting the expansion into central Europe of the first colonisers, previously retreated to the southern European peninsulas. From Randi (2007)

The third scenario described by Randi (2007) involves north-to-south changes of the population range (Figure 1.7). Populations would have moved northwards, from their refugia, during interglacial periods. The refugia populations would have become extinct due to changes in the environmental conditions. In this scenario there would not be a bottleneck involved in the recolonization of the continent and the southern European populations would be relicts or they would be derived from a secondary colonisation from the continent.

In this case, the higher genomic diversity would be found on the continent while the European peninsulas would be less genetically diverse.

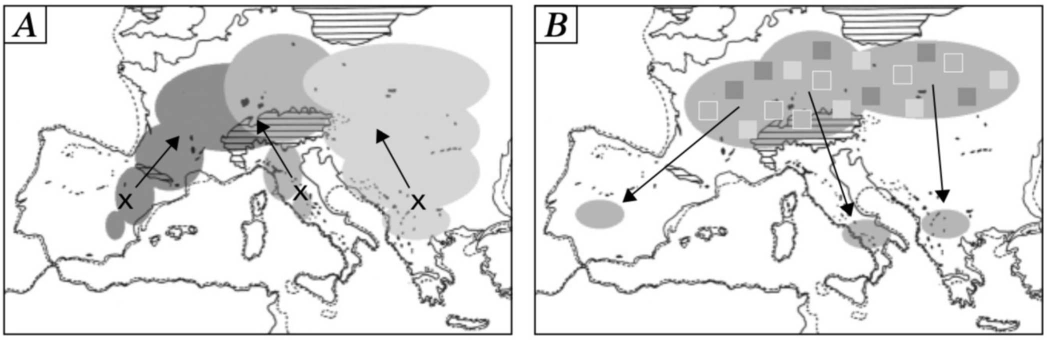


FIGURE 1.7: Third recolonization scenario: North to south movements. (A) During interglacial periods populations from the different refugia (dark grey: Iberian peninsula, grey: Apennine peninsula and light grey: Balkan peninsula) move northwards and the refugia populations went extinct. (B) The expanding populations from the southern European peninsulas admix in central Europe. The colour of the squares represents the origin of the different central European populations, which current distribution range is not longer connected with their phylogeographic story. Current southern European populations will be derived from the central European admix population. From Randi (2007)

Finally, the last phylogeographic pattern described by Randi (2007) is the lack of phylogeographic structure, due to an extensive dispersal with continuous gene flow.

1.1.6 Markers for the study of phylogeographic signals

The Pleistocene ice age occurred between 2.3 Ma and 15 Ka. Genomic and genetic investigations on such timescales requires markers with a relatively fast evolutionary rate for an accurate discrimination of patterns of population movement (Hewitt, 1999). Several different markers have been used to analyse phylogeographic patterns.

1.1.6.1 Mitochondrial DNA (mtDNA)

Probably the most widely used is the mitochondrial DNA (mtDNA), a double-stranded circular molecule, covalently closed, with a length between 15 to 20 Kilobase pairs (Kbp) in animals (Boore, 1999). The average length in mammals is 16.6 Kbp (Gustafsson *et al.*, 2016). This genome codes for genes related to the respiratory chain, electron transport and oxidative phosphorylation (Stuart and Brown, 2006). In mammals, it codes for 37 genes: 13 protein-coding genes, 2 ribosomal RNA genes (rRNA), and 22 transfer RNA (tRNA) (Gibson *et al.*, 2004). Its inheritance is generally maternal and it is characterised by lack of recombination, a high mutation rate and a high copy number in the cells.

These characteristics make this genome suitable for population genetics, phylogeography and phylogenetics studies (Yu *et al.*, 2008). Some of the studies have been performed using complete mitochondrial genomes (Pala *et al.*, 2012; Lippold *et al.*, 2011), cytochrome b sequences (CytB) (Stojak *et al.*, 2016; Stojak *et al.*, 2015; Stone and Cook, 2000; Herman *et al.*, 2017), cytochrome c oxidase subunit I (COX1) fragments (Lunt *et al.*, 1998), NADH dehydrogenase subunit I (NAD1) (Consuegra *et al.*, 2002) and/or the control region (Bernatchez, 2001). Allio *et al.* (2017) showed mitochondrial mutation rates in mammals to be between 0.033 and 0.0655 mutations per million years, values between 10 to 28 times higher than in nuclear markers.

The haploid and non recombinant nature of mtDNA make mutations in this genome to become fixed faster than in a diploid nuclear genomes. This feature enables mtDNA to resolve recent divergence processes, however its capability to resolve older relationships is low due to genetic saturation, a process in which the genetic distance between two groups is reduced due to the effect of reversal mutations or homoplastic changes. Homoplastic sequences can be very similar on their nucleotide composition but have indeed evolved from different ancestral sequences. Despite the general suitability of this genome for phylogeographic studies, it only allows for the female genealogy to be tracked. In species where there is a sex-biased dispersal, as, for example, rodents from the species *Microtus arvalis*, where dispersal is driven by males (Ratkiewicz and Borkowska, 2006), mtDNA can recover a biased history. In this case, it is necessary to use it in combination with Y chromosome in order to analyse the complete phylogeographic history of the species. Furthermore, the lack of recombination on mtDNA make this genome behave as a single locus, whose history can differ from the species or population history.

Moreover, it has been shown that, at least, seven families of bivalves (Theologidis *et al.*, 2008), present other type of mithocondrial inheritance: Doubly Uniparental Inheritance (DUI). DIU is a phenomena in which males inherit the maternal mitochondrial genome (F) and the paternal mitochondrial genome (M) but only transmit the M one, while females can inherit only the F genome or both, M and F genomes, but only transmit the F one. Somatic tissues are, consequently, heteroplasmic in males and can be heteroplasmic in females (Machordom *et al.*, 2015). Biparental mtDNA inheritance have also been described in humans (Luo *et al.*, 2018), where it is known as patternal leakage due to failure of the egg-sperm mitochondrial recognition mechanism (Ladoukakis and Zouros, 2017), but is thought to be extremely rare. mtDNA heteroplasmy by paternal leakage has also been found in chickens (*Gallus gallus)* (Alexander *et al.*, 2015), the Rock Partridge (*Alectoris graeca*), the Chukar Partridge (*Alectoris chukar*) (Gandolfi *et al.*, 2017), fruit flies (*Drosophila melanogaster*)(Nunes *et al.*, 2013), (*Drosophila sumlans*)(Wolff *et al.*, 2013) and in interspecific crosses between female *Mus musculus* and male *Mus spretus* (Shitara *et al.*, 1998).

1.1.6.2 Microsatellites

Microsatellites are fast-evolving nuclear markers that are tandem repeats of 1 to 4 nucleotides (nt), which are distributed throughout the genome of most eukaryotic organisms (Abdul-Muneer, 2014). They are codominat markers with a high polymorphism level (Chistiakov *et al.*, 2006). Microsatellite polymorphism denotes differences in length and not differences in sequence. Their mutation rate is highly variable, changing among loci, lengths, alleles and species (Ellegren, 2004). The difference in mutation rate, in a set of markers from the same organism, can be up to two orders of magnitude, as observed by Ellegren (2004). These characteristics make them suitable for a wide range of studies, from population genetics and phylogeography to medical genetics. Multiple studies have used microsatellites for phylogeographic purposes in a wide range of organism: elephants (Eggert *et al.*, 2002), cattle (MacHugh *et al.*, 1997), *Trypanosoma* (Llewellyn *et al.*, 2009) or fish (Koskinen *et al.*, 2002). Despite their broad use in phylogeographic studies, microsattellites can not determine phylogenetic lineages, one of the main objectives in the field (Edwards *et al.*, 2015). Furthermore, microsatellites can have high levels of homoplasmy, making it difficult to estimate the number of mutation events that have taken place (Edwards *et al.*, 2015).

Work performed with microsatellites usually comprises of two steps: an initial screening of multiple microsatellites and the selection of the most variable ones, which can lead to an overestimation of population genetics parameters. This method is usually time- and cost- consuming when used in a new species or populations, reducing the number of different loci that can be used for a study (Miah *et al.*, 2013).

1.1.6.3 Amplified Fragment Length Polymorphism (AFLP)

Other nuclear markers traditionally used in this field are Amplified Fragment Length Polymorphisms (AFLP). In this method genomic DNA is digested by restriction enzymes and ligated to adapters. Afterwars, fragments containing adapters are amplified by PCR, using primers complementary to the adapter sequence. Fragments can then be visualized through agarose gel electrophoresis or sequenced. This method can only detect dominant markers but it is able to amplify between 50 to 100 fragments in a single PCR reaction (Chial, 2008). The use of a higher number of markers increases the chances to detect diverse history signals trough the genome. Multiple studies have used these markers, either alone (Wang *et al.*, 2003; Pleines and Blattner, 2008; Lambertini *et al.*, 2006) or in combination with other markers (Martínez-Nieto *et al.*, 2013; Creer *et al.*, 2004; Schönswetter *et al.*, 2007). As well as in microsatellites, this method can be time consuming, but the results are more reliable than in other DNA-based techniques (Costa *et al.*, 2016).

1.1.6.4 Restriction-site-associated DNA sequencing (RAD-seq)

Recently, restriction-site-associated DNA sequencing (RAD-seq) protocols have been developed to resolve the phylogeographic history of different species (Hodel *et al.*, 2017; Cao *et al.*, 2018). This method, described in more detail in Section 1.3 can produce hundreds of thousands of loci across the genome located next to restriction-enzyme recognition sites. While many of these loci will be monophormic, and therefore, non informative for the applications described so far, it is possible to obtain thousands of polymorphic loci in the form of single nucleotide polymorphisms (SNPs).

RAD-seq approaches allow sampling of multiple individuals, populations and loci in a time and cost-effective way. This technique surveys a small portion of the genome but targets different regions subject to different selective pressures, such as exons and introns in the nuclear genome or fragments of the mitochondrial genome. No previous genetics knowledge is required in order to apply this technique and the number of possible applications of the generated data is higher than in any of the previously described markers. However, as for other markers, RAD-seq approaches also present drawbacks. Mutations in the restriction sites used for digestion of the DNA can introduce a bias called allelic dropout. Some alleles will not be sequenced due to those mutations and the corresponding SNPs will no be observed, which could cause an overestimation of the genetic diversity (Gautier *et al.*, 2013). Comparisons between traditional methods and RAD-seq have shown that RAD-seq provides a much higher resolution and is able to detect even a fine scale structure among populations (Jeffries *et al.*, 2016) and resolve phylogeographic breaks not identified by other markers (Hodel *et al.*, 2017).

1.2 On *Apodemus*

1.2.1 Introduction

Rodents are the most prolific order of mammals (Adkins *et al.*, 2001). From 5674 species of mammals (IUCN, 2017), 2283 are rodents, a 40% of them. They live on all the continents with the exception of the Antarctica (Fabre *et al.*, 2012) and can inhabit both natural and man-made environments. Molecular studies revealed that rodents and lagomorphs diverged around 61.7 Ma (52.8–71), shortly after the Cretaceous/Paleogene boundary (Wu *et al.*, 2012). This boundary is associated with the mass extinction event that wiped out most of the Mesozoic species, including the dinosaurs. Rodents have since diversified to fill available niches.

Rodents from the genus *Apodemus (Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Apodemus*) are the most common mammals of the Palearctic region (Michaux *et al.*, 2005). At least 21 different species have been described occupying environments as diverse as Mediterranean, North Africa and Siberia. They are classified into two subgenera:

Apodemus senu stricto, which includes most species from Asia, and *Sylvaemus*, which includes European and Near East species. Based on molecular data, the two subgenera diverge 9.87-7 Ma (Liu *et al.*, 2004; Michaux *et al.*, 2002), 2 to 4 Ma later than the divergence between *Apodemus* and *Mus*, (Chevret *et al.*, 2005).

Apodemus are an important group of rodents not only due to their role in many ecosystems, but also for molecular and medical reasons. They contribute to seed dispersal by moving and hiding seeds (Sunyer *et al.*, 2013) and are an important part of the diet of multiple carnivores, such as the tawny owl (*Strix aluco*) (Luka and Riegert, 2018), the wild cat (*Felis silvestris*) (Piñeiro and Barja, 2011) and pine martens (*Martes martes*) (Kleef and Wijsman, 2015), among others. They present an interesting karyotype feature, B chromosomes (extra-autosomal chromosomes), with unknown heritability and function, that could be related to cellular metabolism, due to the relationship between increase cell size, by the accumulation of noncoding DNA and low metabolic rates (Maciak *et al.*, 2014; Kozłowski *et al.*, 2003). Furthermore, they also contribute to the spread of human diseases like Lyme disease and tick-borne encephalitis (Netušil *et al.*, 2013; Welc-falęciak *et al.*, 2010; Randolph *et al.*, 1999) and are a reservoir of hantaviruses that cause hemorrhagic fever with renal syndrome (HFRS) (Heyman *et al.*, 2009; Klempa *et al.*, 2003).

The two apparently separate species, *Apodemus flavicollis* (Melchior, 1834), also known as the yellow-necked mouse, and *Apodemus sylvaticus* (Linnaeus, 1758), also known as the wood mouse, live in sympatry in the forests and fields of the European Plain (Jojić *et al.*, 2014) (Figure 1.8).

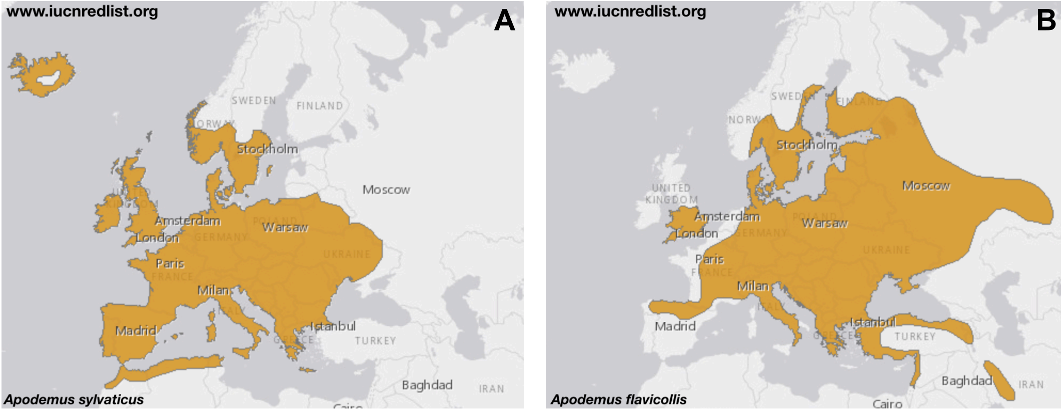


FIGURE 1.8: Distribution maps for *Apodemus sylvaticus* (A) and *Apodemus flavicollis* (B). From IUCNredlist

These two sibling species are phylogenetically related and diverged about 4 Ma (Michaux *et al.*, 2003), probably in an allopatric speciation (Michaux *et al.*, 2005) event from an eastern ancestor that arrived in Europe at the end of the Pliocene (Jojić *et al.*, 2014). In central and northern Europe, the two species are easily distinguishable (Bugarski-Stanojević *et al.*, 2008) by their external characteristics: a yellow collar around the neck of *A. flavicollis* not present in *A. sylvaticus*. However, their differences are reduced in the southern part of their distribution range to such an extent that the species identification in the absence of molecular data relies on craniofacial morphometrics (Bugarski-Stanojević *et al.*, 2013).

Despite the similarities between the two species, there have been no reports of hybridisation between *A.flavicollis* and *A. sylvaticus*. Their relatively long divergence time of 4 Ma makes it unlikely, however on account of their similarity in southern Europe and the lack of information about hybridisation under laboratory conditions, it cannot be completely excluded.

1.2.1.1 Ecology of *Apodemus*

Apodemus sylvaticus and *Apodemus flavicollis* show similar behaviour and morphology, but they differ slightly in their ecology. *A. flavicollis* is considered an edge forest species, associated with deciduous and mixed forest areas with high canopy. It prefers forest with high production of seed, which accounts for the largest source of nutrients along with invertebrates (Juškaitis, 2002; Butet and Delettre, 2011). The home range of this species varies from 100 to 3950 m^2 (Matić *et al.*, 2007) and their mortality during the winter season can reach up to 86% of the spring population (Pucek *et al.*, 1993).

A. sylvaticus is characterized by a greater ecological plasticity than *A. flavicollis* (Michaux *et al.*, 2005). It can inhabit different environments as woodland, moorland, steeped, arid Mediterranean scrubland and sand dunes. It can also appear in man-made habitats such as pastures, arable fields, forest plantations, gardens or urban parks (Schlitter *et al.*, 2010). Their diet is similar to the yellow-necked mice, feeding mainly on seeds and invertebrates (Montgomery and Montgomery, 1990). Their home range varies from 275 to 6150 m^2 , being bigger in males than in females (Korn, 1986). Despite being more generalist, when both species co-occur in the environment, *A. sylvaticus* is usually dominated by *A. flavicollis* (Michaux *et al.*, 2005).

1.2.1.2 Breeding behaviour

Both species have a similar breeding behaviour, being a classic example of r-strategists, organism with a high fecundity rate, early maturity, short generation time, short period of parental care and short lifespan which prioritise high growth rates. The breeding season usually occurs between March and October, but under certain circumstances it may continue throughout the year (Macdonald and Tattersall, 2001). Mating is usually limited to a dominant male and any receptive female in his territory, but other males can mate with females in that territory if they have opportunity to do so. Males produce ultrasound vocalisations in order to attract females (Leach 1990). Gestation lasts around 25 days (Macdonald and Tattersall, 2001). *A. flavicollis* have an average of 3 litters per year, producing from 2 to 11 pups in each. *A. sylvaticus*, however, can have from 4 to 7 litters per year, but the number of pups produced per litter ranges from 2 to 9 pups (Macdonald and Tattersall, 2001). The number of litters and pups can vary between geographical regions. Juveniles born at the beginning of the reproductive season can reproduce within the same year. Lifespan can reach up to two years but the mortality during winter is very high, with few adults surviving for more than one reproductive season (Macdonald and Tattersall, 2001). Males are aggressive with juveniles, especially during the breeding season, and often kill them (Leach, 1990).

1.2.1.3 Genetics of *Apodemus*

The genomes of *Apodemus flavicollis* and *Apodemus sylvaticus* contain 23 pairs of autosomal chromosomes and a pair of sexual chromosomes (XX for females, XY for males) determining a diploid chromosome number of 2n=48 (Figure 1.9). Their genome, despite their close relatedness to *Mus musculus*, includes 4 pairs of chromosomes more than the house mouse (2n=40).

However, this number can change due to the presence of B chromosomes (Adnađević *et al.*, 2012), supernumerary chromosomes that are thought to be dispensable for the organism. Only 1.2% of mammals have B chromosomes, mainly rodents, although they are quite common in *Apodemus*, where they have been found in 6 species (Jojić *et al.*, 2011), including *Apodemus flavicollis* and *Apodemus sylvaticus*. Their frequency is higher in *Apodemus flavicollis* than in *Apodemus sylvaticus* and differ between different populations (Bugarski-Stanojević *et al.*, 2016). In *Apodemus flavicollis*, B chromosomes are thought to have originated from the pericentromeric region of the sex chromosomes

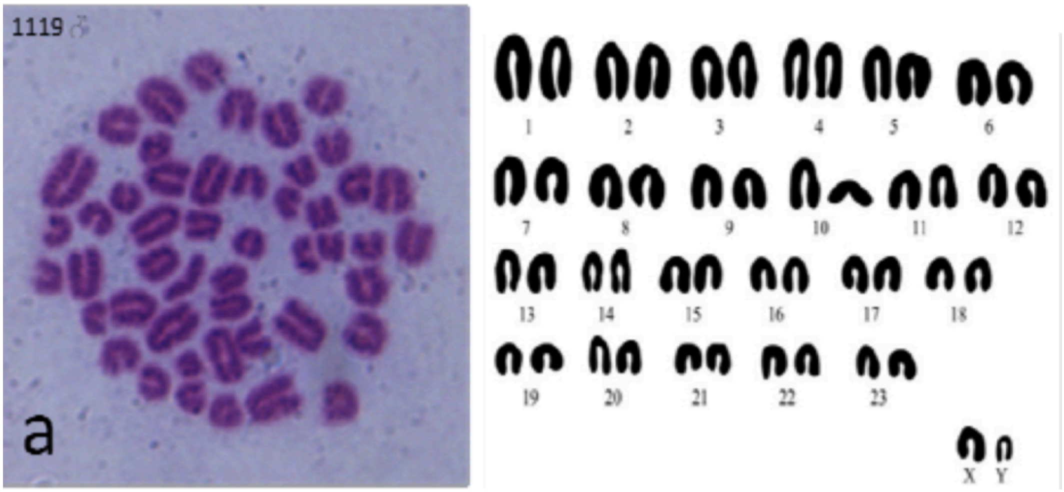


FIGURE 1.9: Karyotype of *Apodemus flavicollis*. From Eyison and Kıvanç (2016)

(Rajičić *et al.*, 2017). However, other species of *Apodemus*, such as *Apodemus peninsulae*, present B chromosomes with different DNA content, suggesting multiple origins of their supernumerary chromosomes (Rajičić *et al.*, 2017) The frequency of B chromosomes in natural populations is stable over time (Vujošvić, 1992). Two different models have been proposed to explain their maintenance. The parasitic model considers the equilibrium frequency of B chromosomes could be maintained by accumulation by meiotic drive and elimination due to their deleterious effects on the individuals carrying them. The heterotic model, however, considers that, without drive or an accumulation mechanism, the number of B chromosomes could be maintained due to the beneficial effects of having a small number of B chromosomes and the deleterious effects of higher numbers of them. There are different hypothesis about the possible benefits of B chromosomes on their carriers. Vujošević *et al.* (2007) found that the frequency of B chromosomes in *Apodemus flavicollis* was the lowest in the habitat predicted to be the optimal one. This discovery, along with the lack of a known accumulation mechanism for B chromosomes, support the maintenance of B chromosomes by an heterotic model. Furthermore, B chromosomes have been previously linked to a better winter survival (Zima *et al.*, 2003) and could be related to metabolism (Kozłowski *et al.*, 2003; Maciak *et al.*, 2014). Despite the amount of studies on B chromosomes, little is known about their roles, content or heritability.

Previous work on *Apodemus* has focused on short fragments of mtDNA, such as Cytb, (Michaux *et al.*, 2003; Michaux *et al.*, 2004; Michaux *et al.*, 2005), 12S (Michaux *et al.*, 2002) or COX1 (Panculescu-Gatej *et al.*, 2014; Štefka and Hypša, 2008); microsatellites (Czarnomska *et al.*, 2018; Bartmann and Gerlach, 2001; Berckmoes *et al.*, 2005; Makova *et al.*, 1998), and some nuclear genes, for example the interstitial retinol binding protein (IRBP) (Michaux *et al.*, 2002), I7 gene, an olfactory receptor (Suzuki *et al.*, 2008) or RAG1, a recombination activating gene (Suzuki *et al.*, 2008). These markers have been used for phylogenetic reconstruction and to study the phylogeographic relationships between populations of Apodemus in Europe and Asia, but are insufficient for more complex studies, such as detecting adaptations or investigating genetic basis of complex traits due to low genomic resolution. Adaptation processes generally involves few genes of large effect and multiple genes of small effect. Only using whole genome approaches it is possible to detect sequences that will have extreme differentiation between ecotypes, morphotypes or populations without *a priori* knowledge of a selected phenotype (Stapley *et al.*, 2010). Even when genome-wide reduced representation methods might not identify genes under selection, they can help to identify regions under selection where further sequencing efforts should be concentrated.

1.2.2 European phylogeography of *Apodemus*: state of the art

The European phylogeography of species from the genus *Apodemus* has been broadly studied for the last two decades. Michaux *et al.* (2004) analysed the possible refugia used by *A.flavicollis* during the Quaternary glaciations, using sequences of 972 bp from CytB from 124 individuals from 53 different locations in Europe. Neighbour-joining trees (Figure 1.10) revealed the existence of two different clades that diverged 2.2-2.4 Ma: one clade includes samples from Turkey, Near and Middle East (Clade 2, Figure 1.10), whereas the other includes three subclades (Subclades 1a, 1b and 1c, Figure 1.10) from the western Palearctic region.

Michaux *et al.* (2004) interpreted this as the existance of two main refugia, into which ancestral populations split when the climate started to fluctuate in Europe: one in the Balkan region and a second one in the Near East region. The highest nucleotide diversity, in *A. flavicollis*, was found in the Balkan region. Michaux *et al.* (2004) explained that the current European population is composed of three different subclades that survived the Quaternary glaciations in three independent refugia within the Balkan region and merged later during the interglacial periods. The Iberian peninsula and

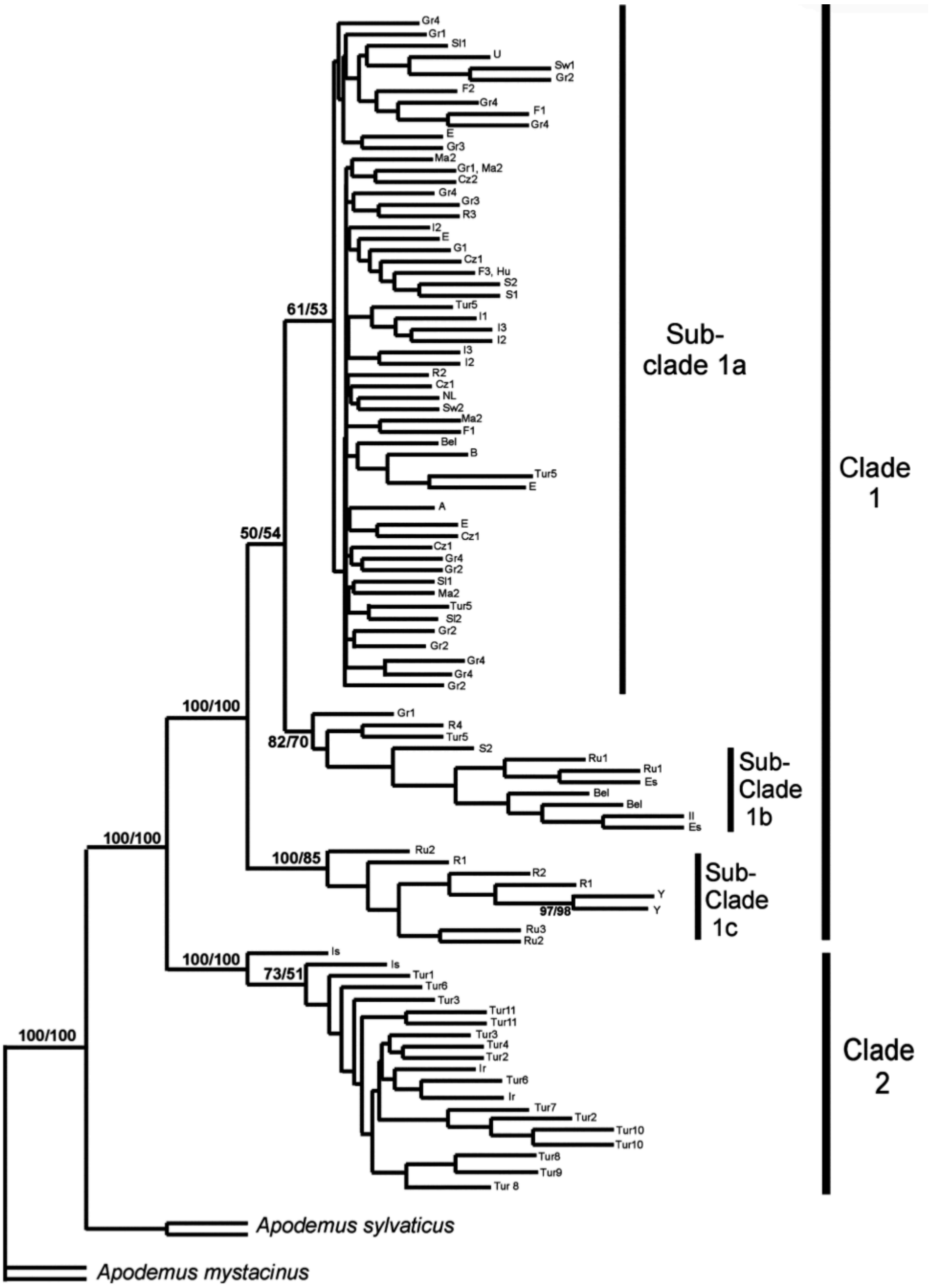


FIGURE 1.10: Neighbour-joining tree for *Apodemus flavicollis* cytB mtDNA haplogroups. Only bootstrap support values higher than 50% in neighbour joining and maximum parsimony analysis are shown. From Michaux *et al.* (2004)

southern France populations, however, had a low nucleotide diversity compared to the other European populations, suggesting that they have been generated by recolonisation from the Italo-Balkan populations and have suffered a strong bottleneck. These populations, therefore, did not survived the LGM (23–19 Ka) there (Michaux *et al.*, 2005).

Michaux *et al.* (2004) suggested that individuals surviving in Turkey could have done it in the western coastal regions, however refugia with Turkey were not cleared located yet . They proposed that changes in vegetation during the last 20000 years would have allowed *A. flavicollis* to expand through the Oriental regions. However, this lineage did not contribute to the recolonization of Europe, probably due to the presence of strong barriers, such as the Black Sea or the Caucasus, that could stop their expansion into northern Europe (Michaux *et al.*, 2004). Michaux *et al.* (2004) also suggested that the presence of *A. flavicollis* already in the Balkan region could have impeded the arrival of new colonisers through intraspecific aggressiveness towards newcommers.

Michaux *et al.* (2003) also studied the possible refugia used by *A. sylvaticus* to survive to the Quaternary ice ages. They used 965 bp of CytB from 102 individuals from 40 different European locations.

Their analyses revealed the existence of two main clusters, each one representing a genetic lineage with a non-overlapping distribution (Figure 1.11). The two main lineages belonged to continental Europe and split 1.5–1.6 Ma. The first lineage was divided into two subclades with non overlapping distributions: one included samples that ranged from south Spain to Sweden and Ukraine while the second one was restricted to northern Africa. The second lineage was also divided into two main groups or subclades, with non overlapping distributions: one including samples from Italy, Balkans and western Turkey and a second one including samples from Sicily. Michaux *et al.* (2003) interpreted the higher genetic diversity found in Iberia and southern France as a signal of a refugium in this region, from where *Apodemus sylvaticus* recolonized Europe at the end of the LGM (19 Ka). The North African group that clusters within the Spanish-western Europe lineage, could have originated from a recent anthropogenic introduction from south-western Europe.

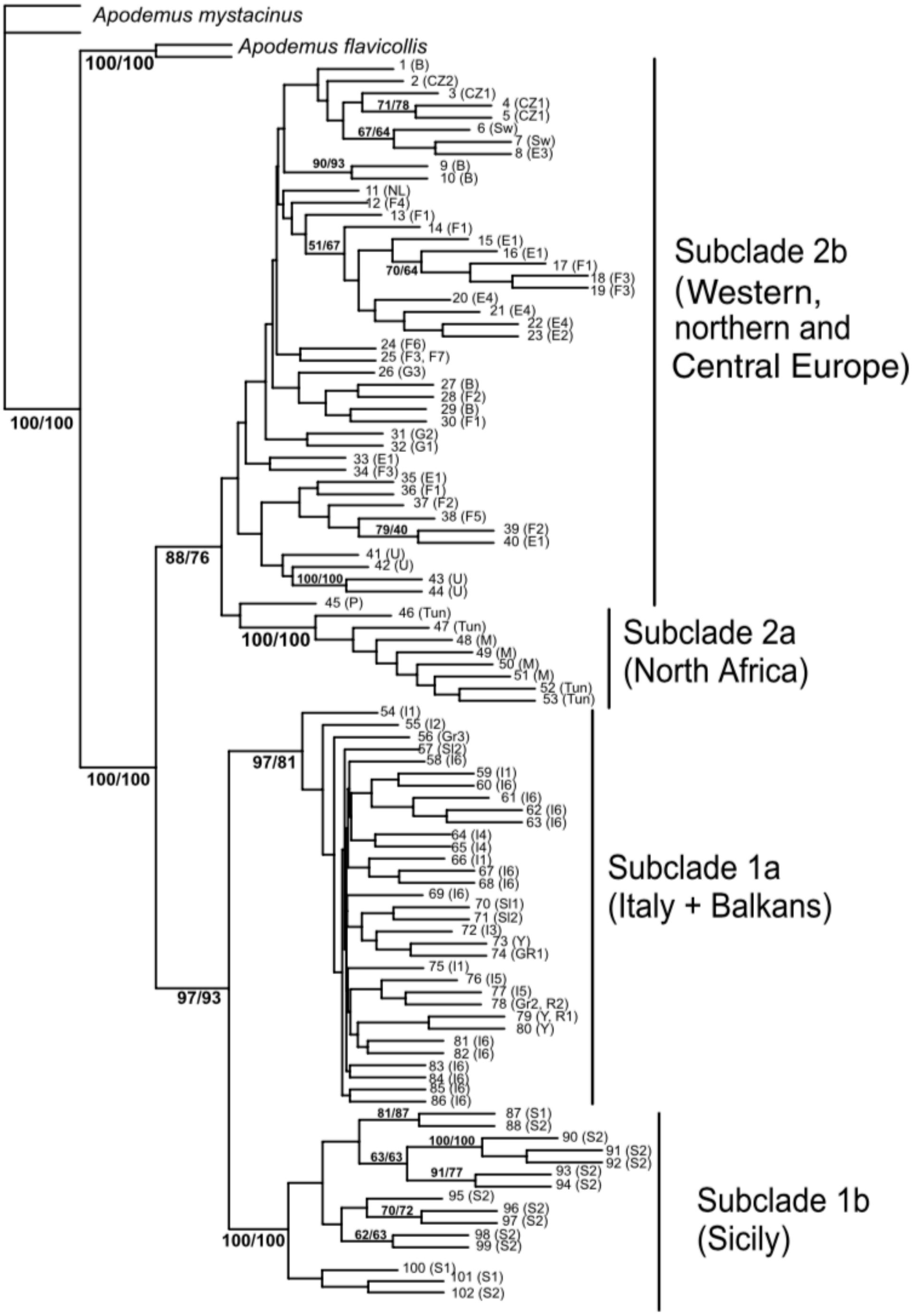


FIGURE 1.11: .
Neighbour-joining tree for *Apodemus sylvaticus* CytB mtDNA haplogroups. On the top of the branches are shown the bootstrap support values higher than 50% in neighbour joining and maximum parsimony analysis. From Michaux *et al.* (2003)

Michaux *et al.* (2003) also found a low genetic diversity in the Italo-Balkan peninsulas, probably due to the effect of a strong bottleneck. Furthermore, the Italo-Balkans region was used as refugia for *A. flavicollis*, so interspecific competition could occur. Michaux *et al.* (2003) interpreted these results along with palaeontological and palaeoclimatological data as a signal of an Italo-Balkan refugia for *Apodemus sylvaticus* during the Quaternary ice ages.

The Alps have likely limited the expansion of the Italo-Balkan populations, although the drops in the level of the Adriatic and Marmara Seas could have allowed genetic exchanges between the populations confined to these areas. The Sicilian lineage appeared in the same clade as the Italo-Balkan one, from which it separated around 0,8 Ma. This lineage seems to be a hot-spot of genetic diversity for *A. sylvaticus*, probably due to the isolation of an old Italian lineage. Michaux *et al.* (2003) suggested that this old lineage could have entered Sicily 70 Ka, where it has been trapped until present, while the continental population was replaced by modern lineages from the Italo-Balkan refugia.

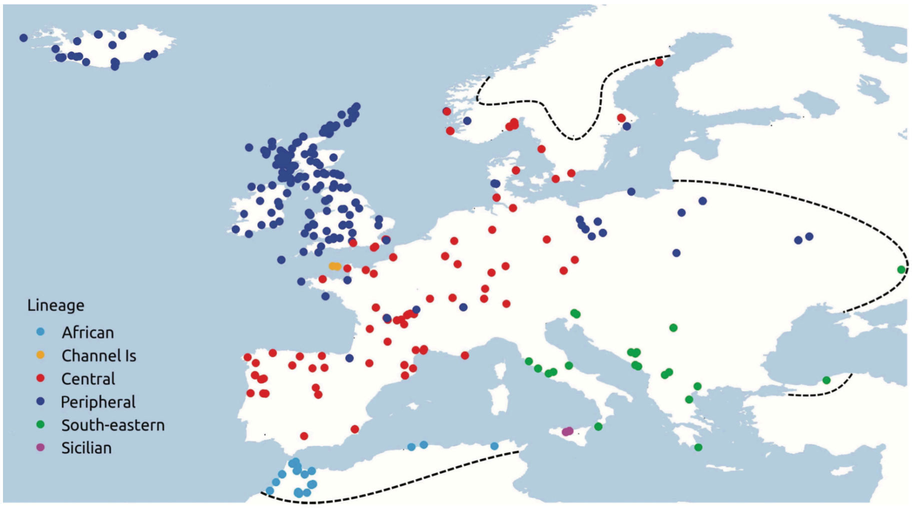


FIGURE 1.12: Distribution of six CytB lineages identify in *Apodemus sylvaticus*. Each colour represents one of the lineages. From Herman *et al.* (2017)

The most recent studies, which included a larger number of samples, supported these conclusions (Herman *et al.*, 2017) (Figure 1.12).This study was performed with sequences between 818-1140bp long from CytB of 981 individuals from different locations in Europe. It demonstrated that the Spanish-western Paleartic lineage (Michaux *et al.*, 2003) actually comprises of two well-supported lineages (Figure 1.13) with overlapping distributions, one in central Europe (In red in Figures 1.12 and 1.13) and another one with a more peripheral distribution (In blue in Figures 1.12 and 1.13) (Herman *et al.*, 2017).

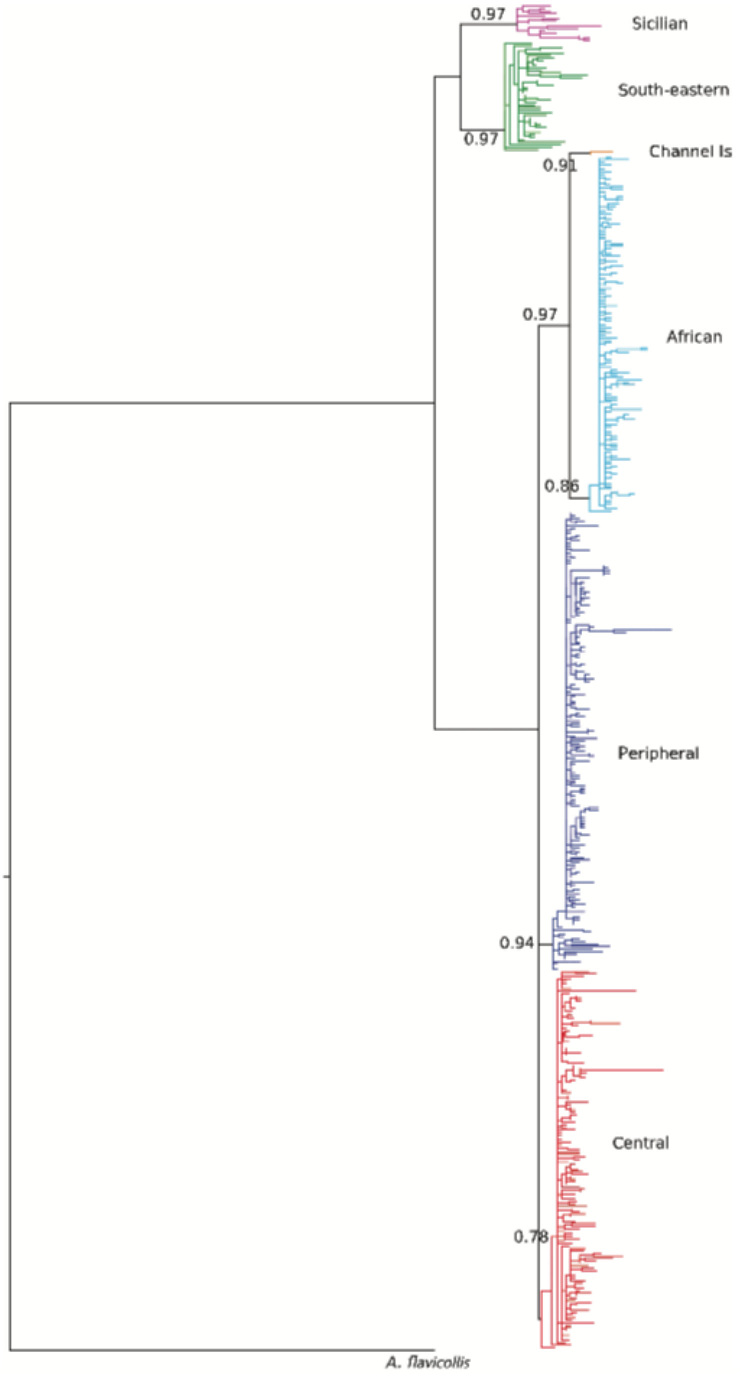


FIGURE 1.13: Maximum likelihood tree for *Apodemus sylvaticus* CytB mtDNA haplogroups. Maximum likelihood support values (SH-aLRT) are shown for the six identified mtDNA lineages and deeper splits. From Herman *et al.* (2017)

This peripheral lineage appears mainly in the British Isles, Iceland, Poland and Russia, but also in some populations from France, Germany, Denmark, Sweden and Norway. In addition, Herman *et al.* (2017) suggested the existence of three refugia for *A. sylvaticus*, two in the south of Europe and another one in a more northern location, the exact position of which could not be determined with their dataset.

This possible northern refugium has also been detected in *Heligmosomoides polygyrus*, a nematode which parasitises *Apodemus sylvaticus*, in a study includying 687 bp long fragments of CytB from 136 indivuals from 22 different european locations (Nieberding *et al.*, 2005) (Figure 1.14). They found a clade including only samples from Ireland and Denmark, whose divergence time was estimated between 2.02 ± 0.21 and 1.46 ± 0.19 Ma, and hypothesise this clade could have survived in southern England, a region that was not always covered by glaciers, or somewhere else further south in the continent.

However, this postulated northern refugium is not supported either by the fossil records of small mammals on Younger Palaeolithic sites in Europe (between 23 and 16 Ka, summarized by Sommer and Nadachowski (2006)), or by the distributions models calculated by Fløjgaard *et al.* (2009). Furthermore, the calibrations obtained for the most recent common ancestor of the Peripherial group (HPD lower =8.689 Ka, median=16.363 Ka, HPD upper=32.252 Ka) is more recent than the ones obtained for the two traditional refugia (Central: HPD lower =9.404 Ka, median= 22.254 Ka, HPD upper=37.355 Ka and South Eastern: HPD lower =13.799 Ka, median=19.868 Ka, HPD upper=29.079 Ka).

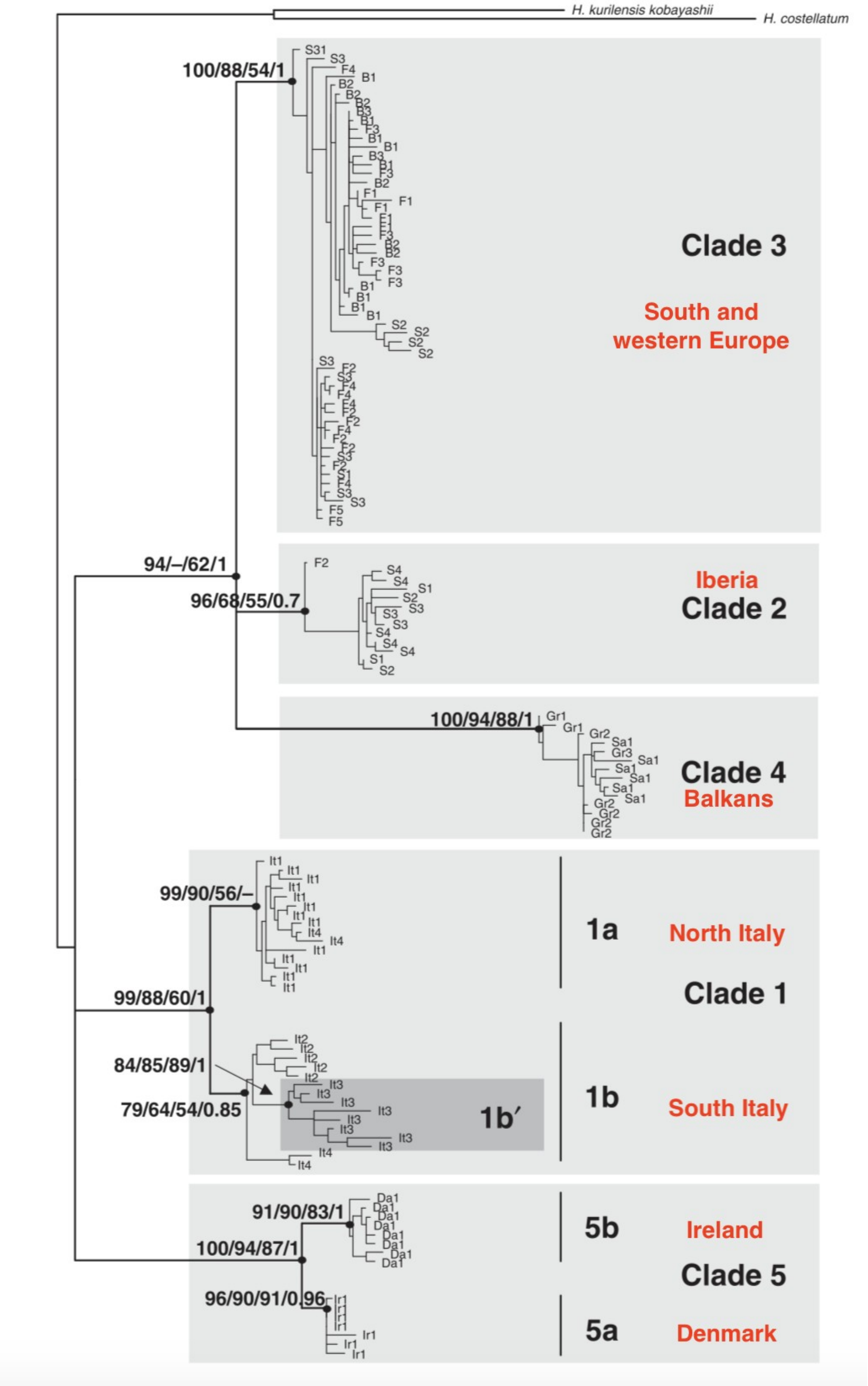


FIGURE 1.14: CytB maximum likelihood tree of *Heligmosomoides polygyrus*. Four different support values are shown on top of the branches: 1- Neighbour joining bootstrap support values, 2- phyml bootstap support values, 3- Bootstrap support in Bayesian Analysis and 4- posterior probabilities from MrBayes analysis. Only bootstrap values higher than 70% and posterior probabilities higher than 0.5 are shown. The origin of the samples is indicated in red. Modified from Nieberding *et al.* (2005)

1.3 Evolution of RAD-seq protocols

1.3.1 Evolution of sequencing technologies

During the 1970s, different DNA sequencing technologies were developed: "plus and minus" (Sanger and Coulson, 1975), Sanger (Sanger *et al.*, 1977) and Maxam-Gilbert sequencing (Maxam and Gilbert, 1977). These techniques completely revolutionised genetics. However, during the next 20 years only viral and organellar genomes were sequenced completely. It was not until 1995 when the first free-living organisms' genome was sequenced, the bacteria *Haemophilus influenzae Rd* (Fleischmann *et al.*, 1995).

The Human Genome Project (HGP), launched in 1990, which required huge amounts of time and resources, stimulated the developement of faster, cheaper and higher-throughput sequencing technologies (Van Dijk *et al.*, 2014). Some of the most successful, known as next-generation sequencing (NGS), were 454-pyrosequencing (Pyrosequencing AB, Uppsala, Sweden, now Roche Company, Branford, CT, USA), Illumina's sequencing-by-synthesis (San Diego, CA, USA), SOLiD sequencing-by-ligation (Life Technologies, Carlsbad, CA, USA), Ion Torrent's semiconductor sequencing (Life Technologies, Carlsbad, CA, USA) and PacBio Single Molecule Real-Time sequencing (Pacific Biosciences, Menlo Park, CA, USA, now Illumina, San Diego, CA, USA). Almost all these methods were able to perform thousands to millions of sequencing reactions in parallel, produce relatively short reads, and did not rely on electrophoresis for base calling (Van Dijk *et al.*, 2014). The improvements in these technologies have reduced the price of sequencing by many orders of magnitude during the last 20 years, making it affordable for most laboratories around the world (Figure 1.15).

One of the milestones of the Human Genome Project was achieving a cost of US$1000 per human genome, which was reached in 2017 with Illumina's High X Ten system. Nevertheless, the cost of this system in 2014 was close to 10 million dollars, a price that only a few customers could pay (Hayden, 2014). However, even when the US$1000 per human genome is not yet achievable for most laboratories, it is a clear sign of the continued improvement of NGS technologies. Nowadays, technological advances are focused on long reads and single molecule sequencing (10x Genomics, Pleasanton, CA and Nanopore, Oxford, United Kingdom) as well as on the reduction of sequencing errors.

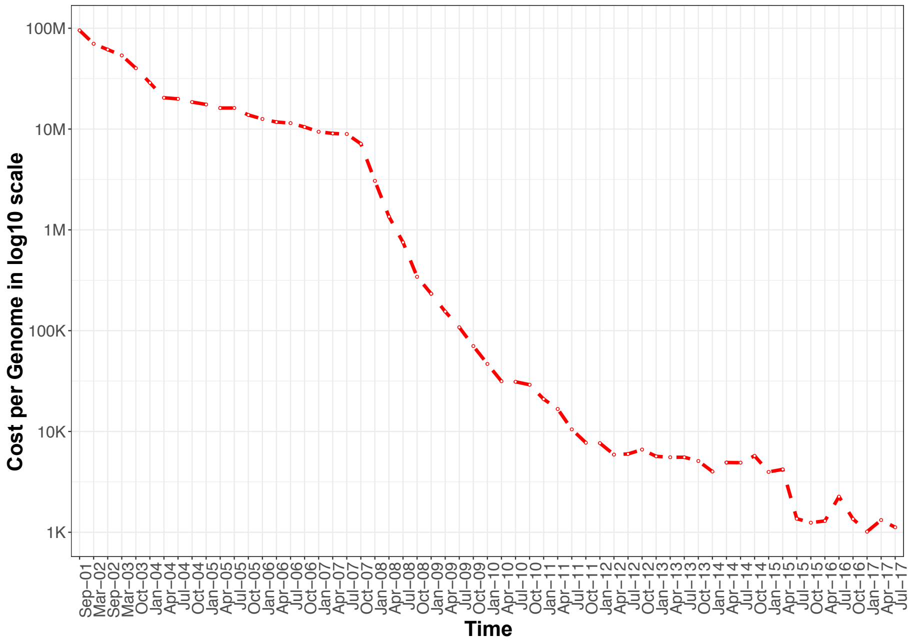


FIGURE 1.15: Evolution of the cost of sequences per per genome since 2001 to 2017 . Modified from National Human Genome Research Institute (NHGRI)

In addition to advances in whole genome sequencing, falling costs enabled developments and application of multiple reduce representation sequencing methods: genotyping-by-sequencing (GBS), restriction-site-associated DNA sequencing (RAD-seq), reduced-representation bisulfite sequencing (RRBS), exon capture or transcriptome sequencing. These methods allow sequencing of only targeted parts of the genome, reducing the amount of sequences required per individual, and therefore increasing the number of individuals that can be sequenced while increasing the coverage of a single individual per unit of cost.

1.3.2 Restriction-site-associated DNA sequencing (RAD-seq)

In 2006, Eric A Johnson's group at the Institute of Molecular Biology at the University of Oregon described a rapid and cost-effective method that allowed them to identify polymorphisms and genotype populations using sequences produced by restriction enzyme digestion, the restriction-site-associated DNA (RAD) markers (Miller *et al.*, 2007). This method creates a reduced representation of a genome by digestion of DNA with a restriction enzyme and ligation of biotinylated linkers. After random shearings, only fragments containing biotinylated linkers will be kept and identified by differential hybridization patterns on a microarray (Miller *et al.*, 2007). These fragments are largely homologous between different individuals and enable screening of thousands of markers in parallel. This method was subsequently adapted to be used with the new sequencing technologies and became known as restriction-site-associated DNA sequencing (RAD-seq) (Baird *et al.*, 2008).

The main advantage of this method is its capacity to select a small fraction of the whole genome to identify unbiased sets of thousands of SNPs (Rašić *et al.*, 2014) without requiring almost any previous knowledge about the genome of the organism selected for the study. The number of markers obtained after the digestion with a restriction enzyme could be controlled by the selection of a specific enzyme and also by the selection of a specific size range of digested fragments. Moreover, samples from different individuals could be combined ("multiplexed") to fit on a single lane of an Illumina sequencer (Baird *et al.*, 2008), reducing the cost of sequencing per sample.

In the RAD-seq method developed by (Baird *et al.*, 2008), high molecular weight (HMW) DNA is first digested with a restriction enzyme. Next, Illumina adapters, including individual barcodes, are ligated to the sticky ends generated by the enzymes. At this point, DNA from multiple individuals can be pooled together and the fragments are randomly sheared. A second adapter is ligated to the 3′ end of the sequences afterwards and the different fragments are amplified in a PCR reaction (Figure 1.16).

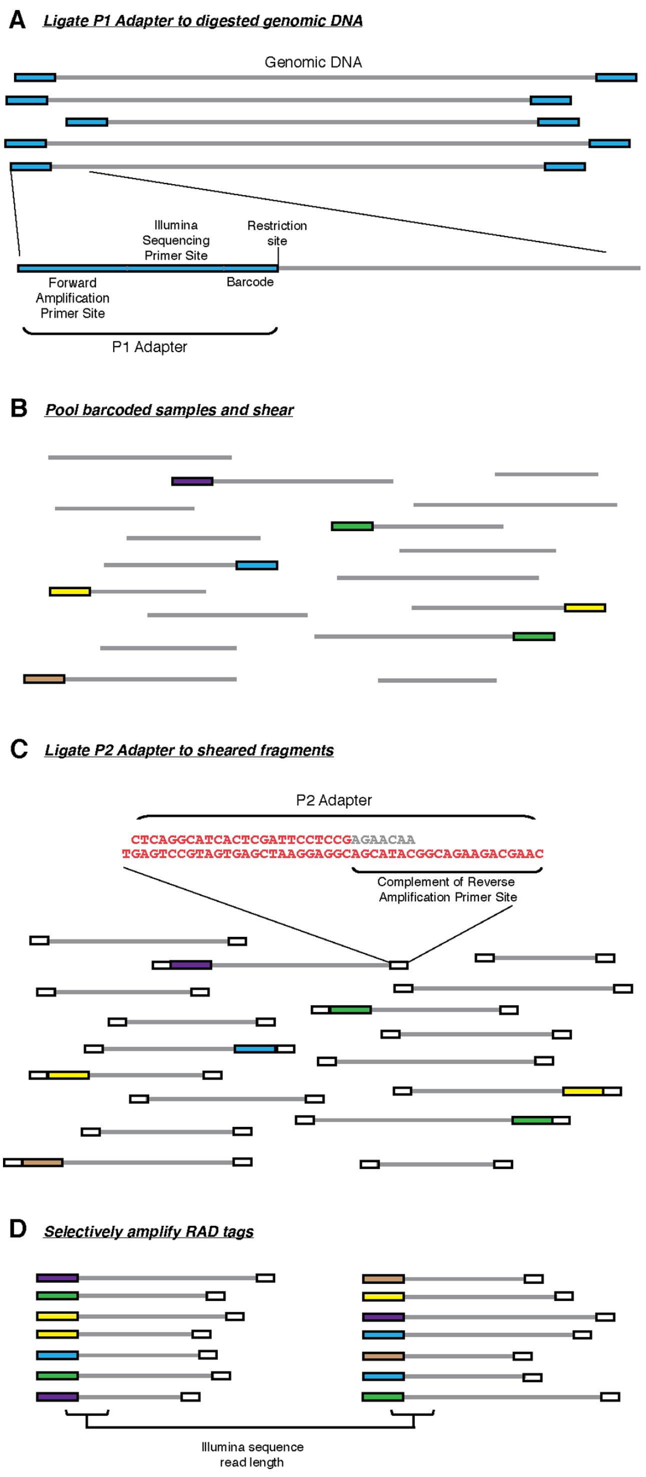

A
Ligate P1 Adapter to digested genomic DNA
Genomic DNA
Illumina Sequencing Primer Site
Restriction site
Forward Amplification Primer Site
Barcode
P1 Adapter
B
Pool barcoded samples and shear
C
Ligate P2 Adapter to sheared fragments
P2 Adapter
CTCAGGCATCACTCGATTCCTCCGAGAACAA
TGAGTCCGTAGTGAGCTAAGGAGGCAGCATACGGCAGAAGACGAAC
Complement of Reverse Amplification Primer Site
D
Selectively amplify RAD tags
Illumina sequence read length

FIGURE 1.16: Rad-seq protocol from Baird *et al.* (2008). A- Digestion of the DNA and ligation with adapters P1. B- Multiplexing of samples ligated to different adapters and shearing of the DNA. C- Ligation to P2 "Y" adapters. D- Amplification of fragments containing P1 adapters.

The random shearing step in this method allows the identification of PCR duplicates generated during the PCR step. Two fragments with identical sequence and length will be considered as copies of the same fragment, and therefore can be removed from further analysis. However, random shearing has to be performed by a sonicator, an expensive piece of equipment that is not commonly found in standard molecular laboratories. Furthermore, different length libraries hybridise to the sequencing flowcells with different efficiencies, producing fragments sequenced at different coverages. Nonetheless, this approach enabled a revolution in ecological genomics. It enabled generation of a large number of SNPs independently of the knowledge of the genomic sequences of an organism. These markes have opened previously unavailable research directions, such as population genetics (Blanco-Bercial and Bucklin, 2016; Cromie *et al.*, 2013; Hohenlohe *et al.*, 2013; Combosch and Vollmer, 2015), phylogeography (Jeffries *et al.*, 2016; Reitzel *et al.*, 2013; Alter *et al.*, 2017), phylogenetics (Eaton and Ree, 2013; Hipp *et al.*, 2014; Hou *et al.*, 2015; Razkin *et al.*, 2016), marker development (Pegadaraju *et al.*, 2013) and linkage mapping studies (Henning *et al.*, 2014; Gonen *et al.*, 2014; Baxter *et al.*, 2011), among others, for a fraction of their past cost and effort.

1.3.2.1 Double-digestion RAD-seq

Original RAD-seq and GBS protocols were based on the use of a single restriction enzyme. Despite the suitability of this approaches to genotype multiple individuals, some improvements were needed in order to allow researchers to more precisely select the fraction of the genome to sample (Peterson *et al.*, 2012). Using two restriction enzymes (double digestion or ddRAD-seq (Peterson *et al.*, 2012; Poland and Rife, 2012)) and by changing the combination of enzymes and the range of sizes of interest, we can produce between hundreds to hundreds of thousands of SNPs at different coverages (Puritz *et al.*, 2014b). ddRADseq protocols (Peterson *et al.*, 2012; Poland and Rife, 2012) therefore include an additional step of agarose gel-based size selection. As ddRAD-seq does not require a shearing step or enzymatic end repair, it can be 5-10-fold cheaper than single digestion RAD-seq protocols (Peterson *et al.*, 2012).

Despite all the advantages of ddRAD-seq, it is important to be aware of its limitations or possible complications. During the PCR step, PCR duplicates are produced but are now indistinguishable as the random shearing step is removed. PCR duplicates can be confounded with real reads, producing false genotype calls. PCR duplicates, therefore, can modify alleles frequencies and increase homozygosity (Pompanon *et al.*, 2005). A solution to this problem is to incorporate random bases to the Illumina adapters in order to make identification of the duplicates possible (Tin *et al.*, 2015; Hoffberg *et al.*, 2016; Franchini *et al.*, 2017). Methods that do not allow recognition of PCR duplicates usually reduce to the minimum possible the number of cycles during PCR reaction, reducing the potential error (Andrews *et al.*, 2016). Previous studies have shown a very different percentage of duplicate reads on the libraries sequenced. Schweyen *et al.* (2014) identified a 33.48% of the reads as PCR duplicates while Franchini *et al.* (2017) detected between 0.17% and 31.46% of duplicated sequences, with the higher percentages obtained in those libraries prepared with a low amount of input DNA and a high number of PCR cycles.

Some protocols, in order to increase multiplexing capacity without increasing the cost associated with adapters synthesis, include a four barcodes strategy (Franchini *et al.*, 2017). Using all the possible combinations of 15 barcodes (3 inner-5′ barcodes, 4 inner-3′ barcodes, 4 outer-5′ barcodes and 4 outer-3′ barcodes) it is possible to multiplex 192 samples (Figure 1.17) on a single lane of Illumina sequencing. However, jumping PCR can produce chimeric or hybrid sequences that will be misidentified as a wrong sample. The copy number of these chimeric sequences is not expected to be high, so they could potentially be filtered out, along with low coverage reads, during the bioinformatic analysis.

High multiplexing methods, such as quaddRAD, (Franchini *et al.*, 2017) have the advantage of reducing the gel interlane size variation. Briefly, this variation originates when the size selection step needs to be performed in more than a single gel lane. The loci selected in each one of the lanes can be slightly different, reducing the proportion of overlapping loci (Franchini *et al.*, 2017). Increasing the amount of individuals that can be multiplexed reduces this bias.

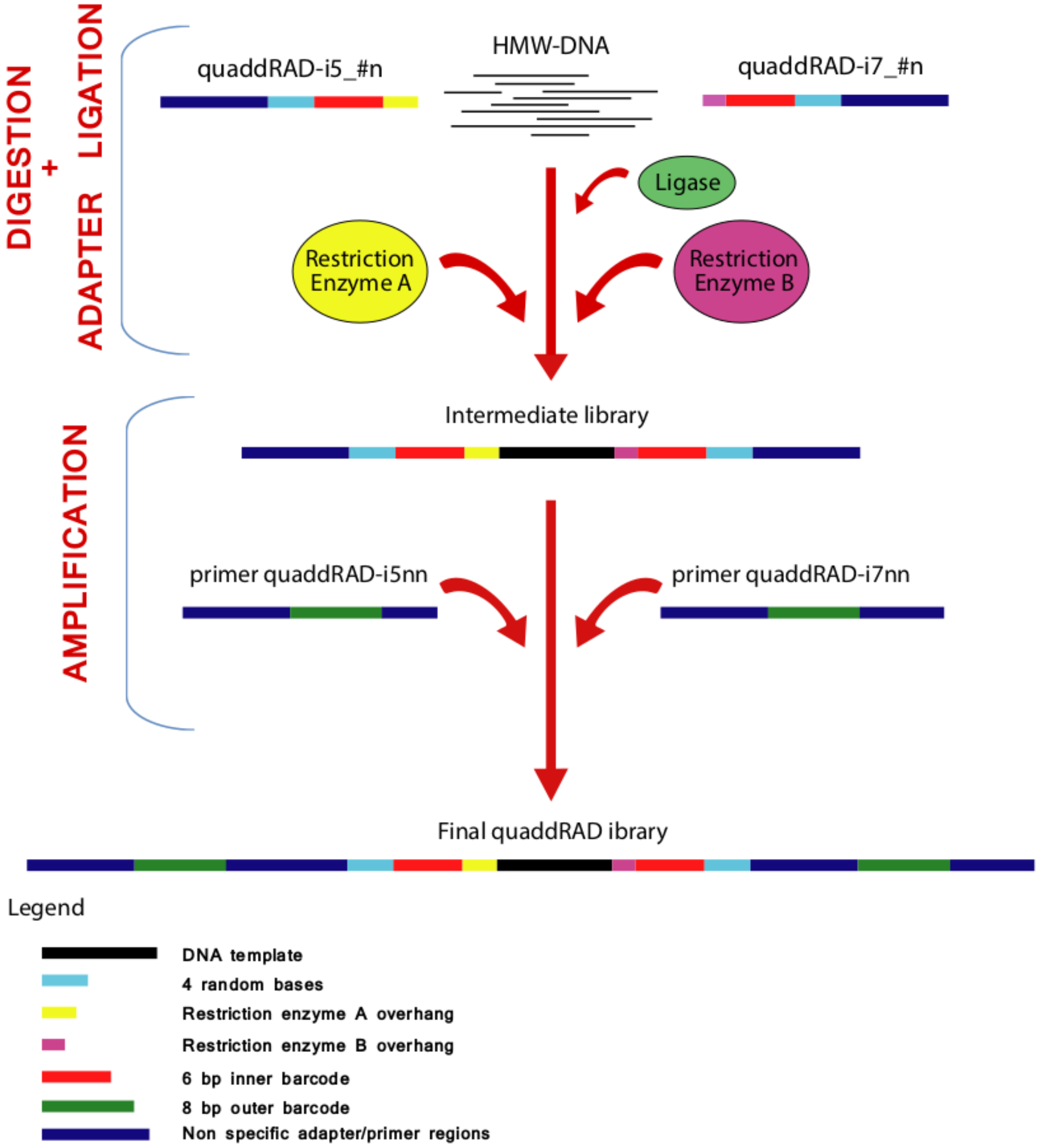


FIGURE 1.17: QuaddRAD protocol from Franchini *et al.* (2017). Schematic representation of the structure of the structure of the DNA during different steps of the library preparation. HMW DNA is digested and ligated to inner adapters including an overhang with the enzymes used, inner barcodes, 4 random nucleotides and a primer region. During the amplification step other pair of barcodes will be added to the DNA, generating DNA fragments with 4 barcodes and 2 degenerated base regions.

1.3.3 Bioinformatics analysis and STACKS

Aside from the challenges during the library preparation, RAD-seq analyses also require accurate bioinformatic analyses. Multiple pieces of software have been developed to analyse RAD-seq data (e.g. Stacks (Catchen *et al.*, 2011), dDocent (Puritz *et al.*, 2014a), AftrRAD (Sovic *et al.*, 2015), PyRAD (Eaton, 2014), Rainbow (Chong *et al.*, 2012)), in addition to others, designed for specific steps related to Illumina sequencing (e.g. FASTQC (Andrews, 2010), trimmomatic (Bolger *et al.*, 2014) or cutadapt (Martin, 2011). Quality control and filtering of sequences with low quality is one of the main steps of RADseq data analysis. One of the most used software for quality control is FASTQC (Andrews, 2010), which generates a graphic report including information about per base quality, per tile sequence quality, per sequence quality score, per base sequence content, GC content, per base N content, sequence length distribution, duplication level, overrepresented sequences, adapter content and kmer content. Once the quality of the data has been estimated, it is necessary to remove low quality reads.

Another important step is the demultiplexing of the individuals included on the same lane of sequencing. A good design of the barcodes used during the library preparation is essential for an accurate separation of the reads, as sequencing errors within barcodes can potentially result in misalocation of reads to wrong individuals. Stacks (Catchen *et al.*, 2011), probably the most commonly used pipeline for RAD-seq analysis (Paris *et al.*, 2017), combines quality check, filtering and demultiplexing.

When a reference genome is available, reads should be aligned to it, and SNPs are called based on the alignment. However, as previously said, a reference genome is not required for RAD-seq data analysis. A *de-novo* assembly of loci can be performed and SNPs can be called in each assembled loci. Stacks pipeline provides three main parameters to control the number of loci and polymorphisms found: m, M or n (Table 1.1).

The first two parameters act at the individual read level, setting the minimum number of identical reads needed to create a stack (m) and the number of mismatches allowed between stacks to merge them into a locus. The third parameter is used to build a catalogue with all the loci found between individuals from the different populations. This parameter controls the number of differences allow between stacks from different individuals. An inaccurate use of these parameters may have important consequences on the results. For example, sequence errors can be misidentified as polymorphism, alleles with

TABLE 1.1: Parameters controlling the number of loci and polymorphism used in Stacks. Modified from (Paris *et al.*, 2017)

Parameters	**Default value**	**STACKS component**	Description
m	3	ustacks	Minimum number of raw reads required to form a stack (a putative allele)
M	2	ustacks	Number of mismatches allowed between (putative alleles) to merge them into a putative loci
n	1	cstacks	Number of mismatches allowed between stacks (putative loci) during the construction of the catalog

a coverage lower than the m parameter will be missed, or different alleles fixed in different populations can be analysed as different loci.

Various approaches have been proposed to identify the best parameters for a dataset. Mastretta-Yanes *et al.* (2015) use replicated samples that should produce identical genotypes, to optimize the parameter selection for RAD-seq projects without a reference genome available. Following this approach, it is possible to quantify different error rates, as loci error rate, allele error rate and SNPs error rate. The information obtained by this method can be useful to determine the quality of the library and also the quality of the bioinformatic analysis, but increases the price per sample, due to the need of sequencing the same samples multiple times.

Paris *et al.* (2017) proposed a method based on the repeated runs of denovo_map package from Stacks (Catchen *et al.*, 2011), changing only one parameter value each time. After calculating the number of assembled loci, polymorphic loci and number of SNPs for each individual as well as for the loci shared between the 40, 60 and 80% of the samples, the selection of the best parameters can be performed by choosing the values that produced the higher number of polymorphism in loci that are shared between the 80% of the individuals in a population. Loci that are shared between a high percentage of the population are unlikely to be derived from paralogous or repetitive sequences and will not incorporate a significant amount of sequencing errors, making them the best proxy of a genome (Paris *et al.*, 2017). The highest number of polymorphism was always found for values of n equal to ± 1 iteration of M, so they recommend to set n following this general rule.

A mixed approach has also been described by (Paris *et al.*, 2017), consisting of a *de-novo* assembly of loci and posterior alignment of the consensus sequences to the reference genome. This method has been shown as more effective than a direct alignment to a reference genome. Insertions and deletions with respect to the reference genome can result in a poor alignment of raw reads that can be improved when aligning consensus reads obtained after the *de-novo* assembly.

1.4 Scope of the thesis

The relatively short life, high reproductive rate, wide-spread ecology and easy trapping make *Apodemus flavicolllis* and *Apodemus sylvaticus*, and other members of the genus, a rich target for evolutionary studies on hybridisation, host–pathogen interactions, adaptations and heritability. Such studies require well-established population genetics and genome-wide variation parameters for the species. However, *Apodemus* are very underdeveloped in terms of the genomic and genetic resources available for their study.
The main aim of this project is to investigate the phylogeographic history of two different species of wild mice, *Apodemus flavicollis* and *Apodemus sylvaticus* in Europe, using, for the first time in this genus, whole-genome, high-density genotyping using restriction-site-associated DNA sequencing. The content of each Chapter and the associated supplementary information will be shortly summarized.

- **Chapter 2**: Includes a small-scale project, where we investigated the population structure and genetic diversity of three populations of *Apodemus flavicollis* and compared them to two populations of *Apodemus sylvaticus*. This approach enabled the development of an analysis pipeline and identify and incorporate improvements in the RAD-seq protocols, which were then implemented in the major part of the work. This work is also a basis of a manuscript that will be shorty submitted for publication (Authors: Maria Luisa Martin Cerezo, Marek Kucka, Karol Zub, Yingguang Frank Chan & Jarek Bryk)

- **Appendix A**: Includes information about the localities, coordinates and type of environment for each sample included in the study in Chapter2. All the tables produced for the calculation of the best parameters for the combined dataset and for *Apodemus flavicollis* alone, as well as graphs which summarise that information are also available through a link to Github or directly in the appendix. A list of the 117 loci with the highest divergence and PCA plots performed using those loci for the Polish dataset and the 20 European and Tunisian samples have also been included. All code used to run the analyses and plot the results, as well as the complete catalogue used to differentiate between species has been made available also through link to Github or Dropbox.

- **Chapter 3**: This chapter comprises a technical discussion of the challenges encountered during library preparation in Chapter 2. Here we also describe an alternative protocol using custom adapters which were thoroughly tested before use in the main part of the project.

- **Appendix B**: Includes the laboratory protocol followed for the preparation of the trial library used on Chapter 3 as well than the sequences of all the new adapters and barcodes. Also included are tables summarising the results of demultiplexing, as a total number of reads, percentage of PCR duplicates, ambiguous barcodes and ambiguous radtags, among others.

- **Chapter 4**: Here I describe all methods used to produce and analyse the data generated to investigate the phylogeographic history of *Apodemus flavicollis* and *Apodemus sylvaticus* in Europe, from sample collection and library preparation to bioinformatic analysis.

- **Appendix C**: A complete list of samples sequenced for this project is included in Appendix C, including information about samples that were sequenced by duplicate and triplicate.

- **Chapter 5**: Chapter 5 describes all the results obtained for the main project of this thesis, from data manipulation and cleaning, to species differentiation and phylogeographic patterns observed for *Apodemus flavicollis* and *Apodemus sylvaticus*.

- **Appendix D**: Includes all the tables generated to estimate the optimum parameters for genotyping *Apodemus flavicollis* and *Apodemus sylvaticus* samples. A population map with the samples that passed filtering and were used in the analysis is also available here. Finally, tables with F_{st} pairwise comparisons and the estimation of Π, H_e, H_o and % polymorphic loci have been made available through Github links.

- **Chapter 6**: Chapter 6 discusses the accuracy and effectiveness of the newly developed method, and the importance of PCR artefacts such as PCR duplicates and chimeric sequences with regards to library preparation and analysis of the sequenced data. Furthermore, comparisons are made between phylogeographic patterns observed in this project and previously publish data, giving new insights about the potential existence of northern refugia for *Apodemus* species. Finally, I summarize the future direction of this project.

Chapter 2

Population structure of *Apodemus flavicollis* and comparison to *Apodemus sylvaticus* in north-eastern Poland using ddRAD-seq

2.1 Introduction

In the Western Palearctic, the yellow-necked mice *A. flavicollis* (Melchior, 1934) and the woodmice *A. sylvaticus* (Linnaeus, 1758) are widespread, sympatric and occasionally syntopic species. They are often difficult to distinguish morphologically in their southern range (Jojić *et al.*, 2014), but in central and northern Europe, both are easily distinguishable by the yellow collar around the neck of *A. flavicollis*, that it is absent in *A. sylvaticus*.

Their prevalence in Western Palearctic and common status in western and central Europe following the last glaciation (Michaux *et al.*, 2005; Herman *et al.*, 2017) made them one of the model organisms to study post-glacial movement of mammals. Both species present a host of characteristics that also make them suitable for ecological genomics studies: they are wide-spread, common, not commensal to humans and have a history of ecological studies. They have traditionally been studied in a parasitological context, as one of the vectors of *Borellia*-carrying ticks *Ixodes ricinus*, who often feed on *Apodemus* (Cull *et al.*, 2017; Richter *et al.*, 2011), tick-borne encephalitis virus (Mlera and Bloom, 2018) and hantaviruses (Kolodziej *et al.*, 2018; Papa *et al.*, 2016). They have been used as markers for environmental quality (Martiniaková *et al.*, 2010; Velickovic, 2007). Lastly, *Apodemus* also have extra-autosomal chromosomes, called B chromosomes, with varied distribution among the populations (Rajičić *et al.*, 2017). The role of B chromosomes is unknown, although it has been suggested they play a role in cellular metabolism (Kozłowski *et al.*, 2003; Maciak *et al.*, 2014).

Such studies, however, require a genome-wide approach. In the absence of high-quality reference genome, which remains cost-prohibitive for complex genomes, whole-genome marker discovery enabled by restriction site-associated DNA sequencing presents a cost-effective method to study species on a population scale even with no previous genetic and genomic resources available.

Previous studies on *Apodemus* typically employed a small number of microsatellite (Rico *et al.*, 2009; Czarnomska *et al.*, 2018) and mtDNA markers (Michaux *et al.*, 2003; Michaux *et al.*, 2004; Michaux *et al.*, 2005; Herman *et al.*, 2017), which are insufficient to learn about the species' population structure, admixture patterns and to identify loci under selection in appropriate detail. Here we employ, for the first time in *Apodemus*, the whole-genome, high-density genotyping using double-digest restriction site-associated DNA sequencing (ddRAD-seq) to elucidate the genetic structure and connectivity of three populations of *A. flavicollis* and compare it to a population of *A. sylvaticus* in Poland. We demonstrate clear divergence between the two species and very low differentiation within populations of *A. flavicollis*, suggesting wide-range gene flow within this species. Our results provide the first whole-genome-based estimate of population structure in *A. flavicollis* divergence between the two *Apodemus* species, as well as a selection of loci enabling their accurate identification.

2.2 Materials and methods

2.2.1 Sample collection and DNA extraction

Eighty two individuals (10 *Apodemus sylvaticus* and 72 *Apodemus flavicollis*) from four locations in north-eastern Poland spanning 500km were trapped between 2012 and 2015 (Figure 2.1). *A. flavicollis* were collected in Białowieża (E23.8345814, N52.7231935), an oak-lime-hornbeam forest (n = 35), Bory Tucholskie (E17.5160265, N53.7797608), in an oak-lime-hornbeam and pine forest (n = 23) and Haćki (E23.1793284, N52.834369), in a xerothermic meadow (n = 14). *A. sylvaticus* were trapped in Kadzidło (E21.3778496, N53.2089113) in a dry pine forest (n = 5) and in Bory Tucholskie, mainly in a pine forest (n = 5) (Appendix A, Section A.1, Table A.1). While A. flavicollis are present in all sampled locations, there have been no trappings of A. sylvaticus in Białowieza for the last 20 years, despite Białowieza being within the European range of this species (Dr. Karol Zub, personal communication). The sampling procedures were approved by the Local Ethical Commission on Experimentation on Animals in Białystok, Poland, under permission number 2015/99.

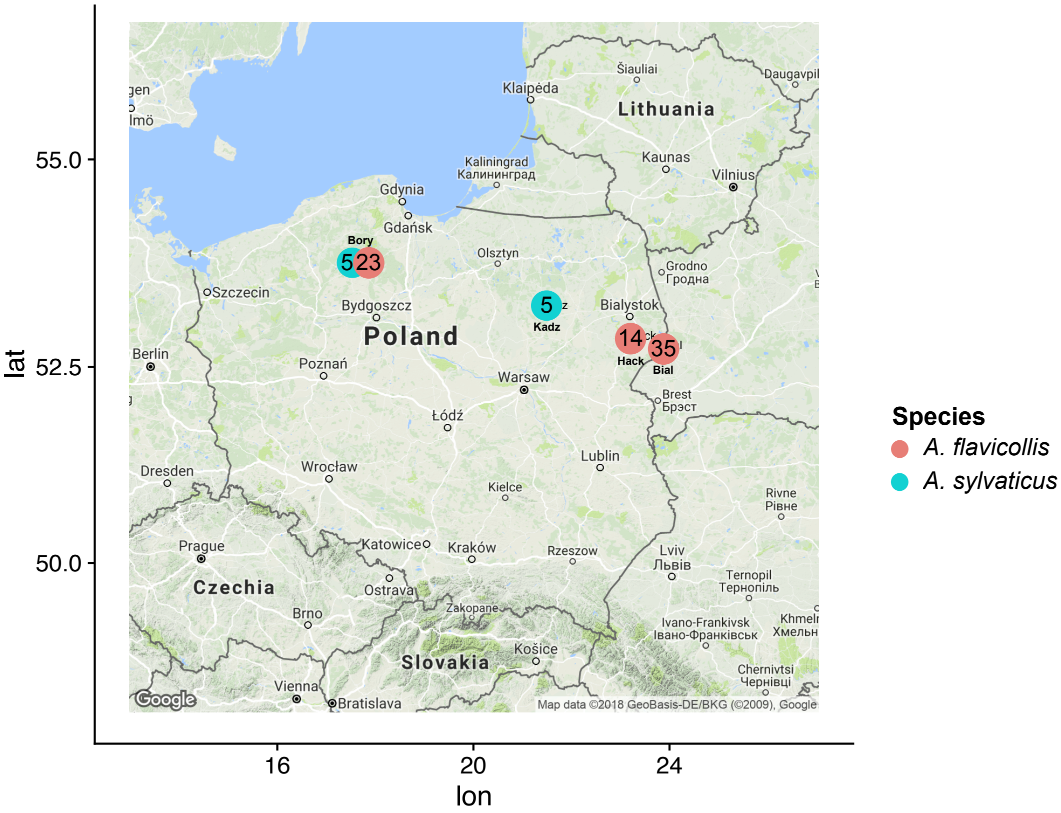


FIGURE 2.1: Locations of the Polish samples. Red circles represent samples from *Apodemus flavicollis* while blue dots represent samples from *Apodemus sylvaticus*. The number inside the circles are the number of samples from each locality.

Tail clippings were collected, preserved in $\geq$ 95% ethanol and stored at -20°C until DNA extraction. The tissues were digested by incubating at 55°C overnight with lysis buffer (10mM Tris, 100mM NaCl, 10mM EDTA, 0.5% SDS) and proteinase K (20mg/ml). Subsequently, potassium acetate and RNAse A were used to remove protein and RNA contamination. Three ethanol washes were performed using Sera-Mag SpeedBeads solution (GElifesciences, Marlborough, MA, USA). The quality and integrity of the DNA was tested in a 2% agarose gel. Twenty-fold dilutions of the samples were used to measure the DNA concentration using Quant-iT PicoGreen dsDNA assay kit (Invitrogen, Carlsbad, CA, USA) and concentration of each sample was then normalised to 10 ng/μl in 20μl volume. Four samples were used as technical duplicates (F06-B02, G02-D01, H11-G06, F12-A12). Technical duplicates had the same DNA but were digested and ligated to barcodes independently.

2.2.2 ddRAD-seq library preparation

ddRAD-seq library was prepared following the protocol from Poland and Rife (2012), adapted to utilise a different combination of enzymes. Briefly, genomic DNA was digested in a 20 μl reaction with CutSmart® buffer, 8 units of *SbfI* and 8 units of HF-*MseI* (New England Biolabs, Frankfurt am Main, Germany). Digestion was performed at 37°C for 2 hours. Enzymes were inactivated at 65°C for 20 minutes and the reactions were kept at 8°C. Adapter ligation was performed at 22°C for 2 hours and the ligase was inactivated by incubating the samples at 65°C for 20 minutes. Samples were cooled down to 8°C and multiplexed by combining 5μl of each sample. P1 adapters contained barcodes with a length between 5 and 10 bp (Appendix A, Section A.2, Table A.2).

PCR amplification was conducted in 25μl with 1μl of each primer at 10mM, 0.5μl of 10 mM dNTPs, 13.25μl of PCR-grade water, 5μl of 5x Phusion HF Buffer, 0.25μl of Phusion DNA Polymerase (New England Biolabs, Frankfurt am Main, Germany) and 4μl of the multiplexed DNA. After an initial denaturation step of 30s at 98°C, PCR reaction was carried out for 12 cycles (10s at 98°C, 20s at 58°C and 15s at 72°C). Final elongation step was performed at 72°C for 5 minutes.

PCR products were loaded into a single lane on a 1% agarose gel with 100 bp DNA ladder (New England Biolabs, Frankfurt am Main, Germany). Fragments between 200 and 500 bp were cut from the gel with a scalpel and purified using the QIAquick gel extraction kit (QIAgen, Hilden, Germany), followed by a second cleanup step with Sera-Mag SpeedBeads (GElifesciences, Marlborough, MA. USA). Sizing, quantification and quality control of the DNA was performed using Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) before paired-end sequencing on an Illumina HiSeq 3500 with 151 bp read length.

2.2.3 Processing of RAD-tags

Sequences were analysed with Stacks version 1.48 (Catchen *et al.*, 2011). Samples were demultiplexed using process_radtags allowing no mismatches in barcodes and cutting sites. Sequences with uncalled bases and low quality scores were removed and all reads were trimmed to 141 bp. The four files generated per sample by process_radtags were concatenated using a custom bash script. The best parameters for building and calling SNPs *de novo*, using denovo_map, were calculated following Paris *et al.* (2017) approach, using either samples from both species or only from *A. flavicollis*. Secondary reads were not used to call haplotypes in denovo_map (option -H).

2.2.4 Genotyping error rates

We estimated the error rates by analysing a set of four samples that were prepared and sequenced in duplicates. Sequences for 52494 loci from both species, were extracted using –fasta_samples option from the populations package in Stacks. We extracted sequences for each of the duplicated samples with a custom script and calculated different errors as described by (Mastretta-Yanes *et al.*, 2015). Briefly, the locus error rate is the percentage of missing data at the locus level, calculated by dividing the number of loci found only in one of the duplicates by the total number of loci in each sample. The allele error rate is the percentage of mismatches between the International Union of Pure and Applied Chemistry (UIPAC) consensus sequences between homologous loci from each pair of duplicates. These sequences, which code heterozygotes with special characters from the UIPAC code, in the same way than degenerate bases are coded, were built using Consensus.pl script (Hughes, 2011). Finally, the two SNP error rates: the percentage of different SNPs called in each of the duplicated samples using either all 10178 SNPs or using the SNPs called with missing data between duplicate samples excluded (see Table 2.1).

2.2.5 Variant calling and filtering

We first combined the data from *A. sylvaticus* and *A. flavicollis* to establish species differentiation. We filtered the SNPs using the populations package from Stacks (Catchen *et al.*, 2011) and VCFtools (Danecek *et al.*, 2011). We retained SNPs common to 80% of the individuals in each species (p=1, r=0.8) and excluded SNPs with minor allele frequencies MAF<0.05 and which deviated from Hardy-Weinberg equilibrium (HWE) at P<0.05. We also removed sites with mean depth values lower than 20. We manually modified the chromosome numbers in the vcf file to input it into SNPhylo (Lee *et al.*, 2014), which we used to build a maximum likelihood tree. This program runs

DNAML programs from the PHYLIP package in order to generate the tree (Lee *et al.*, 2014). We set a missing rate (-M) of 1, minor allele frequencies (-m) of 0, linkage disequilibrium threshold (-l) of 1 and the -r option to skip the step of removing low quality data. Confidence values were estimated using 1000 bootstrap replicates. The root was manually fixed to separate both species. Principal Component Analysis (PCA) was performed using the R package Adegenet (Jombart, 2008).

The catalogue generated by Stacks on Polish samples was tested for its ability to differentiate the two species with an extra set of samples from other locations in Europe. Ten *Apodemus flavicollis* (2 samples from Austria, 5 from Lithuania and 3 from Romania) and 10 samples of *Apodemus sylvaticus* (4 samples from Wales, 3 from Tunisia and 3 from Scotland) were kindly provided by Dr Jeremy Herman, National Museums Scotland, Dr Johan Michaux, University of Liege and Dr Karol Zub, Mammal Research Institute of the Polish Academy of Sciences (MRI) (Information about samples in Appendix A, Section A.8). We considered all the 20 test samples as a different group from Polish *Apodemus sylvaticus* and *Apodemus flavicollis* for SNP calling. We kept SNPs common to the 80% of the individuals in each group (p=1, r=0.8) and excluded SNPs with minor allele frequencies MAF<0.05, SNPs which deviated from the Hardy-Weinberg equilibrium (HWE) at P<0.05, sites with mean depth values lower than 20 and with more than 5% of missing data.

2.2.6 Population divergence

To analyse genetic diversity and population connectivity within *A. flavicollis*, we analysed the three populations (Bory Tucholskie, Białowieża and Haćki) separately (p=3, r=0.8), while keeping the other parameters as described above. Due to the lack of outgroup, a mid-point root was chosen in the phylogenetic tree. Individual ancestries were estimated following a maximum likelihood approach with ADMIXTURE (Alexander *et al.*, 2009), after conversion of the VCF file to ped with plink version 1.9 (Chang *et al.*, 2015; Purcell and Chang, 2018). ADMIXTURE analysis were run 10 times, from K=1 to K=5, each using 10 different seeds. Weighted (Weir-Cockerham) Fst was calculated with VCFtools v0.1.13. Heterozygosity, Pi and Fis were calculated with the population package from Stacks (Catchen *et al.*, 2011).

2.2.7 Species divergence

To calculate the divergence between the two species, a set of 21377 common loci were extracted with a custom script and the strict consensus sequences for each species were calculated with Consensus.pl script (Hughes, 2011). Sequence divergence was then calculated using a custom R script.

2.2.8 Effect of group size

In order to evaluate the effect of unequal number of samples on the estimation of genetic diversity parameters and Admixture analysis, 100 permutations with resampling of 15 individuals per population were performed.

2.3 Results

2.3.1 Sequencing and variant calling

The sequencing produced a total of 92741120 reads. The number of reads per individual varied from 346810 to 4157586, with an average of 1078385 reads per individual and median of 905786,5. The best parameters for calling the stacks and variants for the entire dataset were: minimum number of identical, raw reads required to create a stack m = 2, number of mismatches allowed between loci for each individual M = 4 and number of mismatches allowed between loci when building the catalogue n = 5 (Appendix A, Section A.3, Figure A.1). The best parameters calculated for *A. flavicollis* samples only were: m = 2, M = 4 and n = 3 (Appendix A, Section A.4, Figure A.3). The coverage per sample ranged from 4.95x to 26.20x with an average of 10.13x and median of 9.32x for the entire dataset (Appendix A, Section A.3, Figure A.2 and Section A.4, Figure A.4).

2.3.2 Error rates

Analysis of the duplicated samples showed that loci and allele misassignment rates were of similar magnitude, on average, between all the pairs of duplicates. The duplicate pair F06-B02 showed the highest discrepancy between loci of 10%, and also between alleles, of 8%. When only shared loci were included in the comparisons, all four sets of duplicates showed on average 0.5% ± 0.2% SNPs called differently (Table 2.1). The difference on the number of reads obtained for each pair of duplicates, as indicated by D1/D2 proportion on Table 2.1, show a high discrepancy on the pairs of duplicates F06-B02 and A12-F12, were one of the duplicates was sequenced more intense than the other. Pairs of duplicates with a disproportionate number of reads are the ones with the highest error rates.

TABLE 2.1: Error rates calculated by comparing four sets of duplicated samples. For explanation of different errors please see Methods. D1/D2: ratio of reads from Duplicate 1 to Duplicate 2.

	F06-B02	**A12-F12**	**H11-G06**	**G02-D01**	**MEAN**	**SD**
Reads (D1/D2)	0.19	3.54	1.29	1.380		
Coverage	8.9/11.2	15.9/10.2	8.0/10.5	7.6/8.5		
Locus error rate	0.10	0.08	0.04	0.04	0.07	0.031
Allele error rate	0.08	0.06	0.07	0.05	0.07	0.01
SNP error rate 1	0.15	0.12	0.05	0.07	0.10	0.04
SNP error rate 2	0.006	0.004	0.007	0.004	0.005	0.002

2.3.3 Comparison of *A. flavicollis* and *A. sylvaticus*

The number of assembled loci per individual ranged from 46286 to 117366 (mean: 73711, median: 71395, standard deviation: 29917).

52494 loci passed the population filters established for species differentiation, representing 8,3% of the total 632063 loci included in the catalogue. Out of 158144 SNPs called, 60366 (38.1%) were removed after filtering for MAF and 52298 (33%) were removed after failing the HWE test at p<0.05; further 35302 (22.3%) were removed due to a minimum mean depth lower than 20, leaving 10178 SNPs (6.6%) to be used in the downstream analyses (Figure 2.2).

PCA plot (Figure 2.3) of the first two components, accounting for 13.13% of the total variance, shows differentiation of the two species but also distinguish different populations of *A. flavicollis*. Similarly, the phylogenetic tree shows *A. sylvaticus* as a separate clade to the three populations of *A. flavicollis*, with *A. flavicollis* from geographically closer regions (Białowieża and Haćki, 50 km) grouped closer than a population from Bory Tucholskie, 450 km away from Białowieża (Figure 2.4). The *A. sylvaticus* and *A. flavicollis* clusters have high bootstrap value support (100% and 99% respectively).
We then investigated the suitability of the SNPs we identified on Polish populations to distinguish *A. sylvaticus* and *A. flavicollis* from other European populations.

The genotypingof the extra ten samples from each species (see Methods) produced 179763 SNPs. 62158 (34.58%) were removed after filtering for MAF and 69125 (38.45%) were removed after failing the HWE test at p<0.05; further 42054 (23.39%) were removed due to a minimum mean depth lower than 20 and 5203 (2.89%) were removed due to more than 5% missing data, leaving 1223 SNPs (0.68%) to be used in the downstream analyses.

The first two axis of the PCA plot (Figure 2.5) constructed from this data accounts for the 65.73 % of the total variance and shows clear differentiation between the two species from other European populations. Moreover, all the *A. flavicollis* samples cluster with the Polish *A. flavicollis* samples, while all but Tunisian samples of *A. sylvaticus* cluster with the Polish samples of the same species. Tunisian *A. sylvaticus* appear as a separate cluster but nevertheless still clearly closer to the *A. sylvaticus* group.

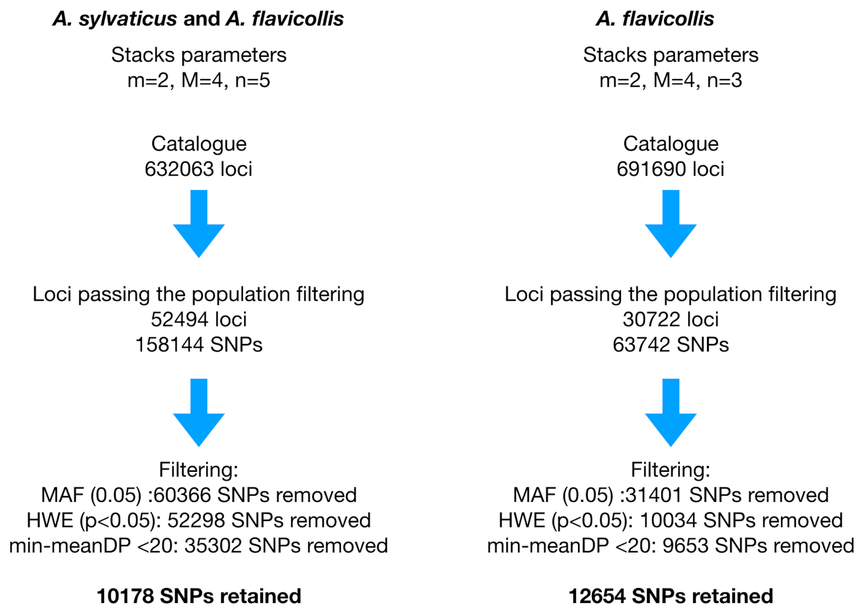

A. sylvaticus and A. flavicollis
Stacks parameters
m=2, M=4, n=5
Catalogue
632063 loci
Loci passing the population filtering
52494 loci
158144 SNPs
Filtering:
MAF (0.05) :60366 SNPs removed
HWE (p<0.05): 52298 SNPs removed
min-meanDP <20: 35302 SNPs removed
10178 SNPs retained
A. flavicollis
Stacks parameters
m=2, M=4, n=3
Catalogue
691690 loci
Loci passing the population filtering
30722 loci
63742 SNPs
Filtering:
MAF (0.05) :31401 SNPs removed
HWE (p<0.05): 10034 SNPs removed
min-meanDP <20: 9653 SNPs removed
12654 SNPs retained

FIGURE 2.2: Summary of cataloque construction and SNP filtering steps for the complete dataset (left) and Apodemus flavicollis dataset (right). The graphic includes: Stacks parameters values (m, M, n), number of loci in the catalogue, number of SNPs filtered by minor allele frequency (MAF), which failed the Hardy-Weinberg equilibrium test at p<0.05 (HWE), SNPs removed due to an average depth, across individuals, lower than 20 (min-meanDP) and the total number of SNPs retained for further analysis

The catalogue of tags, alleles and SNPs used for species differentiation is included in the Appendix A, Section A.6.

2.3.4 Genetic diversity and population structure of *A. flavicollis*

The number of assembled loci per individual in the Polish populations ranged from 46286 to 117366 (mean: 72738, median: 70592, stdev: 12575). 30722 loci passed the population filters established for population differentiation, representing and 4.43% of the total 691960 loci included in the catalog. Out of 63742 SNPs called, 31401 (49.26%) were removed after filtering for MAF and 10034 (15.74%) were removed after failing the HWE test at $p<0.05$. Further 9653 (15.14%) were removed due to a minimum mean depth lower than 20, leaving 12654 (19.85%) SNPs to be used in the downstream analyses.

PCA plot (Figure 2.6) shows differentiation between the three Polish *A. flavicollis* populations, with PC1 and PC2 cumulatively explaining 10.47% of

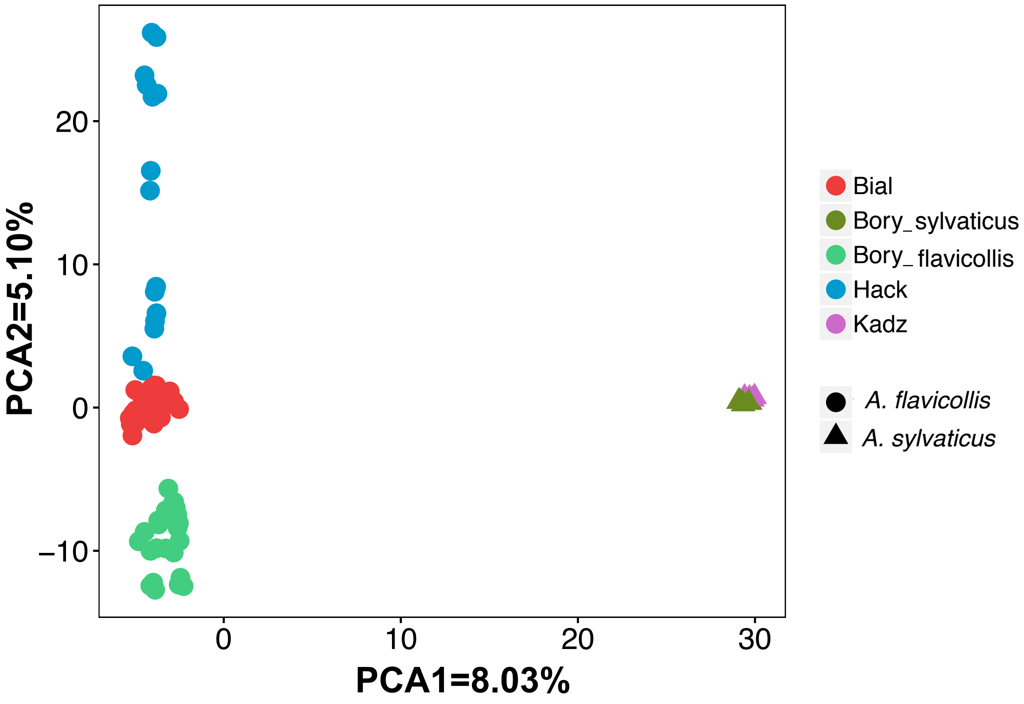


FIGURE 2.3: Principal Component Analysis of all samples analysed in the study. Each point represents one sample; the shape of the point represent the species (circles: *Apodemus flavicollis*, triangles *Apodemus sylvaticus*, whereas the colour represents the location where the samples were collected.

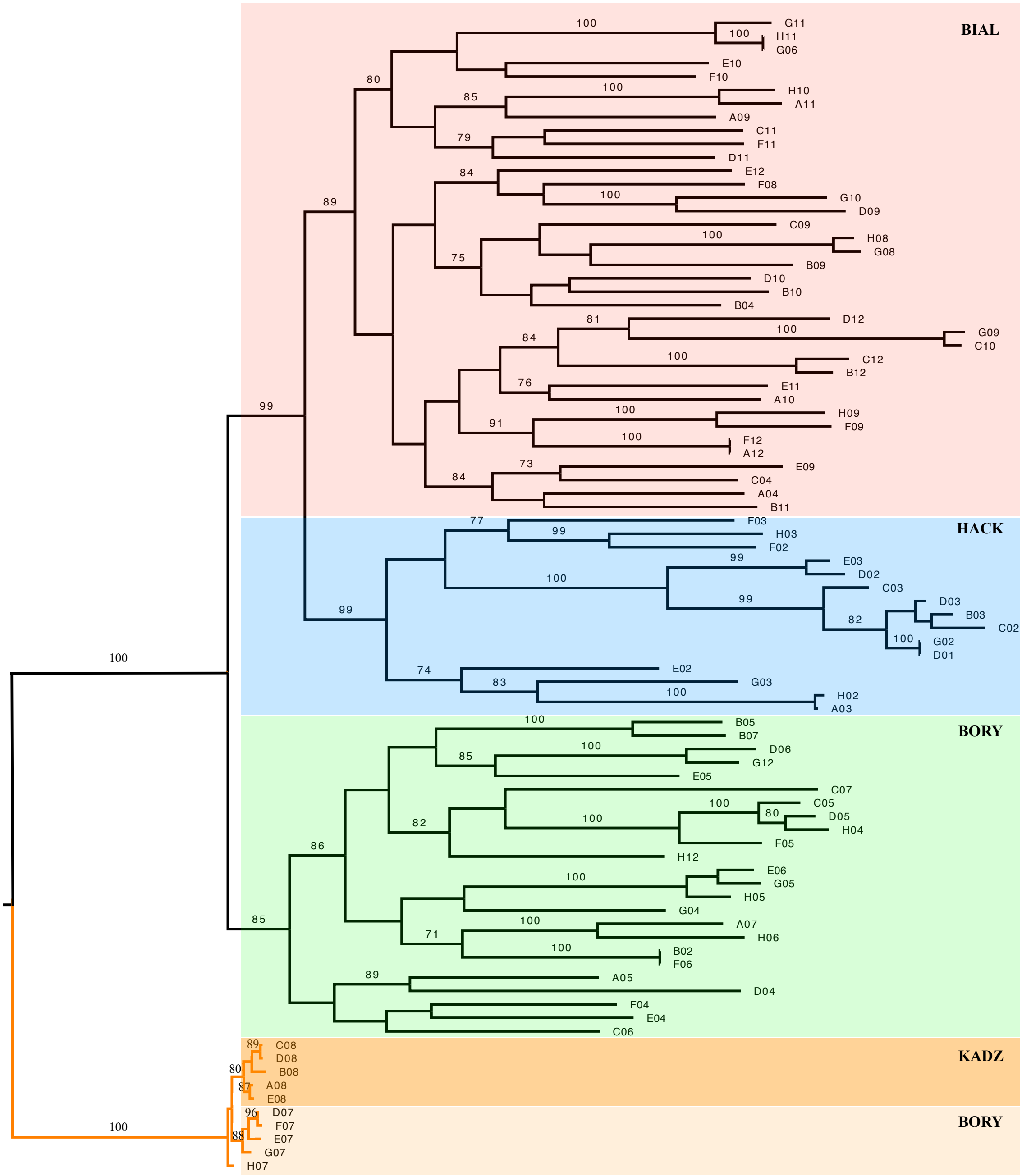


FIGURE 2.4: Maximum likelihood phylogenetic tree of all the samples analysed in the study. Line colour represents the species: A. sylvaticus (n=10) in orange and A. flavicollis (n=72 + 4 duplicates) in black. Colour boxes represent the different populations: Bial - Białowieża, Kadz - Kadzidło, Hack - Hacéki, Bory - Bory Tucholskie. Duplicates samples are included: F06-B02 from Bory Tucholskie, F12-A12 and H11-G06 from Białowieża and G02-D01 from Hacki. Bootstrap support values from 100 replicates are indicated at the branches of the tree. Only bootstrap values higher than 70% are shown.

the total variance. Haćki population shows higher diversity than the other populations, with some Haćki individuals closer to Białowieża individuals than to others from this location. The phylogenetic tree (Figure 2.7), due to the lack of an outgroup, can not indicate the relationship between the different populations, but shows each population as a monophiletic group. Bory Tucholskie and Haćki populations each form a cluster with 100% bootstrap support value, whereas Białowieża forms a cluster with 95% bootstrap support.

In the ADMIXTURE analysis, the lowest cross-validation errors (Alexander and Lange, 2011) were always found at K=3 (Appendix A, Section A.5, Figure A.5) indicating contribution of three ancestral populations (Figure 2.8). The majority of samples from each of the populations show a single dominant component of ancestry with little contribution from other populations, with the exception of four individuals from Haćki, which show clear admixture of the Białowieża population.

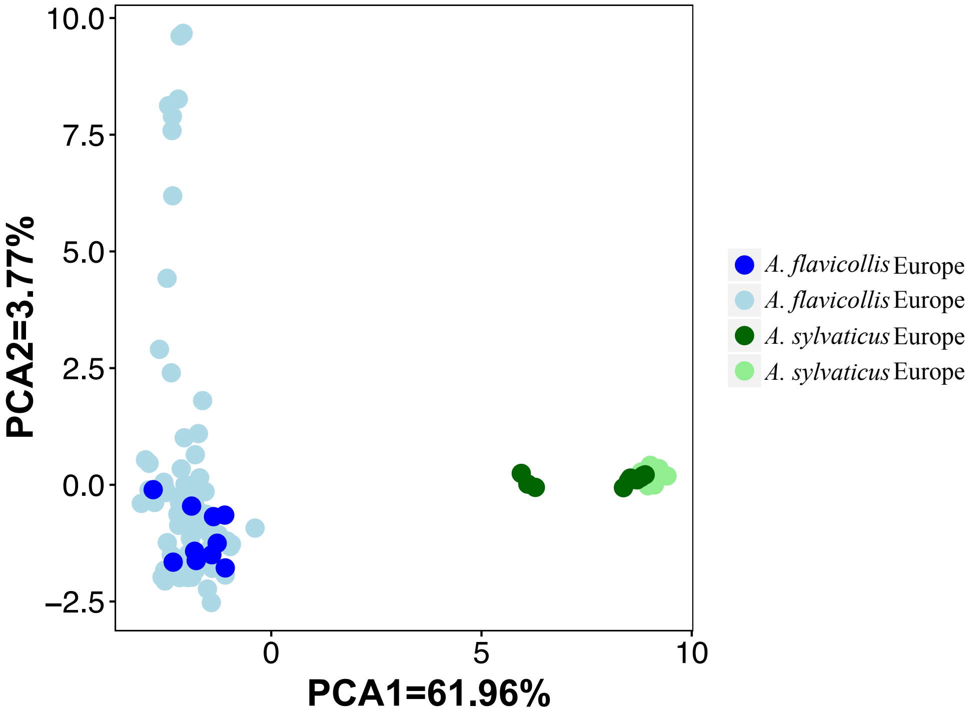


FIGURE 2.5: Species identification through Principal Component Analysis. Light colours represent the Polish samples while dark colours represent samples from other European regions and Tunisia (collectively named "Europe"). Green: *Apodemus sylvaticus*, blue: *Apodemus flavicollis*

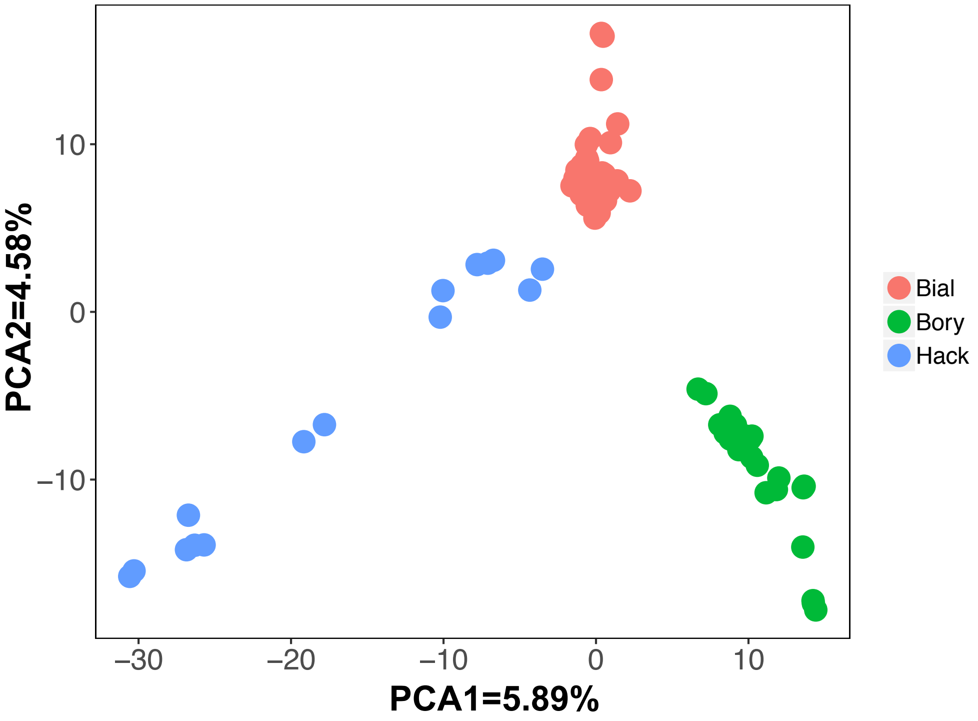


FIGURE 2.6: PCA plot showing Polish samples of A. flavicollis from Białowieża (red) (n=35 + 2 duplicates), Hacéki (blue) (n=14 + 1 duplicate) and Bory Tucholskie (green) (n=23 +1 duplicate). Bial - Białowieża, Bory - Bory Tucholskie , Hack - Haćki

Recognising that STRUCTURE-type analyses (on which ADMIXTURE is based) may be sensitive to the effects of uneven number of samples in compared groups (Puechmaille, 2016), we repeated the ADMIXTURE analysis 10 times, each time randomly drawing the same number of individuals (n = 15) from each population. In all cases, the lowest cross-validation errors were found for K = 2, followed by K = 3 (Appendix A, Section A.5, Figure A.6). At even sampling, ADMIXTURE pattern found for K = 3 was the closest to the observed ecological and geographical distribution of the samples and closely matched our results when all samples were included.

The patterns of heterozygosity point out to Haćki as the only population where the values of observed heterozygosity (H_o) is higher than the expected heterozygosity (H_e), where the inbreeding coefficeint (F_{IS}) is negative (Table 2.2). As parameters such as the number of private alleles, nucleotide diversity and heterozygosity vary with sample size, we performed 1000 calculations of the above parameters using random samplings of the same number

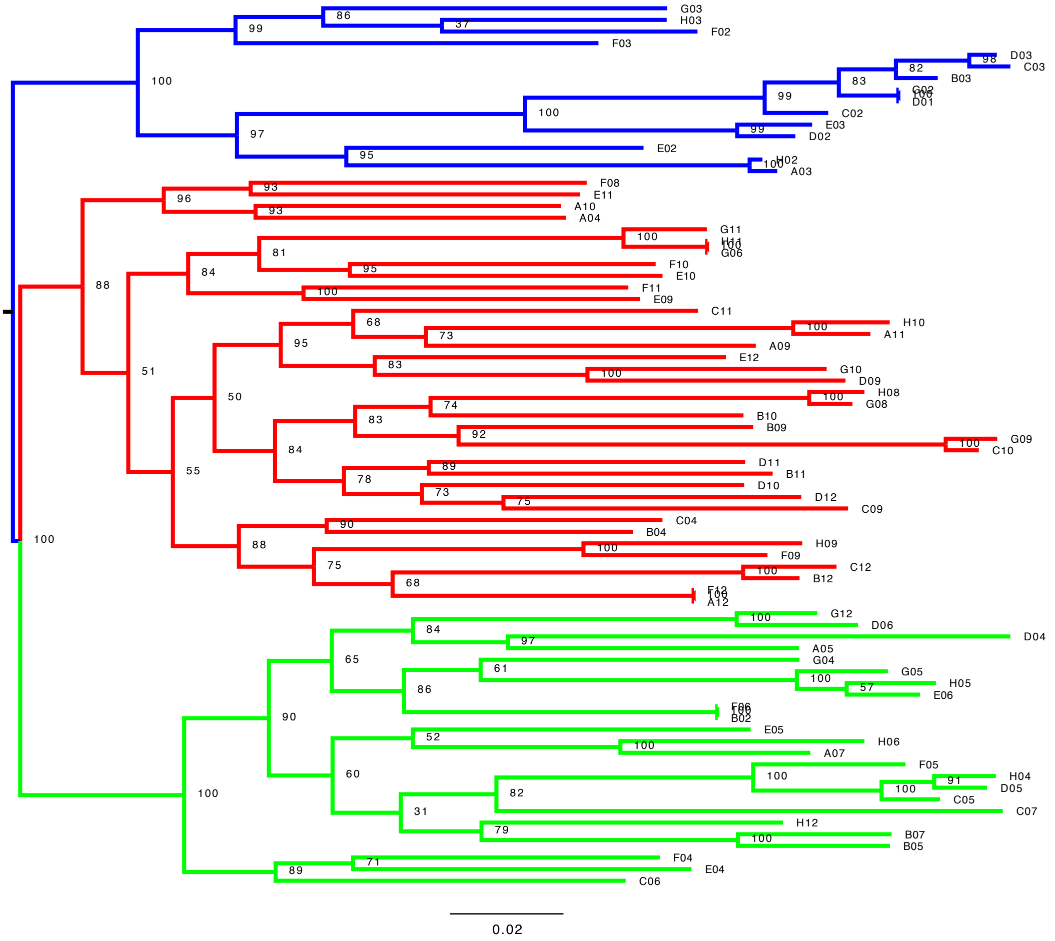


FIGURE 2.7: Maximum likelihood phylogenetic tree of n = 72+4 *A. flavicollis* samples from Białowieża (red, n = 35 + 2 duplicates), Haćki (blue, n = 14 +1 duplicate) and Bory Tucholskie (green, n = 23 +1 duplicate). Bootstrap support values (from 100 replicates) are indicated at the nodes of the tree.

of individuals from every population. The parameters showed similar relationships except for the number of private alleles (data not shown).

F_{st} values are consistently very low between all the populations, even though populations from Haćki and Bory Tucholskie show two-fold higher F_{st} values that for the other two pairs of populations (Table 2.3).

2.3.5 Species divergence

Finally, the calculated average divergence between *A. flavicollis* and *A. sylvaticus* based on 21377 shared loci is 1.51% (min=0%, max= 6.38%, median= 1.42%, stdev= 1.10%). We also identified 117 loci with divergence larger than 4.9% (The loci ID are provided in the Appendix A, Section A.7, Table A.3)

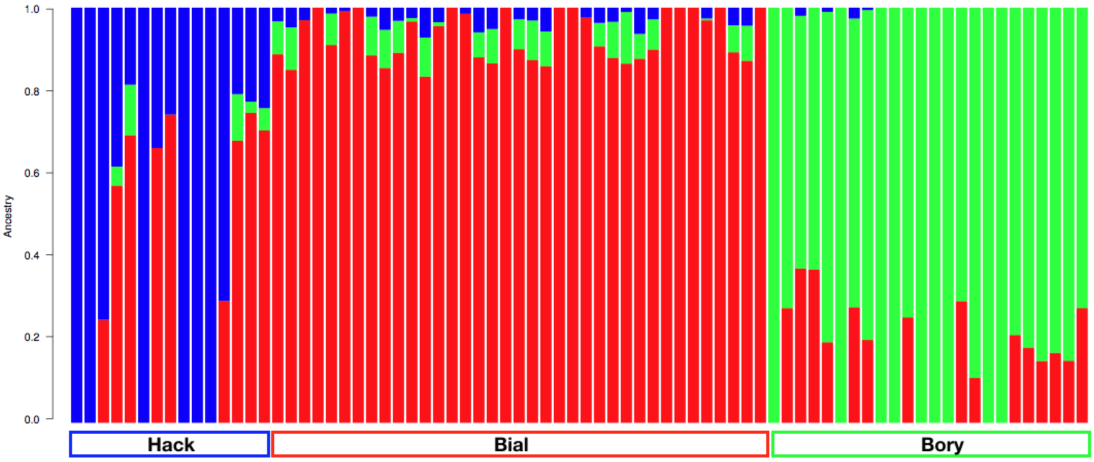


FIGURE 2.8: Maximum likelihood Admixture analysis of all *A. flavicollis* samples for the optimal K = 3. Each bar represents an individual and each colour represents its ancestry (red: Białowieża, blue: Haćki, green: Bory Tucholskie).

TABLE 2.2: Genetic diversity parameters calculated based on 12654 SNPs from all 72 individuals (+4 Duplicates) of *A. flavicollis*. N, number of individuals; Npa, number of private alleles; Ind per loci, Mean number of individuals per locus in this population; H_o observed and H_e expected heterozygosity; π, average nucleotide diversity; F_{IS} inbreeding coefficient

Pop ID	N	Npa	Ind per loci	Obs Het	Exp Het	Pi	Fis
Haćki	15	32	14.42	0.30	0.27	0.28	-0.04
Bory Tucholskie	24	74	22.93	0.28	0.28	0.29	0.02
Białowieża	37	148	35.13	0.29	0.30	0.30	0.01

TABLE 2.3: Pairwise F_{ST} values for the three populations of *A. flavicollis*

	Bory Tucholskie	Białowieża
Haćki	0.085	0.055
Bory Tucholskie		0.045

and checked whether these loci alone allow for accurate assignment of samples to the two species. We constructed PCA plots from the Polish samples only and from the Polish, other European and Tunisian samples together. They demonstrate that while the 117 loci are sufficient to clearly assign Polish samples to the two species (Appendix A, Section A.7, Figure A.7), some uncertainty remains when we use these loci for the broader set of samples.

Whereas all A. flavicollis samples do cluster together, A. sylvaticus samples do not form a clearly differentiated group (Appendix A, Section A.7, Figure A.8).

2.3.6 Effect of group size

Permutations performed for the calculations of genetic diversity parameters (Table 2.4) have shown that with the exception of the number of private alleles, the results are comparable, regardless of the number of samples included per population.

,

TABLE 2.4: Average genetic diversity parameters for *Apodemus flavicollis*calculated from 100 permutations of 45 individuals each one (15 samples per population, 12654 SNPs). N, number of individuals; Npa, number of private alleles; Ind per loci, Mean number of individuals per locus in this population; H_o observed and H_e expected heterozygosity; π, average nucleotide diversity; F_{IS} inbreeding coefficient

Pop ID	**N**	**Npa**	**Ind per loci**	**Obs Het**	**Exp Het**	**Pi**	**Fis**
Haćky	15	115.53	14.42	0.31	0.28	0.29	-0.05
Bory Tucholskie	15	183.95	14.33	0.29	0.28	0.29	0.02
Białowieża	15	204.84	14.23	0.30	0.29	0.31	0.02

The F_{ST} values obtained when using equal number (Table 2.5) of samples per population are even lower than before, with the smallest differentiation found this time between the populations of Bory-Tucholskie and Białowiezża.

TABLE 2.5: Pairwise F_{ST} values for the three populations of *A. flavicollis* including equal number of samples

	Bory Tucholskie	**Białowieża**
Haćky	0.086 ± 0.002	0.057 ± 0.002
Bory Tucholskie		0.045 ± 0.002

2.4 Discussion

RAD-sequencing approaches, including double-digest RAD-seq and its variants (Miller *et al.*, 2007; Baird *et al.*, 2008; Poland and Rife, 2012; Peterson *et al.*, 2012; Franchini *et al.*, 2017), have allowed a cost-effective discovery of thousands of genetic markers in both model and non-model organisms

(Rodriguez-Ezpeleta *et al.*, 2017; Hammerman *et al.*, 2018), proving to be a transformative research tool in population genetics (Blanco-Bercial and Bucklin, 2016; Cromie *et al.*, 2013; Hohenlohe *et al.*, 2013), phylogeography and phylogenetics (Jeffries *et al.*, 2016; Reitzel *et al.*, 2013; Alter *et al.*, 2017; Hipp *et al.*, 2014), marker development (Pegadaraju *et al.*, 2013), linkage mapping studies (Baxter *et al.*, 2011), species differentiation (Pante *et al.*, 2015) and detecting selection (Shultz *et al.*, 2016). However, despite the widespread use of this approach to genome-wide marker discovery, only few studies have used RAD-seq in mammals (Fernández *et al.*, 2016; Lanier *et al.*, 2015; Knowles *et al.*, 2016; Moura *et al.*, 2014; Shafer *et al.*, 2017). Here, we have identified over 10000 markers in two closely related and common species of *Apodemus* in western Palearctic, characterised the population structure of *A. flavicollis* and compared it to *A. sylvaticus*, for the first time providing genome-wide estimates of the species divergence and population genetic parameters.

2.4.1 Technical considerations

We have used four pairs of technical duplicates to check the accuracy of the RAD-seq genotyping based on the Poland protocol (Poland *et al.*, 2012). By far the biggest source of discrepancy in SNP calls between the duplicates is caused by unequal identification of loci. SNPs error rate, considering all the loci sequenced per sample averaged approximately 10% (2.1). However, when considering only shared loci between the duplicates, the discrepancy in SNP calls fell by over an order of magnitude to an average of 0.5%, indicating high accuracy and reliability of calls in once-defined shared loci. Our finding of loci calls being the major source of genotyping error agrees with Mastretta-Yanes *et al.* (2015), although our discrepancies are almost an order of magnitude smaller. Moreover, despite the differences in number of loci included in the analysis, each duplicated pair of samples clustered together with a 100% bootstrap values support and branch length equal to 0 on the phylogenetic tree (Figure 2.7), indicating that the samples were identical. Overall, our finding reiterates the importance of the influence of stochastic events and imprecise size selection in the library preparations on genotyping calls (Mastretta-Yanes *et al.*, 2015); we note that some of these variables could be better controlled with more automated size-selection approaches (Peterson *et al.*, 2012). It also illustrates the usefulness of including technical replicates during library preparation.

2.4.2 Population structure

The F_{ST} values calculated in this study between all three pairs of populations of *A. flavicollis*, based on 12654 SNPs, are consistently low. Previous studies of *A. flavicollis* populations in north-eastern Poland based on a small number of microsatellites showed similarly and consistently low values (Gortat *et al.*, 2010; Czarnomska *et al.*, 2018), even though Gortat *et al.* (2010) suggested some population structure based on statistically significant differences between very low pairwise F_{ST} values. Czarnomska *et al.* (2018), also suggest large, broadly geographically defined clusters of *A. flavicollis* in north-eastern Poland that are separated by highly admixed individuals, but, again, F_{ST} between those clusters are as low as those reported by Gortat *et al.* (2010) and this study.

We would argue, based on a much larger set of genome-wide markers reported here, that *A. flavicollis* has a very limited population structure across the entire area studied. Large number of markers nevertheless allows us to discover evidence for admixture of Białowieża population and Haćki (Figure 2.8), further indicated by relatively high heterozygosity and negative F_{IS} in this population. It is therefore intriguing that such a low differentiation occurs across hundreds of kilometres of varying landscape in a species that typically has a limited range of about 4 km and that suffers close to 90% winter mortality rate (Pucek *et al.*, 1993), which would typically lead to multiple bottlenecks and drift-driven population differentiation. With this in mind, our data suggests a much larger dispersal ability of the species, a much better connectivity between populations, or both and the existence of an effectively single population of *A. flavicollis* in north-eastern Poland.

The heterozygosity values reported in this study are smaller than in previous work by Gortat *et al.* (2010) and Czarnomska *et al.* (2018). They range from 0.27 to 0.30, in comparison to ranges between 0.841 to 0.877 in (Czarnomska *et al.*, 2018) and 0.56 and 0.7 in (Gortat *et al.*, 2010) for most (but not all) of their markers. However, as their work was based on relatively few microsatellites, these differences reflect the higher variability of microsatellites compared to SNPs (Hauser *et al.*, 2011; Fischer *et al.*, 2017; Fernández *et al.*, 2013).

Both low overall F_{ST} and moderate heterozygosity data suggest it would be worthwhile to conduct a genome-wide scan for selection using F_{ST} as a metrics of local genomic differentiation to identify potential, geographically local regions under selection. This, however, is not yet feasible given the lack of high-quality reference genome for *Apodemus*.

2.4.3 Divergence and differentiation of *A. flavicollis* and *A. sylvaticus*

Given that accurate identification of the two species using morphological characters is problematic, especially in their southern range (Bugarski-Stanojević *et al.*, 2013), a large collection of markers identified in this study allowed us to create a catalogue of 632063 loci and 1226 SNPs, which, after filtering, allow for a clear differentiation between species. This identification is somewhat biased, as the catalogue was built using more samples of *A. flavicollis* than *A. sylvaticus* (72 vs 10) and both from a relatively limited geographical range. Nevertheless, it allowed for accurate assignment of species, as we demonstrated on a set of 20 independent samples from other European countries and Tunisia (Figure 2.5). Given the wide distribution of both species in western Palearctic, a more representative sample from both species from a broader geographic range would provide even more accurate set of markers for their identification.

Finally, we calculated the nucleotide divergence between the two species, based on 21377 shared loci, which is 1.51%. Considering a divergence time between *A. flavicollis* and *A. sylvaticus* estimated from archeological data of 4 Ma (Michaux *et al.*, 2003), the evolution rate is 0.0019 substitutions per site per million of years. This estimate of sequence divergence level is in broad agreement with calculations based on mitochondrial 12S rRNA, IRBP and Cytochrome b genes (Michaux *et al.*, 2002). However, as we only used shared loci to calculate divergence, it is likely an underestimate as it does not include the potential impact of insertion/deletion events, which can significantly affect the total genomic divergence between species (Li *et al.*, 1987; Britten, 2002).

2.5 Conclusions

We have successfully applied the ddRad-seq approach to discover tens of thousands of SNPs in widespread and common mammalian species of *A. flavicollis* and *A. sylvaticus*, which has been underdeveloped in terms of genomic resources available for its study. The high resolution data obtained here allowed us to distinguish geographically close populations but suggest that *A. flavicollis* effectively forms a single population in an entire sampling area that spans 500 km in the W-E direction. Comparing *A. flavicollis* and *A. sylvaticus*, we have calculated their genome-wide divergence and identified a set of 632063 loci and 1226 SNPs that enable effective molecular identification of the species. We anticipate that with the development of further whole-genome resources, *Apodemus*, thanks to its common status, broad geographic range and long history of ecological observations, will become an excellent model species for evolutionary and ecological research in the genomic era.

Chapter 3

Protocol troubleshooting and new adapter design

3.1 Introduction

The purpose of this chapter is to present the technical analysis of the performance of the library preparation and sequencing, carried out for the pilot project on the *Apodemus* population structure in north-eastern Poland (described in Chapter 2). It describes the issues related to reads' retention that led to the design of an alternative library preparation protocol and the design of custom adapters for the main project of this thesis, on the European phylogeography of *Apodemus flavicollis* and *Apodemus sylvaticus* (Chapter 4).

3.1.1 Overview of the sources of errors during Illumina library preparation and sequencing

Illumina sequencing generates millions of reads simultaneously in a short period of time, with a relatively high accuracy. Sequences are amplified in order to produce clonal clusters on the flow cell and bases are called one by one through massively parallel synthesis of the complementary strands. Only sequences containing Illumina adapters, which are complementary to the oligonucleotides present and immobilised on the flow cell, are sequenced, and the rest of the DNA is washed out.

One factor affecting sequence quality is overclustering on the Illumina flow cell. An excess on the number of clusters formed during sequencing can reduce the number of reads passing the Illumina chastity filter, that removes the least accurate clusters after the first 25 cycles through the measurement of a brigthness ratio. The chastity of a base call is the ratio of the intensity of the highest signal divided by sum of the two highest signals. Overclustering can also lower the Q30 quality scores, introduce sequencing artefacts to the reads and reduce the accuracy of base calling. One of the most common reasons of overclustering is an inaccurate library quantification (Illumina, 2018a), i.e. having too much DNA in the flow cell. In addition, a poor cleanup of the library can leave traces of adapters, adapter dimers or partial fragments of library constructs that will affect the clustering efficiency. Those sequences will be shorter than complete constructs and, therefore, will cluster more efficiently on the flow cell. Moreover, they can overinflate the DNA concentration leading to the underload of the flow cell.

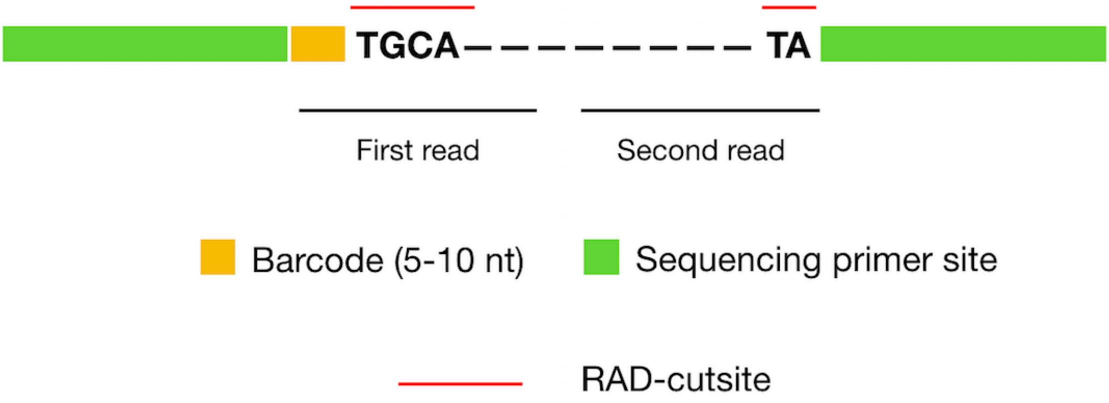


FIGURE 3.1: Representation of the structure of adapters used for the preparation of the library from Chapter 2

In projects requiring multiplexing of samples, specific barcodes are routinely added to the Illumina adapters. Generally, these barcodes will appear at the beginning of the read and will allow the demultiplexing or separation of the reads that belong to different individuals (Figure 3.1). An accurate knowledge of the barcode sequences is essential for a proper demultiplexing process. Reads that can not be assigned to any of the used barcodes are classified as undetermined. Despite its importance for the assessment of library preparation and sequencing quality, the percentage of undetermined reads is commonly neglected in scientific publications and only few scientific publications include them. For example, Chatterjee *et al.* (2012), using Illumina Hiseq 2000, found a 10.9% of undetermined reads in a reduced-representation bisulfite sequencing experiment (RRBS) and a 5.5% in a genomic DNA sequencing experiment used as a control. Bartram *et al.* (2016)

found a 10.9 % of undetermined reads in a MiSeq-based project while Meyer and Kircher (2010) indicated a 15% of undetermined reads when requiring perfect barcodes, which decreases to 5% when allowing one mismatch in the barcode sequence for the Illumina's Genome Analyzer II/IIx/IIe or HiSeq2000 libraries. These results, although limited, indicate an amount of unidentified reads per Illumina lane between 5% and 15%.

Another source of errors during Illumina sequencing is barcode hooping. Barcode hooping causes the incorrect assigment of barcodes to specific samples, and can result in the misassignment of reads from one sample to another (Illumina, 2018b). The best way to reduce barcode hooping is the removal of adaptors that did not ligate to DNA fragments (Illumina, 2018b). Furthermore, preservation of the libraries at -20°C and the use of dual indexing protocols can further reduce this error. Nevertheless, the percentage of reads that can be affected by barcode hooping is very low (from 0.2 to 2 %) (Illumina, 2018b) and, these sequences can be subsequently removed, during the data analysis. Such reads will have lower coverage than real reads, hence can potentially be removed due to their low coverage in *de novo* approaches or when mapping, when a reference genome is available.

Barcode design can also have important consequences for the demultiplex step. Barcodes with a short sequence distance between them limit the number of mismatches allowed during demultiplexing. As a consequence, if the quality of the base calls at the beginning of the sequence is low, or if there are uncalled bases in the barcode sequences, these sequences would be discarded or will contribute to barcode hooping. Designing barcodes with a sequence distance of, at least, 2- or 3-base difference between any pair of barcodes can help to reduce this problem (Illumina, 2018b).

Another problem, which is omitted in Illumina's trubleshooting guides, but that can significantly affect demultiplexing and the number of retained reads, is unspecific ligation of the adapters to digested DNA fragments. DNA ligases preferentially ligate sequences containing perfectly complementary sticky ends. However, some ligases, including T4 ligase, commonly used in library preparation protocols, can ligate overhangs with one or even more mismatches (Wu and Wallace, 1989; Cherepanov *et al.*, 2001). This leads to production of libraries where reads contain perfect barcodes but do not pass the demultiplexing step due to the absence of the correct RAD-cutsite produced by the enzymes used for the digestion.

Nevertheless, one of the advantages of RAD-seq is that it allows for recovery of sufficient number of high-quality reads to call thousands of SNPs, even if technical problems with the library preparation and/or sequencing results in large numbers of incorrect sequences. In the present work, after the first run of the Poland protocol (Poland and Rife, 2012), the proportion of retained reads after sequencing was very low, as more than the 80% of the generated reads had to be discarded. Nevertheless, the remaining 20% of the reads with correct features enabled the analysis described in Chapter 2. However, understanding the reasons behind this loss of sequences became an important and necessary step of the present work, in order to improve the library preparation protocol and to increase the proportion of retained reads for future sequencing experiments.

3.2 Analysis of discarded reads

The adapters used for the library preparation on Chapter 2 were different for the first and the second reads obtained through paired-end sequencing. Adapters ligated to the overhang produced by the enzyme SbfI, included different length barcodes (from 5 to 10 nt) in direct contact with the enzyme overhang, also known as RAD-cutsite, and a TruSeq adapter, complementary to the primers used for sequencing. The adapters ligated to the overhang produced by MseI, however, only contain the TruSeq adapter sequence complementary to the Illumina sequencing primers (Figure 3.1). Due to the lack of barcode on the second read, these reads were demultiplexed based on the cluster position in the flow cell during sequencing.

Four strategies were used to analyse the results from the demultiplexing process, requiring either a) perfect barcodes and cut sites, b) perfect barcodes and one mismatch on the cut site c) allowing one mismatch in the barcode and cut site and d) allowing two mismatches in the barcode and one mismatch on the cut site. When both perfect barcodes and RAD-cutsites were required (strategy a), only 18.05% of the total reads were retained (116619114 retained sequences out of 646058520 total reads)(Figure 3.2, strategy A). This indicates a very low efficiency in the library preparation and sequencing. Of the discarded reads, 3.35 % contained adapters (21671575 out of 646058520 reads) and were removed since the Stacks pipeline is not able to include reads of different lengths. 19.31% of the reads were discarded due to the lack of a perfect RAD-cutsite immediately following the barcode (124773433 out of 646058520 reads) and 59.27% of the reads were discarded due to the lack of a perfect barcode (382948066 out of 646058520 reads). There were no sequences marked by the Illumina chastity/purity filter as failing and the amount of reads discarded due to low quality was below 0.01% (46332 out of 646058520).

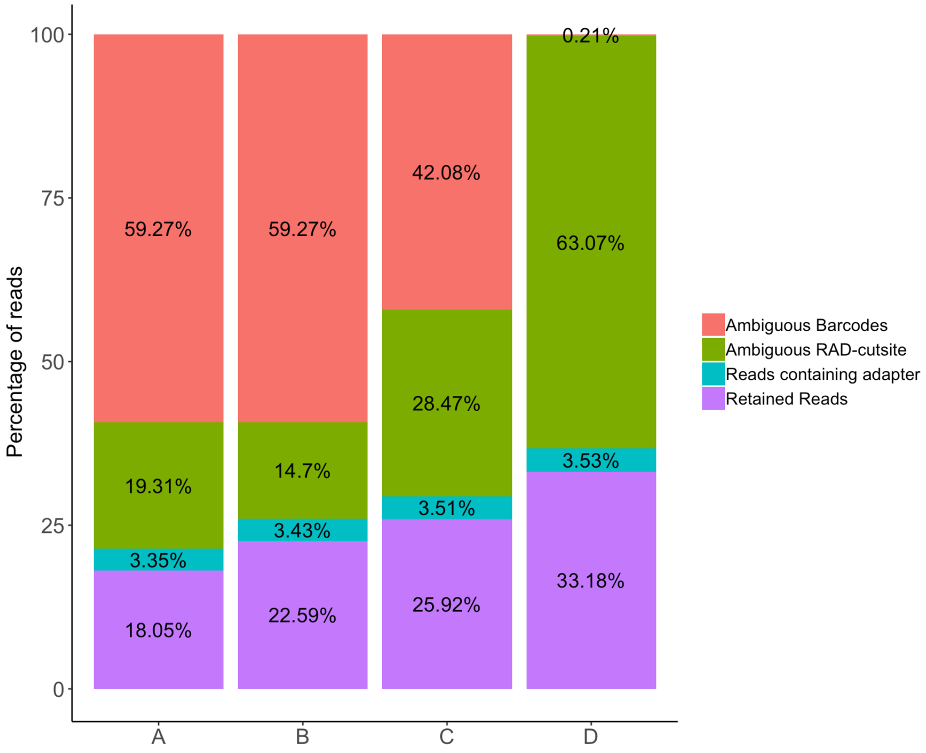


FIGURE 3.2: Classification of retained reads and discarded reads for the different analysis performed with process_radtags. A- Perfect barcodes and RAD-cutsites, B- One mismatch on RAD-cutsite and perfect barcode, C- one mismatch on RAD-cutsite and barcode, D- one mismatch on RAD-cutsite and 2 mistmaches on barcode

Allowing one mismatch on the RAD-cutsite (Figure 3.2, strategy B) increased the number of retained reads from 18.05% (116619114 out of 646058520 reads), to a 22.59% (145927734 reads). This 4.54 % (29308620 reads), comes from sequences that were previously discarded due to the absence of a perfect RAD-cutsites. Allowing one mismatch in the barcodes and one mismatch in the RAD-cutsites (Figure 3.2, strategy C), increased the number of retained reads to 25.92 % (167431226 reads). Even though the number of barcodes rescued with this setting is even larger, at 17.19 % (111065316 reads more), most of them did not contained an accepted RAD-cutsite. Finally, allowing two mismatches for barcode rescue reduced the amount of sequences with ambiguous barcodes to 0.21% (1325242 reads). Even after this relaxed filtering options, the number of retained reads remained quite low at 33.18% (214337206 reads) of the total amount of reads. Most of the discarded reads (63.07%, 407485572 reads) were discarded due to the lack of a correct RAD-cutsite (with one mismatch allowed) immediately following barcode sequence (Figure 3.2, strategy D) .

Analysing the amount of reads retained per barcode showed an unequal distribution of retained reads per sample (Figure 3.3). Although the samples were normalised to have equal amount of DNA of each one of them, these distributions are considered acceptable. Allowing one mismatch for barcode rescue, shows an important increament in the number of sequences recruited by some of the adapters, mainly the 5 nt ones. Allowing two mismatches for barcode rescue exacerbate these differences, with two 5 nucleotide adapters recruiting most of the reads

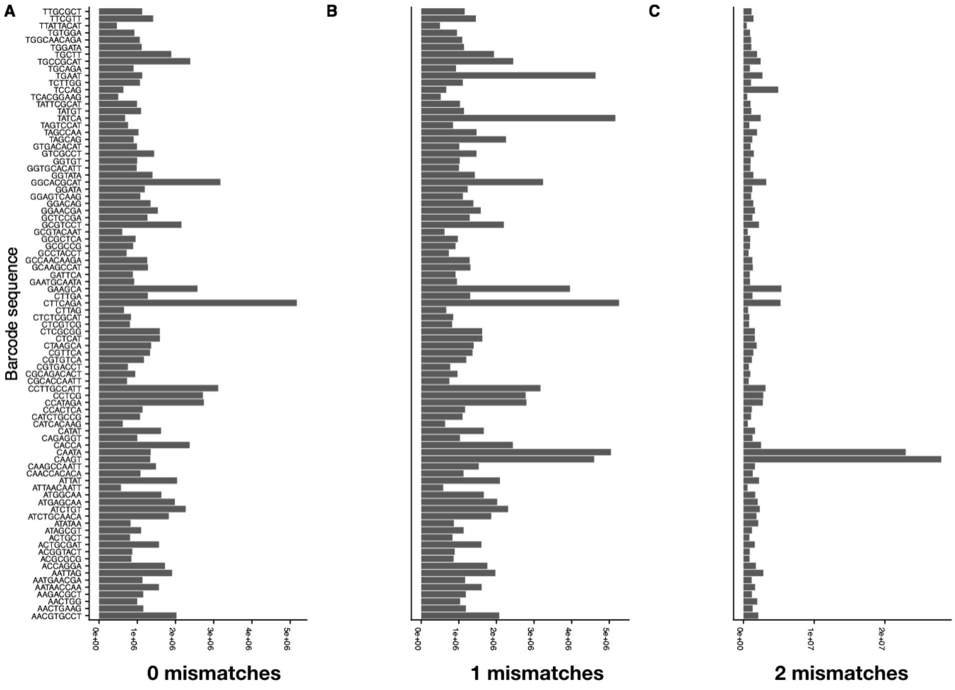


FIGURE 3.3: Sequences recruited per barcode allowing different number of mismatches

Allowing increased number of mismatches for barcode rescue, and also for RAD-cutsite rescue, increased the number of retained reads and also modified the distribution of discarded reads. This pattern is a clear indicator of a low sequence quality at the beginning of the reads, complicating the demultiplexing of the sequences. Most reads that were initially discarded due to ambiguous barcodes, at the end were discarded due to the ambiguous RAD-cutsites. The distribution of reads recruited per barcode clearly indicated a problem during demultiplexing: sequences that had longer barcodes were clasiffied as samples with a 5 nucleotide barcode after allowing two mismatches during barcode rescue. In these cases, after wrongly identifying a 5 nucleotide barcode, the RAD-cutsite could not be found immediately after, as more nucleotides were present from the longer barcode. Therefore, those sequences, apart from being wrongly identified, were excluded due to ambiguous RAD-cutsites.

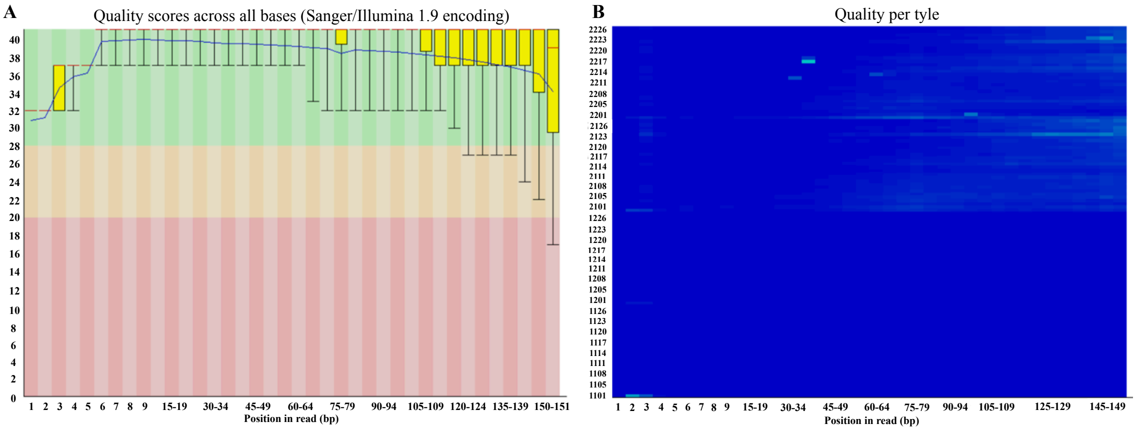


FIGURE 3.4: Per base sequence quality and per tile sequence quality for the second read. A: Box and whisker plot were the yellow boxes represent the 25th-75th percentiles and the whiskers represent the 10th and 90th percentiles. The red line represents the median value while the blue line indicated the mean quality. Background colours show very good quality calls, in green, reasonable good calls, in orange, and poor quality calls in red. B: Quality per tile plot shows the deviation from the average quality for each tile with colours in a cold to hot scale. Blue colours indicate positions were the quality is above average while hot colours indicate positions where the quality is below average.

The quality control of the fastq files showed a different quality pattern in the first and the second reads. The quality of the first read was generally good (Figure 3.4) and therefore poor quality of the reads themselves was likely not responsible for the problems with RAD-cutsites and barcodes. The quality of the second read, however, is very low in the first three nucleotides of the read (Figure 3.5) - exactly those that contain the RAD-cutsite. Fastqc reports showed a 2% of Ns in the first nucleotide position in the first read but over 20% of Ns in the third nucleotide of the second read.

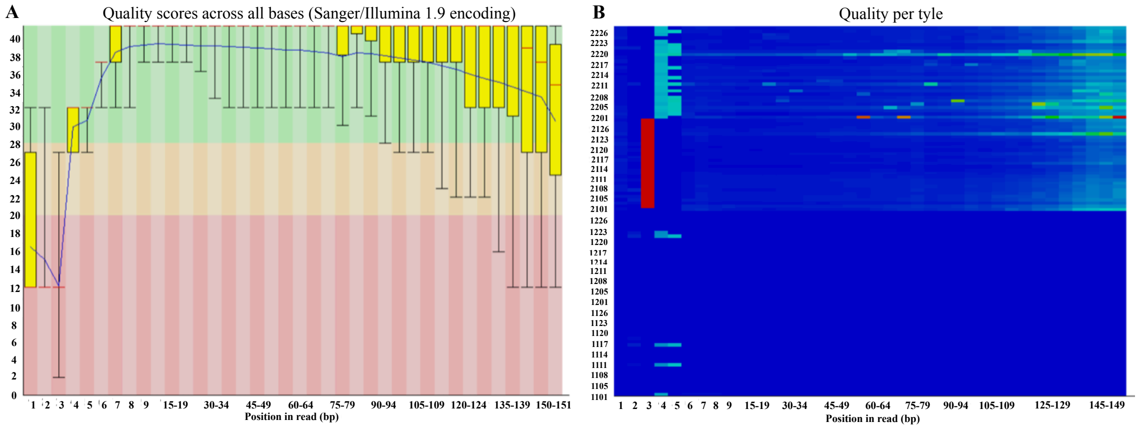


FIGURE 3.5: Per base sequence quality and per tile sequence quality for the second read. A: Box and whisker plot were the yellow boxes represent the 25th-75th percentiles and the whiskers represent the 10th and 90th percentiles. The red line represents the median value while the blue line indicated the mean quality. Background colours show very good quality calls, in green, reasonable good calls, in orange, and poor quality calls in red. B: Quality per tile plot shows the deviation from the average quality for each tile with colours in a cold to hot scale. Blue colours indicate positions were the quality is above average while hot colours indicate positions where the quality is below average.

3.2.1 Troubleshooting summary

Several issues have been identified which are likely responsible for the very low read retention in the first library.

- Low sequence quality at the beginning of the sequences was the major problem, as it affected all other elements of the sequences: RAD-cutsites and barcodes. It is likely that if sequences were of sufficient quality, there would be no issues with RAD-cutsite identification and sample assignment based on barcodes.

- Use of barcodes that are relatively short and of different lengths. Given the issues with sequence quality, a short barcodes can easily be lost; longer barcode can easily be misassigned as a shorted barcode, leading to misidentification of samples. Longer barcodes of equal length and appropriately spaced from one another would allow more effective reads rescue even with low sequence quality.

- The lack of barcodes ligated to the second reads made it impossible to recover the second reads when the barcode on the first read could not be recovered.

While uncovering the actual cause of these problem is not possible given this data, the most probable reasons behind the high number of sequences discarded (apart from unreliable barcode design) are the unspecific ligation of adapters to double digested DNA fragments and the degradation of the adapters due to a prolonged storage of the reconstituted adapters. Degraded adapters likely lost their entire or partial sticky ends and therefore were unable to reconstitute apropriate RAD-cutsites when correctly ligated; in the absence of high number of complementary sticky ends, T4 ligase could unspecifically join sequences together. It is now a standard recommendation in the RAD-seq protocols that the adapters should be reconstituted per single library preparation and stored for no longer than 2 weeks at -20°C (Franchini *et al.*, 2017). Our adapters were stored for much longer period of time, which likely contributed to their degradation.

3.3 New adapter design

In order to improve the performance of future sequencing libraries, we decided to adapt a newer and more robust protocol for library preparation. In particular, we wanted to have longer and same-length barcodes of substantial variation in their sequence. We also wanted the adapters to enable detection of PCR duplicate artefacts formed during the library preparation steps. As previously explained in Chapter 1, in standard ddRADseq protocols, PCR duplicates are indistinguishable and can be confounded with real reads, producing false genotype calls and increasing homozygosity (Pompanon *et al.*, 2005).

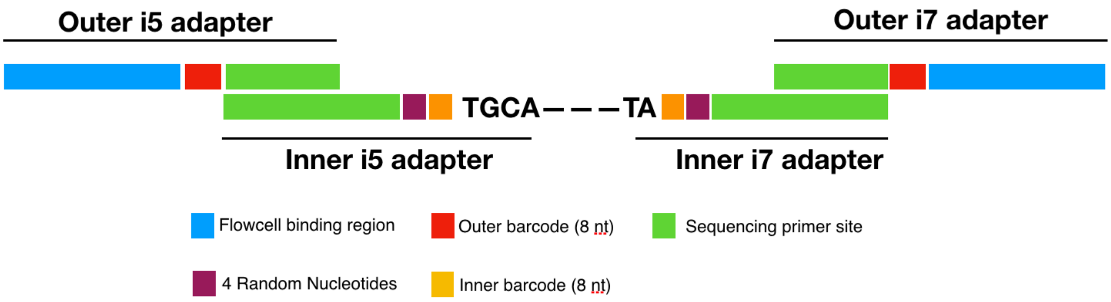


FIGURE 3.6: Representation of the structure of the modified quaddRAD adapters. Based on Franchini *et al.* (2017)

We found out that many of these features were present in a modification of a standard ddRAD-seq protocol by Franchini *et al.* (2017) called quaddRAD. quaddRAD features 4-part barcodes, two outer and two inner, allowing for cost-effective multiplexing of up to 196 samples on a single lane of Illumina sequencing. The quaddRAD adapters also include a sequence of 4 random nucleotides, which allow for detection of PCR duplicates arising during library preparation and sequencing (Figure 3.6).

The inner set of adapters designed by Franchini *et al.* (2017) include four inner i5 adapters and 3 inner i7 adapters. In the original protocol these adapters are combined to be able to multiplex a total of 12 samples using the same outer adapters. We have designed 12 inner i5 and 12 i7 inner adaptors that could allow us to multiplex 144 samples, based only on inner adapters. However, the reason behind this modification is to use fixed pairs of i5-i7 adapters, without combining them, in order to be able to identify chimeric sequences originated during PCR reaction or sequencing, avoiding confounding identification of reads. We decided to modify this protocol further, incorporating insights from our previous library preparation. We increased the length of the inner barcodes to 8 nucleotides and increased the distance between barcodes to 4 nucleotides, making it more likely to rescue them in cases of poor sequence quality at the beginning of the read. Finally, we changed the overhangs of the reconstituted adapters, making them compatible with our set of enzymes.

Barcodes were designed using EDITTAG (Faircloth and Glenn, 2012). 8nt tags were designed with a minimum distance between tags of 4 nt, with a GC content between 40 to 60%, avoiding sequences that were self complementary and that contained more than two adjacent, identical bases. From

TABLE 3.1: Barcodes used in the inner adapters.

Inner adapters	quaddRAD-i5	quaddRAD-i7top
1	AAGACTGG	AGAGTTCG
2	ATGTTGGC	ACCTGTTG
3	ATTGGCTG	AATCGCCT
4	CCTCATCT	CTGGTTCA
5	CGGAATTG	CGACAAGA
6	CAAGGTGA	CAGTCGAA
7	GACTTGAG	GTCAGAAC
8	GAATCACG	GGCAATCT
9	GGATTGTC	GTGGTCTT
10	TCCTTCAC	TTGTTCCG
11	TGTCAGTG	TCGCATTC
12	TTCTGAGG	TCGAACCA

102 tags suggested by EDITTAG, we manually remove sequences that could reconstruct the cut site used by our enzymes. 24 tags were selected for the inner adapters 8 more for the outer adapters. Furthermore the four random nucleotides were changed to 5′VBBN 3′, also to avoid reconstructing the cut sites used by our enzymes. The complete list of new adapters can be found in Appendix B, Section B.1.The inner and outer barcodes designed appear in table 3.1 and 3.2 respectively. Inner adapters have been used in fixed combinations while outer adapters have been used in all possible combinations.

TABLE 3.2: Barcodes used in the outer adapters.

i5 ID	i5 barcode	i7 ID	i7 barcode
i501	AGCATGGA	i701	ACACTCAG
i502	CCTGGAAT	i702	CAGTCGAA
i503	GCAAGCAA	i703	GGCTCAAT
i504	TGAGGATG	i704	TTCCGCTT

3.3.1 Testing of the new quaddRAD protocol and adapters

In order to test the newly designed adapters, 12 pairs of inner adapters and 16 combinations of outer adapters, a trial quaddRAD-seq experiment was run, using 16 samples of varied DNA quality as input for the library preparation (Figure 3.7).

Eight of those samples were of good quality, as indicated by a clear high molecular weight band on an agarose gel; six were of low quality, as they

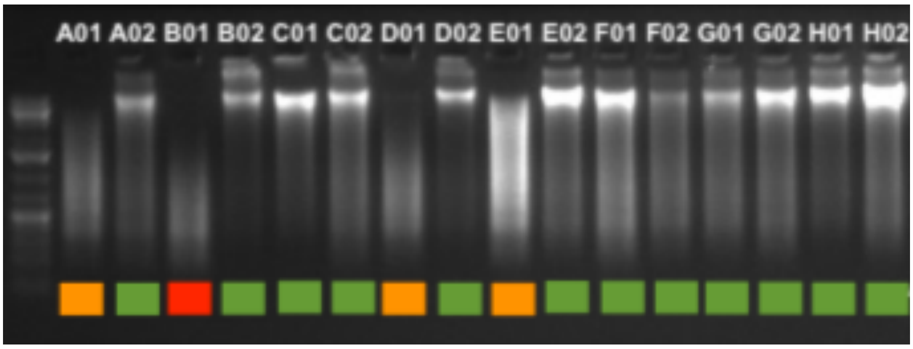


FIGURE 3.7: Example of quality assignment through agarose gel electrophoresis. Samples on this gel are an example and are not the samples selected for the trial run. Green represent HMW samples, orange degraded samples and red very degraded samples.

showed a smear of DNA on an agarose gel and two samples were highly degraded, as evidence by smeared band of low lengths on an agarose gel. Three poor quality samples were considered historical, obtained from preserved skins and stored in a museum collection of the Mammal Research Institute for at least a decade (Karol Zub, personal communication)(Table 3.3)

TABLE 3.3: List of samples used for the trial run including information about the age of the DNA, the quality of the DNA and the set of adapters used during library preparation

Sample	DNA type	quality	Inner	Outer i5	Outer i7
Sample 1	modern	degraded	1	501	701
Sample 2	modern	HMW	2	501	702
Sample 3	modern	HMW	3	501	703
Sample 4	modern	very degraded	4	501	704
Sample 5	modern	HMW	5	502	701
Sample 6	modern	HMW	6	502	702
Sample 7	modern	HMW	7	502	703
Sample 8	modern	HMW	8	502	704
Sample 9	modern	HMW	9	503	701
Sample 10	modern	degraded	10	503	702
Sample 11	modern	degraded	11	503	703
Sample 12	historical	degraded	12	503	704
Sample 13	historical	degraded	1	504	701
Sample 14	historical	degraded	2	504	702
Sample 15	modern	HMW	5	504	703
Sample 16	modern	very degraded	4	504	704

quaddRADseq library was prepared, in collaboration with Dr Marek Kucka and Dr Frank Yingguang Chan, from the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, following Franchini *et al.* (2017) protocol, using a different combination of enzymes (SbfI and MseI) and our own set of adapters. The full description of the library preparation is presented in Appendix B, Section B.2.

3.3.2 Performance of the modified quaddRAD protocol

In the trial quaddRAD library sequencing run, a total of 61056447 paired-end reads was obtained. The number of paired-end reads per individual varied from 2098866 to 5513146, with an average of 3816028 (median = 3682634, stdev =1162291). 2433121 paired-end reads (3.98%) were identified as PCR duplicates and removed from further analysis, which is consistent with previous reports that investigated this issue (Franchini *et al.*, 2017). 58623326 (96,02%) paired-end reads were used as input for demultiplexing. 415522 (0.70%) reads contained adapter sequences, 5129487 (8.74%) were low quality reads, (15.05%) contained ambiguous barcodes and (7.11%) contained ambiguous RAD-cutsites. After discarding those sequences, 49353572 reads were retained, an 80.83% of the reads obtained after the outer demultiplex took place. The results per sample are shown on Table 3.4

Critically, the newly developed adapters and protocol provided performance that not only substantially improved the data compared to our previous attempt, but that performance was in line with previously published data on RAD-seq performance.

More detailed analysis, presented in table 3.4, shows that the presence of adapters per sample varied between 0.22 % and 0.54 % of the input reads, while the number of low quality reads varied from 3.99% to 6.98%. The largest differences between samples were found in the number of ambiguous barcodes, which varied from 2.75% to 19.05% and in the number of ambiguous RAD-cutsites, which varied from 0.8% to 15.01%.

TABLE 3.4: Percentage of retained and discarded reads for each one of the samples included on the trial run

Sample	Paired end reads	% clones	% Retained Reads	% Adapters	% Low Quality	% Ambiguous Barcodes	% Ambiguous RAD-Tag
1	2454996	7.51	75.15	0.49	4.28	10.53	2.04
2	3101859	3.25	83.48	0.33	4.09	6.59	2.26
3	5414329	4.18	86.49	0.34	3.92	4.28	0.80
4	2499536	5.41	61.14	0.26	6.61	18.56	8.02
5	4785820	3.56	88.40	0.39	3.98	2.48	1.19
6	4641668	3.15	88.37	0.37	4.00	2.74	1.36
7	3018111	2.82	83.35	0.34	4.16	5.52	3.82
8	4472463	3.37	79.18	0.31	3.86	10.12	3.15
9	4710475	3.18	86.70	0.34	4.04	4.41	1.33
10	2098866	2.25	65.83	0.27	4.36	15.62	11.68
11	5513146	4.40	86.13	0.33	4.18	3.35	1.60
12	5204083	3.94	80.90	0.30	4.21	7.16	3.49
13	2339829	5.61	78.13	0.52	3.83	9.20	2.72
14	3750238	4.79	81.11	0.33	4.02	6.84	2.91
15	3615030	3.05	85.70	0.36	4.02	4.67	2.19
16	3435998	4.90	56.07	0.22	4.76	19.05	15.01

3.4 Discussion

The analysis of the discarded reads from the first library preparation using the Poland and Rife (2012) protocol and adapters provided by our collaborators showed low sequence quality at the beginning of the reads, compounded by barcode design that is not robust to sequence ambiguity. The most likely explanation for these observations is degradation of adapters and unspecific ligation of adapters to digested DNA.

Given this experience, new adapters were designed based on the quaddRAD approach of (Franchini *et al.*, 2017). The new design has four-part barcodes on both paired-end reads, random nucleotides to detect and remove PCR duplicates and increase the accuracy of genotype calling. The use of outer and inner barcodes allows us to multiplex a high number of samples without the cost of adapter synthesis. The increased length of the inner barcodes to 8 nucleotides and a large distance of 4 nucleotides between any pair of barcodes makes it more likely to rescue them, even in cases of poor sequence quality at the beginning of the read. A fixed combination of inner barcodes was used, instead of combinations of them, to allow the identification of chimeric sequences. Furthermore, in order to keep the integrity of the adapters, they were reconstituted and used within two weeks and stored at -20°C.

The new adapters have clearly improved the efficiency of the sequencing run, increasing the proportion of retained reads from less than 20% to an 80% on the trial quaddRAd run. The exact percentage of reads retained for this trial run is impossible to calculate, due to the use of a shared lane with another project and to the number of reads filtered by the sequencing facility.

3.4.1 Performance of the quaddRAD protocol for genotyping of degraded samples

Traditionally, RAD-seq has been performed only using high molecular weight DNA in order to produce a concordant set of loci for all the individuals in an experiment. Degraded samples typically contain shorter DNA fragments that are less likely to contain both restriction sites, which is required in double digestion protocols. The number of fragments containing both cutting sites produced by the selected enzymes, and therefore ligating to the adapters used for sequencing, will be lower than in samples with high molecular weigh DNA, leading to a lower number of reads. As one of the elements of testing of the new protocol was to assess its suitability for genotyping samples with degraded DNA, during the trial run performed to test the efficiency of the newly developed adapters, I sequenced eight good quality samples, characterised by a high molecular weight DNA band on an agarose gel, six degraded samples, which showed a smear of DNA on an agarose gel and two highly degraded samples, with smear of short lengths of DNA fragments (Figure 3.7).

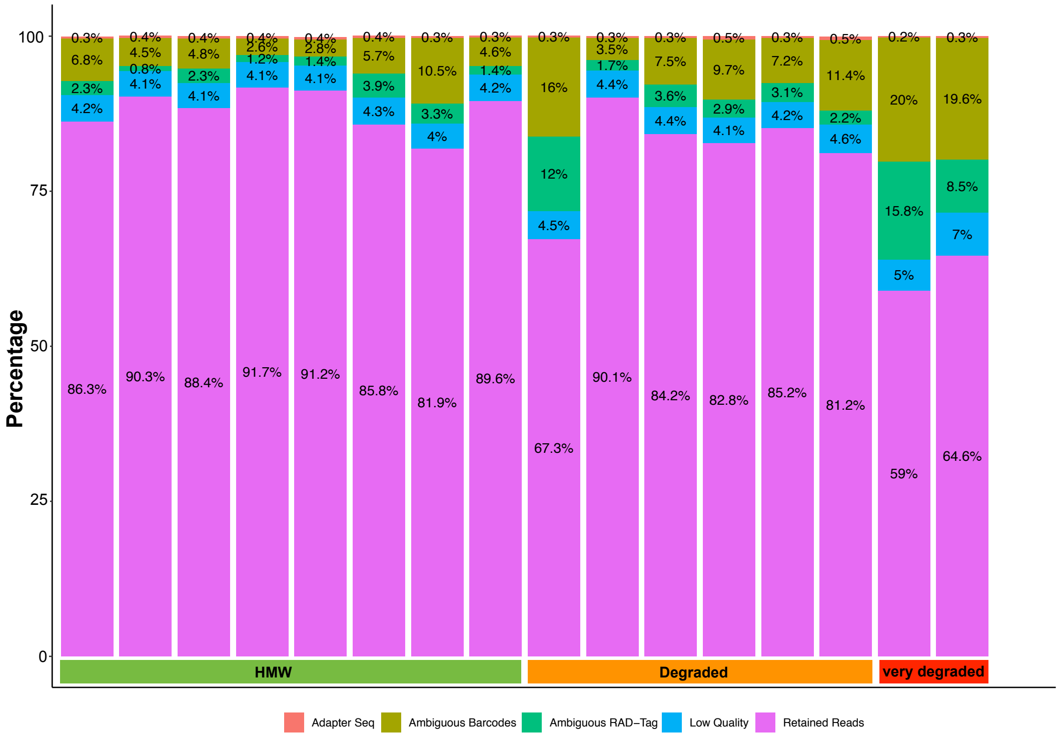


FIGURE 3.8: Results for the QuaddRAD test library: Percentage of adapter sequences, ambiguous barcodes, ambiguous RAD-Cutsites, low Quality reads and retained reads for the samples sequenced on the quaddRAD test library. Colour line on the bottom of the barplot indicate the quality of the samples: green,good samples with high molecular weight DNA, orange, samples with degraded DNA and red, samples with very degraded DNA

The amount of retained reads obtained from process_radtags differed between samples presenting HMW DNA and very degraded samples (Figure 3.8). Samples considered as degraded presented a lot of variation on the amount of retained reads, with Sample 10 exhibiting a similar pattern to the one observed on very degraded samples and Sample 11 showing a similar pattern to samples with high molecular weight. Most of the reads discarded in very degraded samples were removed due to the presence of ambiguous barcodes or ambiguos RAD-cutsites. Despite these differences, all the samples retained a large fraction of their reads, with a minimum of 59% of retained reads.

Although the observed coverage differs between samples (Figure 3.9), these differences do not seem to be related to the quality of the input DNA. The coverage values ranged from 13.35x for a very degraded sample to 36.89x for a high-quality sample. However, degraded samples has, on average, a higher coverage than samples with HMW DNA. The number of assembled loci, polymorphic loci and SNPs were analysed using m=3, M=2 and n=2. The results show a higher number of assembled loci, polymorphic loci and SNPs on HMW samples, with an average of 92472 assembled loci on HMW samples, 71994 on degraded samples and 66597 on very degraded samples (Figure 3.10).

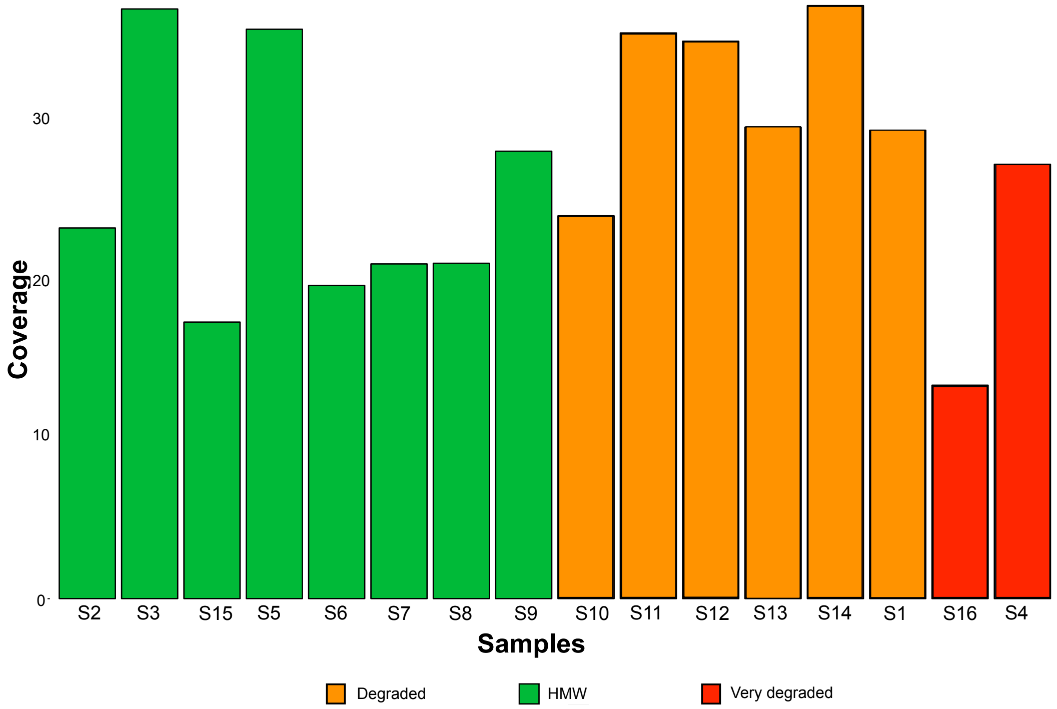


FIGURE 3.9: Final coverage for m=3 for the samples tested on the trial run of the quaddRAD protocol

This test demonstrated that the modified quaddRAD method is suitable for samples with varied levels of DNA quality, despite potentially lower chances of fragments having two restriction enzymes recognition sites.

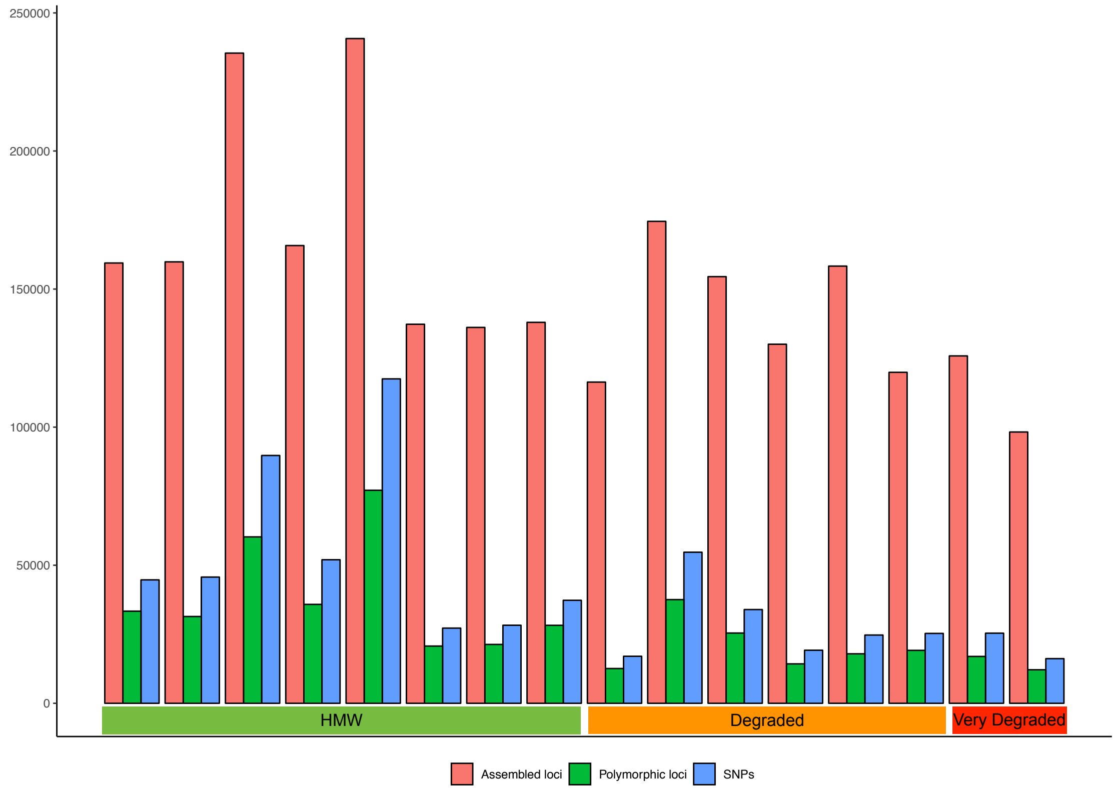


FIGURE 3.10: Number of assembled loci, polymorphic loci and SNPs, for m=3, M=2 and n=0

Thousand of loci can be sequenced in degraded samples and can be used in combination with better preserved samples. Problems can arise when working only with highly degraded samples as they will produce lower number of assembled loci, and therefore, getting the same set of loci in multiple samples can became problematic. In order to be able to successfully work with highly degraded samples, adaptations of ddRADseq protocols should be considered, such as the hyRAD method (Suchan *et al.*, 2016), which uses biotinylated RAD fragments as baits for capturing homologous fragments for sequencing.

Skin DNA extractions not only contained highly degraded DNA, but also very low concentrations of DNA (0 ng/μl to 57.62 ng/μ, Average=8.52, Stdev=15.26). A longer hydration of the samples before DNA extraction could have improved the amount of DNA extracted from each sample and could have increase the lenght of the obtained fragments ((Moraes-Barros and Morgante, 2007)). However, the concentration of the samples should not have affected the sequencing performance, as equal amounts of DNA were used per sample. The important excess on the number of reads recruited by two samples does not seem to be related with the combinations of adapters used for sequencing, as it occurred in two samples using different adapters. The most probably explanation of this phenomena could be a wrong quantification of the samples, causing the input of a higher amount of DNA into the final library.

3.5 Conclusions

The design of new adapters, based on the quaddRAD protocol developed by (Franchini *et al.*, 2017), has considerably improved the percentage of retained reads, from a 20% to an 80%. These results are now comparable to previously published library preparation performance data (Bartram *et al.*, 2016; Chatterjee *et al.*, 2012; Meyer and Kircher, 2010). Due to this positive result, the library preparation for the major project on European phylogeography of *Apodemus* was performed using the above described protocol. Furthermore, sequencing of degraded samples produced positive results, retaining, in the worst case, more than 56% of the reads, allowing us to include degraded samples on the library preparation for the main phylogeography project.

Chapter 4

European phylogeography of *Apodemus flavicollis* and *Apodemus sylvaticus: Material and Methods*

4.1 Introduction

The phylogeographic history of *Apodemus* species in Europe has long been studied (Michaux *et al.*, 2003; Michaux *et al.*, 2004; Michaux *et al.*, 2005; Herman *et al.*, 2017). However, despite the ecological importance of the genus, there has been a lack of genome-wide resources available for their study. Majority of what is known about their phylogeography is based on a single mitochondrial gene, CytB. As previously described on Chapter 1, this marker has proven to be reliable and informative, leading to identification of the major refugia used by *A. flavicollis* and *A. sylvaticus* to survive the Pleistocene glaciations.

Initially, the studies were focused on the Mediterranean peninsulas as they were a well known refugia for many of the temperate species that nowadays occupy the European continent (Michaux *et al.*, 2003; Michaux *et al.*, 2004; Michaux *et al.*, 2005). The detection of more northern refugia in other temperate species of mammals (Bilton *et al.*, 1998; Kotlík *et al.*, 2006; Wójcik *et al.*, 2010; Jaarola and Searle, 2002), mainly in the Carphatians, encouraged more focused studies of *A. sylvaticus* phylogeography, contributing to the postulation of a northern refugium for the species (Herman *et al.*, 2017). However, analyses performed with CytB did not have the power to identify the possible location of this northern group (Herman *et al.*, 2017).

Whether the northern group is unique to *A. sylvaticus* or also appear in *A. flavicollis*, with its own geographic origin, is also unknown. Even when *A.*

flavicollis and *A. sylvaticus* have shown different strategies to survive to the Pleistocene glaciations, due to their similar ecological habitats, it is possible that both species could have survived in more northern regions.

The use of a higher number of markers spread across the genome will provide a greater resolution than the traditional markers used on *Apodemus*. Having thousands of independent loci, subject to different evolutionary pressures will increase the power to detect population structure and phylogeographic patterns. It will also help us to understand the processes responsible of the observed genomic patterns.

In this chapter, I describe the application of the quaddRAD protocol developed in Chapter 3 to analyse the European phylogeography of the two *Apodemus* species.

The main objectives of this project are:

- To determine the number and the relationship between genetically differentiated groups in Europe and analyse their genetic diversity.

- To determine the existence and location of other refugia apart from the ones in the Mediterranean peninsulas.

- To analyse the potential role of humans in the dispersal of the two species.

- To identify geographical regions with particularly highly differentiated populations in order to guide our sampling for future sequencing projects.

- To test the accuracy of the SNPs loci catalogue developed previously to correctly assign species identification to the samples.

4.2 Material and methods

4.2.1 Sample collection and DNA extraction

Samples from *A. flavicollis* and *A. sylvaticus* were provided thanks to collaborations with several colleagues across Europe: Dr Johan Michaux (University of Liège, Liège, Belgium), Dr Jerry Herman (National Museums of Scotland, Edinburgh, Scotland), Dr Joana Paupério (CIBIO-InBIO, University of

Porto, Porto, Portugal), Dr Vladimir Jovanović (Institute for Biological Research "Siniša Stanković", Belgrade, Serbia), Dr Douglas J. Clarke (University of Huddersfield, Huddersfield, United Kingdom) and Dr Karol Zub (Mammal Research Institute, Białowieża, Poland).

In total, 576 samples from different regions in Europe were provided to us (Figure 4.1). 348 samples belonged to *A.sylvaticus* and 286 to *A. flavicollis*. 19 were not identified at the species level and I assigned their species designation using the catalogue of loci developed on samples from the Polish population, described in Chapter 2.

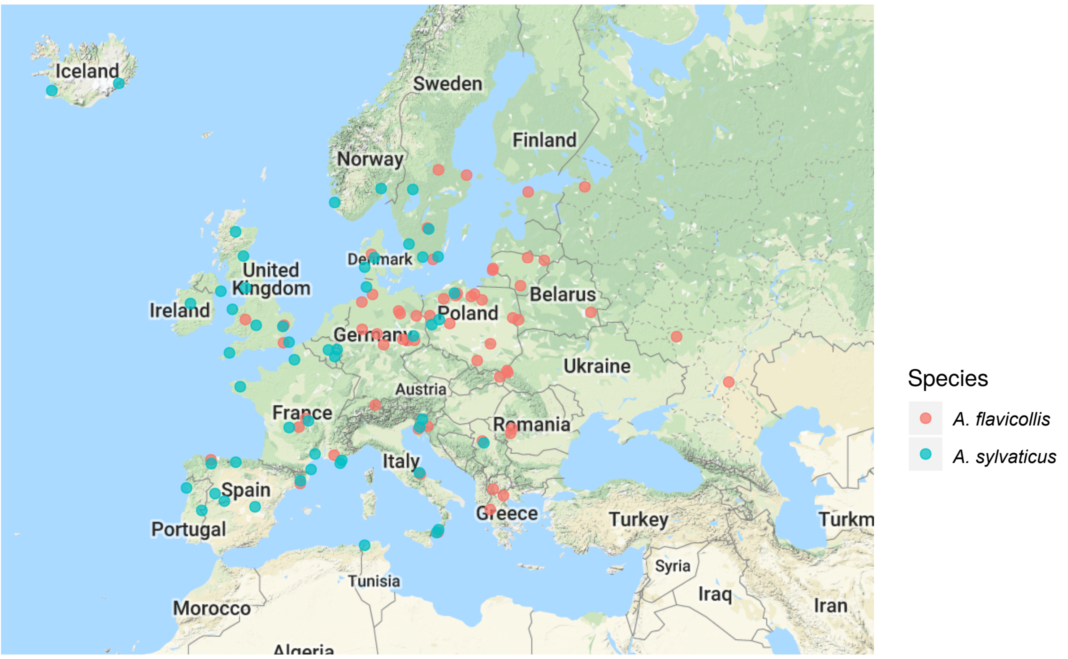


FIGURE 4.1: Location of the samples for the European phylogeography project

Most of the tissue samples were of tail, feet, ear and kidney tissues, preserved in tubes with 96% ethanol. Once the samples arrived at our lab, they were stored at -20°C until DNA extractions. In addition, 34 dried skin from a museum collection at the Mammal Research Institute, Białowieża, Poland were included to check the efficiency of RAD-seq protocols on degraded DNA. Skins were kept in dry and dark conditions at room temperature until DNA extraction. The DNA extractions were performed following the protocol described in Chapter 2. Once the DNA quality was assessed, 352 samples were selected for the library preparation. Those samples were normalised to a final concentration of $6ng/\mu l$ in 10 μl.

4.2.2 Library preparation

quaddRAD-seq library was prepared following the protocol described in Appendix B, Section B.2. and tested on Chapter 3. The three pairs of adapters tested performing digestion and ligation in two different steps were excluded for the library preparation (Inner adapters 3, 8 and 9). Digestion of DNA and adapter ligation was perfomed in a single step. Libraries were prepared in four 96-well plates, each one including 86 samples. 8 individuals, in total, were included in duplicate for quality control, 2 per plate, and 1 individual was sequenced in triplicate (DK1-DK7, ES2-ES4, FR10-FR15, DE19-DE35, FR16-FR32, SC8-SC1, EN11-EN22, PT6-PT7 and ES12-ES15-ES27).

Samples from the same plate were multiplexed by adding the same amount of DNA from each sample. Libraries 1 and 4 were multiplexed at 10ng of DNA per sample, while libraries 2 and 3 at 20 ng per sample due to variable amount of DNA in different sample preparations. The multiplexing scheme can be found in Appendix C, Section C.1. Fragments between 300-600bp were selected with BluePippin (Sage Science) followed by size and concentration check with 2100 Bioanalyzer (Agilent Technologies) (Figure 4.2). Libriaries were sequenced with paired-end protocols on Illumina Hiseq3000 (Illumina) at the Genome Center of the Max Planck institute for Developmental Biology in Tübingen, Germany.

4.2.3 Processing of RAD-tags

Reads were demultiplexed based on outer adapters at the Genome Center of the Max Planck institute for Developmental Biology in Tübingen. This demultiplex will separate groups from 6 to 9 samples sharing the same combination of outer barcodes, but different inner barcodes (Figure 4.3).

A total of 10 sets of 2 files (first and second reads) were received per library. Reads were analysed using Stacks version 1.48 (Catchen *et al.*, 2011). PCR duplicates were removed using clone_filter program. Reads were demultiplexed and quality filtered using process_radtags program. Reads containing adapter sequences, uncalled bases, low quality scores or that were marked by Illumina's chastity/purity filter as failing were discarded. Barcode rescue was enabled, requiring maximum 2 mismatches in the barcode sequence and sequences were truncated to a final length of 136 bp.

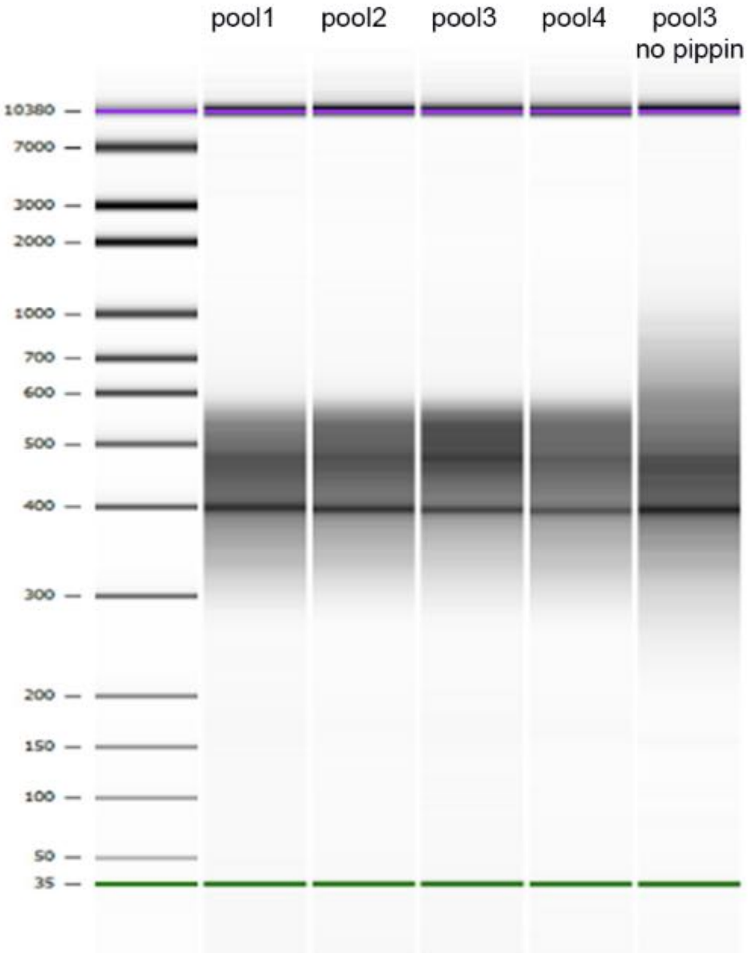


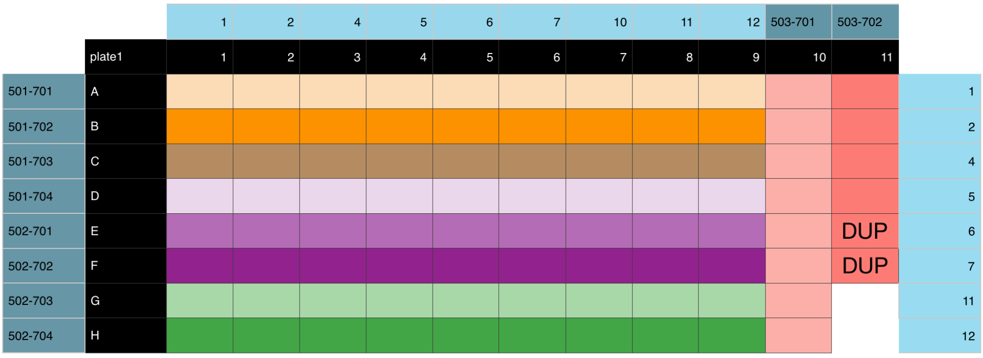
FIGURE 4.2: Comparison of size selection step for the four libraries. Pool 1 to 4 represent the four libraries sequenced, after the size selection step. Pool 3 no Pippin show the distribution of lengths obtained on library 3, after DNA digestion and ligation of inner and outer adapters, but before size selection with Bluepippin. Longer fragments, up to 1000bp, are present in Pool 3 no Pippin and have being removed during size selection, being absents in Pool 3.

.

Chimeric sequences produced during sequencing were extracted using process_radtags, considering pairs of tags used in different samples.

4.2.4 Species identification: undetermined samples

Serbian samples, which species identification was lost, were matched against the catalogue built in Chapter 2. Principal Component Analysis (PCA) was run using the Adegenet package from R (Jombart, 2008) and plotted with ggplot2 (Wickham, 2016) to assign species to each sample.



		1	2	4	5	6	7	10	11	12	503-701	503-702	
	plate1	1	2	3	4	5	6	7	8	9	10	11	
501-701	A												1
501-702	B												2
501-703	C												4
501-704	D												5
502-701	E											DUP	6
502-702	F											DUP	7
502-703	G												11
502-704	H												12

Inner barcodes
Outer barcodes
DUP Duplicate sample

FIGURE 4.3: Example of the multiplex scheme followed combining outer and inner adapters for Plate 1. The colour scheme represents the groups demultiplexed at the Genome Center of the Max Planck Institute in Tübingen

4.2.5 Selection of parameters and variant calling

Best parameters for each species were calculated separately, following Paris *et al.* (2017) approach with a custom bash script. We called SNPs using the populations program from Stacks (Catchen *et al.*, 2011), considering all the samples as belonging to the same population. We kept SNPs common to the 50% of the individuals. Afterwards, SNPs were filtered using VCFtools (Danecek *et al.*, 2011). First, individuals with more than 50% of missing data were removed from further analysis and a new population map was generated. SNPs with MAF smaller than 0.05 and which deviated from the HWE at P <0.05 were also excluded. Subsequently, sites with a mean depth value smaller than 20 were excluded.

4.2.6 Analysis of genetic diversity

Individual ancestries were estimated following a maximum likelihood approach with ADMIXTURE (Alexanderet al., 2009) and a Bayesian approach with fastStructure (Raj *et al.*, 2014). Ten replicates were run with different seeds, and only the highest likelihood, or lower cross-validation error replicate, for each value of K were represented. PCA analyses were performed using the R package Adegenet (Jombart, 2008). Discriminant Analysis of Principal Components (DAPC) was performed using the groups defined by find.clusters function, both also from Adegenet. Fst and heterozygosity were calculated with the populations package from Stacks (Catchen *et al.*, 2011).

Maximum likelihood phylogenetic trees were built using SNPhylo (Lee *et al.*, 2014), a pipeline specialised in building trees from big SNPs datasets. Confidence values were estimated with 1000 bootstrap replicates and the root was fixed on the separation between the outgroup and the ingroup. *A. flavicollis* has being used as outgroup to root *A. sylvaticus* tree while *A. sylvaticus* has being used as outgroup to root *A. flavicollis* tree.

4.2.7 Inference on population history

DIYABC (Cornuet *et al.*, 2014), an Approximate Bayesian Computation software for inference on population history using molecular markers was used to calculate the most likely scenario to explain the presence of Iberian genotypes in Sweden. Four different scenarios have been checked (Figure 4.4). Scenarios 1 and 3 represent the first split between the northern population and the Iberian-Swedish group, with a second split between the Swedish population and the Iberian group. Scenarios 2 and 4, however, show an early separation between Iberia and the northern-Swedish group, followed by a later split between the northern group and the Swedish population. Scenarios 1 and 2 will consider that the time of the first split was at least 17 Ka with the second split taken place around 8 Ka, through Doggerland. Scenarios 3 and 4 considered a more recent time for the second split of around 2 Ka. In the latter case that movement would be consistent with human-related migration. 20000 simulations of the data were performed with DIYABC (Cornuet *et al.*, 2014). 10000 trees were simulated using abcrf (Approximate Bayesian Computation via Random Forests) package in R (Pudlo *et al.*, 2015) and the probability of each DIYABC scenario was calculated using a custom script written by Dr Ilaria Coscia and Dr Allan McDevitt from the University of Salford.

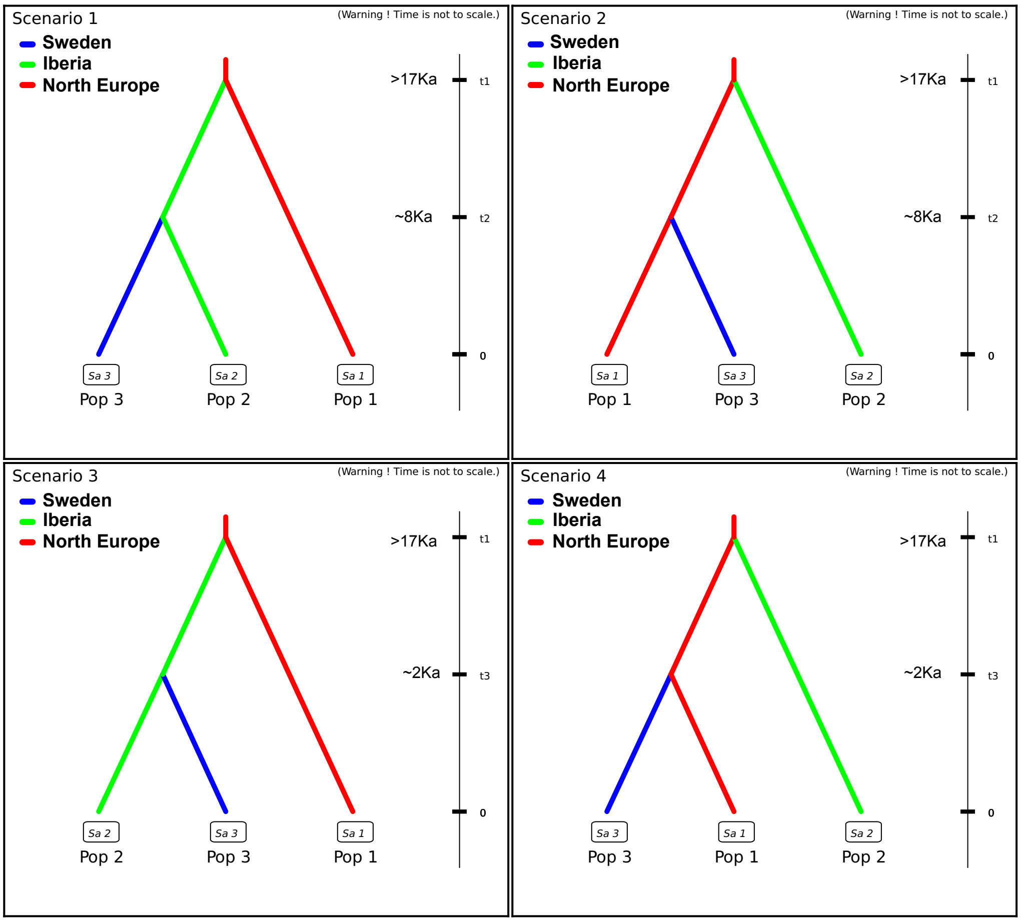


FIGURE 4.4: Models tested with DIYABC to determine the origin of a Swedish population

4.3 Results

4.3.1 Sequencing and data cleaning

1931952960 (1.9B) paired-end reads were generated during the sequencing of the quaddRAD libraries, 48298824 per sequencing lane. 555045356 (37.93%) paired-end reads had an unknown origin after allowing 1 mismatch during barcode rescue. The percentage of unknown reads per lane varied from 37.92% to 38.73% (median= 38.09, stdev= 0.93). In total, 1199225303 (1.1B) paired end reads were identified as belonging to the quaddrad project (62.07% of the sequencing output). The number of paired-end reads per sequencing lane was similar between the 4 lanes, varying from 295900713 to 305305939 (average=299806325, median=299009325, stdev=4526499). 32792436 paired-end reads (2.74%) were identified as PCR duplicates and removed from further analysis. A total of 24571573 chimeric sequences were identified, representing a 1.02% of the sequences generated during sequencing.

1166326553 (97.26% of the total number of reads) paired-end reads (or 2332653106 single end reads), were used as input for demultiplexing and cleaning. 8254446 single-end reads contained adapter sequences, 74620425 were low quality reads, 182081786 contained ambiguous barcodes and 192926555 contained ambiguous RAD-cutsites. After discarding those sequences, 1874769894 single-end reads were retained, an 78.16% of the known reads and a 65.46% of the total number of reads produced during sequencing.

The presence of adapters per sample varied between 0.31 % and 0.42 % of the input reads while the amount of low quality reads varied from 2.90% to 3.84%. The biggest differences between samples were found in the number of ambiguous barcodes, which varied from 5.07% to 11.27% and in the number of ambiguous RADtags, which varied from 6.14% to 9.95%, but this time the differences were smaller than in the trial library preparation (Chapter 3). A table with multiplexing details can be found on Appendix D, Section D.1.

The number of reads per individual varied from 61815879 to 20585 reads, with an average of 5459747 reads (median=5101172, stdev=4503961). These results are heavily influenced by two samples that contained more than 60000000 reads. The total number of reads is higher, on average, for fresh samples, preserved on ethanol, than for dry skins from museum collections (Figure 4.5). Details about the number of reads per individual can be found in Appendix D, Section D.1.

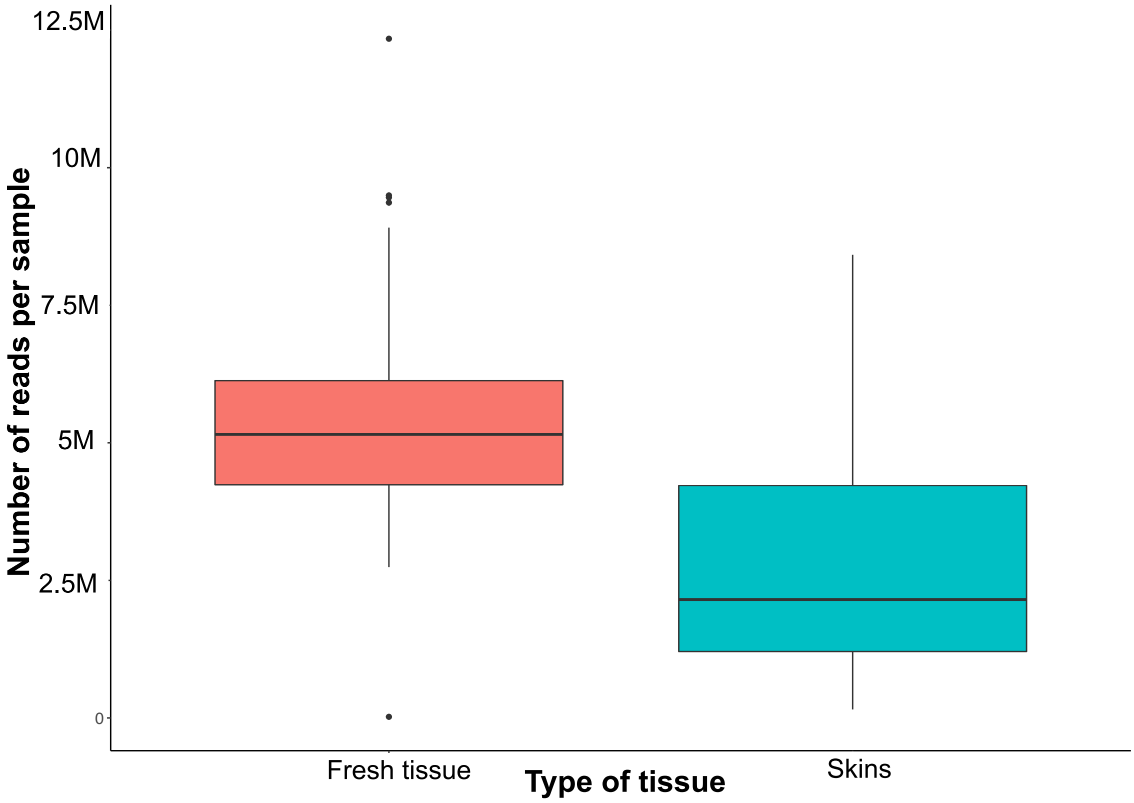


FIGURE 4.5: Distribution of the number of reads per individual based on the type of tissue used for DNA extraction. The line dividing the boxes, or middle quartile, represent the median value of each kind of tissue. Boxes represent the inter-quartile range and include the middle 50% of the values. Upper wishkers represent the top 25 % of the number of reads per samples, while the lower whiskers indicates the 25% of lowest scores. Individuals dots are outliers.

4.3.2 Species identification: undetermined samples

Species assignment for the undetermined samples was performed with Principal Components Analysis. The first two axes of the Principal Components Analysis (Figure 4.6) explained together a 14.28 % of the total variance. The first principal component allowed the differentiation between both species, while the second axis explained the differentiation of the populations of *A. flavicollis*. Serbian samples from *A. flavicollis* appeared closer to the Polish samples from the same species, while Serbian samples from *A. sylvaticus* appeared away from the rest of *A. sylvaticus* samples. Due to the characteristics of the catalogue, that is clearly bias towards *A. flavicollis*, and the fact that we knew in advance the number of samples from each species, the distance between the cluster formed by *Apodemus flavicollis* and the rest of the samples allowed us a clear differentiation of both species, even when *A. sylvaticus* samples from Serbia appear in a middle position to both Polish species.

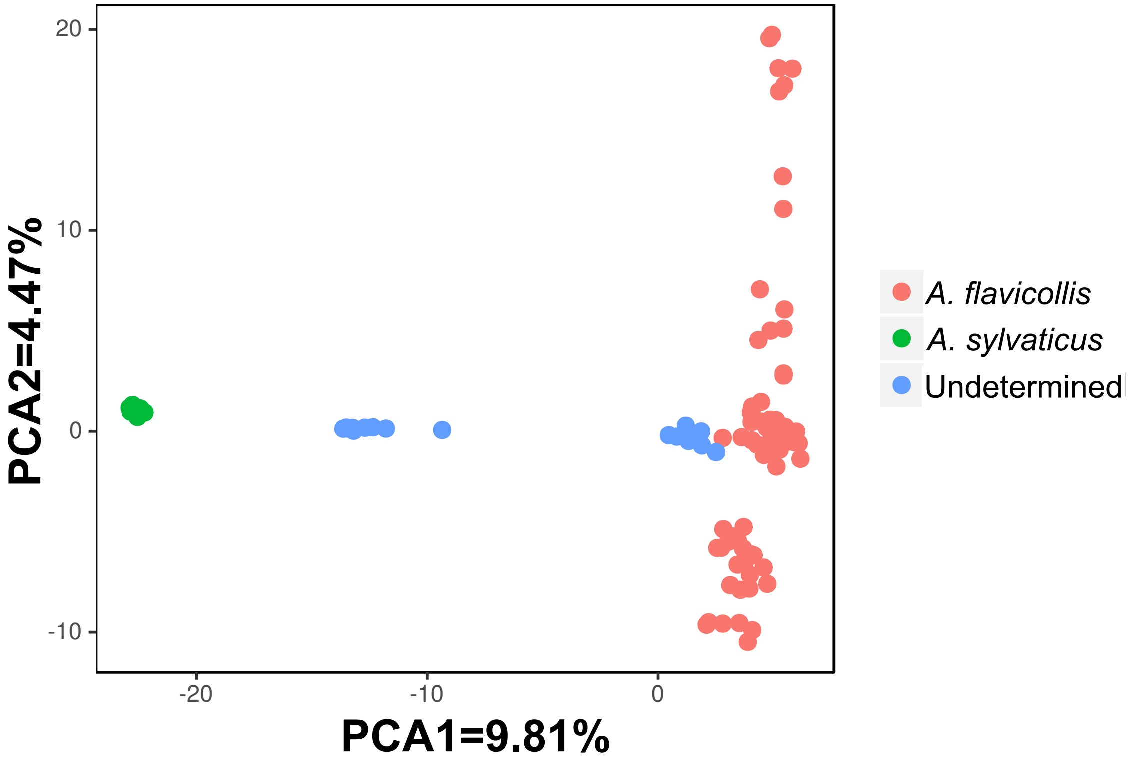


FIGURE 4.6: Species differentiation for undetermined Serbian samples trough Principal Components Analysis. Red dots represent samples from *Apodemus flavicollis* populations in Poland, green dots are *Apodemus sylvaticus* samples from Poland and blue dots are the undetermined samples from Serbia.

This pattern of species differentiation was also evident on the phylogenetic tree (Figure 4.7). Cluster formed contained each one of the species were well supported, with a 100% bootstrap support. Internal relationships within each species had lower support, with *A. sylvaticus* internal relationship being less supported than in *A. flavicollis*. Based on both approaches, samples RS1, RS2, RS4, RS5, RS8, RS10, RS11 and RS15 were identified as *A. flavicollis* and samples RS3, RS6, RS7, RS9, RS12, RS13, RS14, RS16, RS17 and RS18 were identified as *A. sylvaticus*.

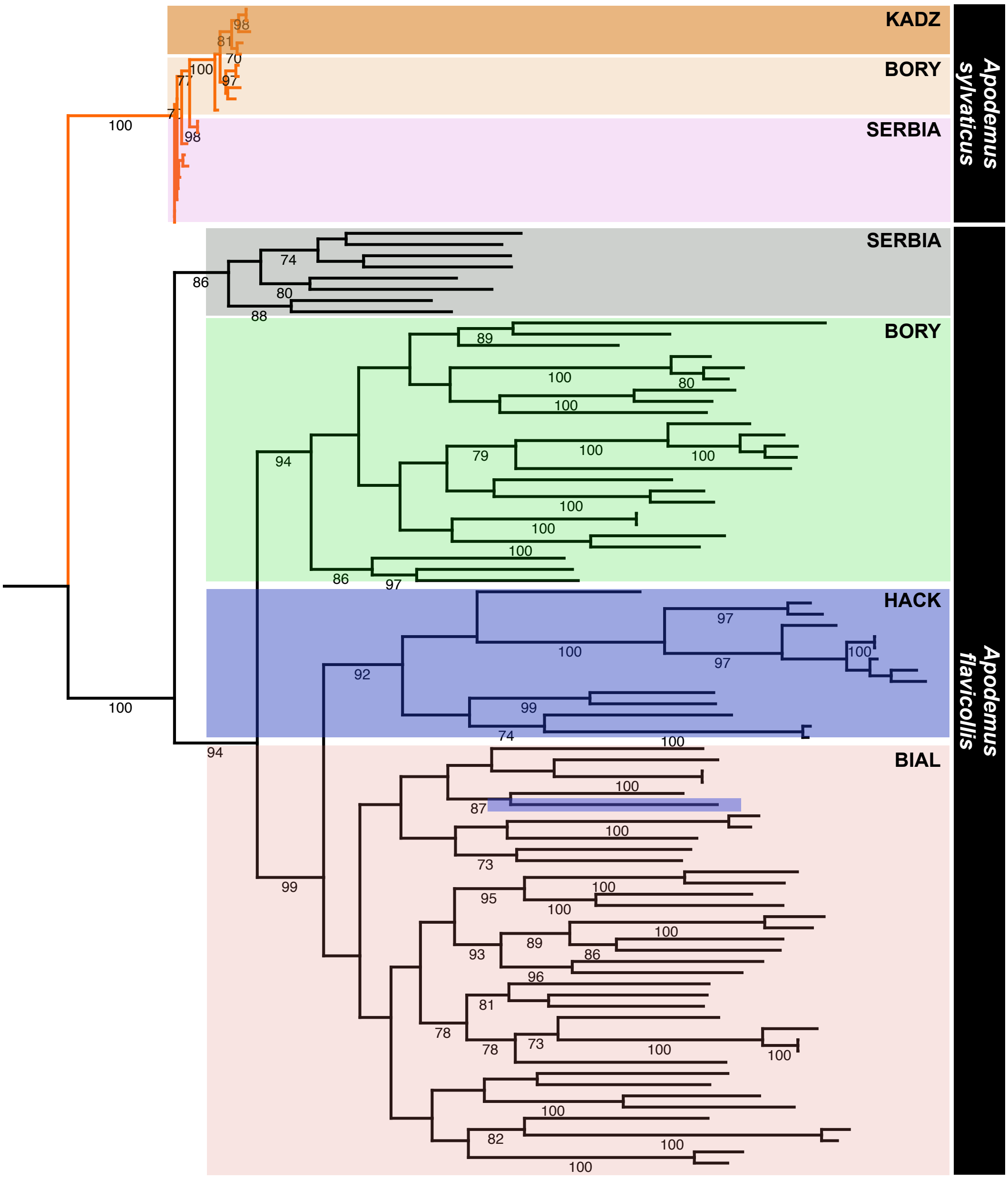


FIGURE 4.7: Species identification of Serbian samples trough maximum likelihood phylogenetic tree built with SNPhylo (Lee *et al.*, 2014). Orange lane: *Apodemus sylvaticus* ; black lane: *Apodemus flavicollis*. Coloured boxes specify the different populations identified for each species. *Apodemus flavicollis* populations: pale read, Białowieża, blue, Haćki, green, Bory Tucholskie and grey, Serbia. *Apodemus sylvaticus* populations: orange, Kadzidło, yellow, Bory Tucholskie and light pink, Serbia. Bootstrap values higher than 70% are shown on the bottom of the branches.

4.3.3 *Apodemus sylvaticus*

4.3.3.1 Selection of parameters and variant calling

The m parameter controls the minimum number of identical, raw reads required to create a stack, than can be then considered as a putative allele. Increasing values of m increases the coverage per sample, which in turn increases even more after merging the putative alleles into loci (Figure 4.8)(Data available in Appendix D, Section D.2.1).

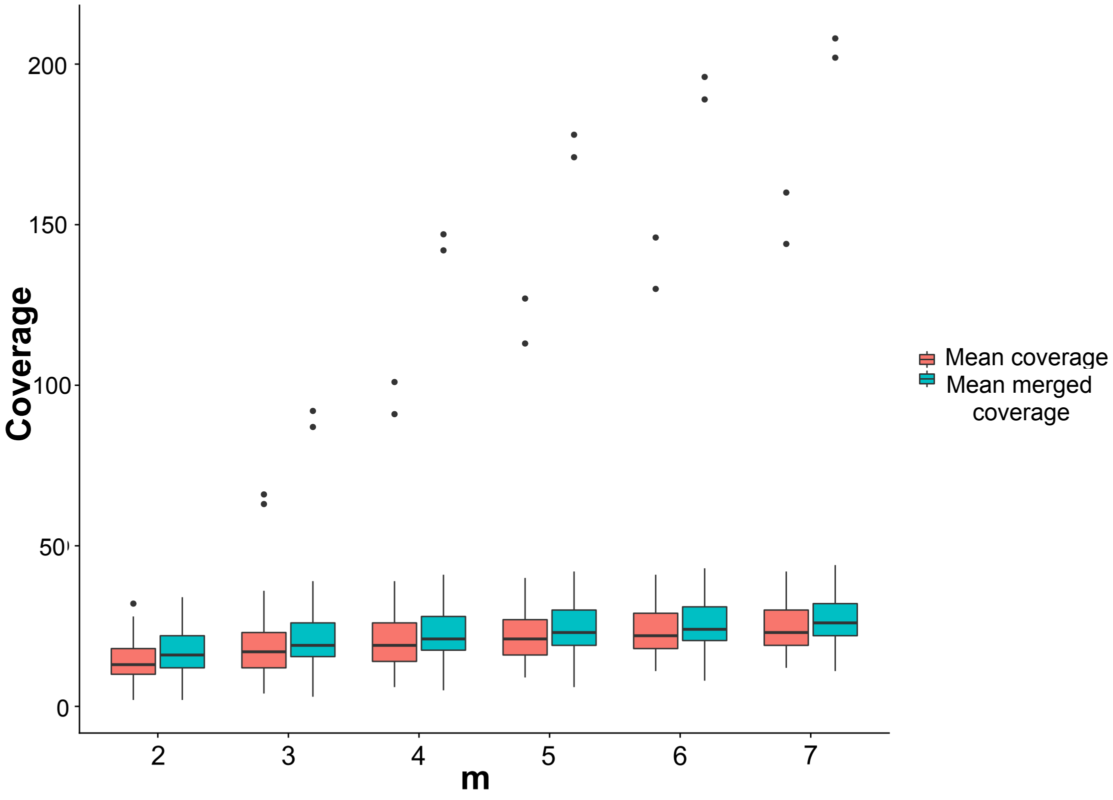


FIGURE 4.8: Distribution of mean coverage for each iteration of the m parameter for *A.sylvaticus*. Mean coverage, in red, is the average value obtained for each sample using only primary reads while mean merged coverage, in blue, is the average coverage value after merging alleles into loci. The boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range. Black dots are outliers outside the 10th and 90th percentile.

However it also reduces the number of assembled loci and SNPs, and, from m=3, the number of polymorphic loci. A value of three was selected in order to increase the number of polymorphic loci generated.

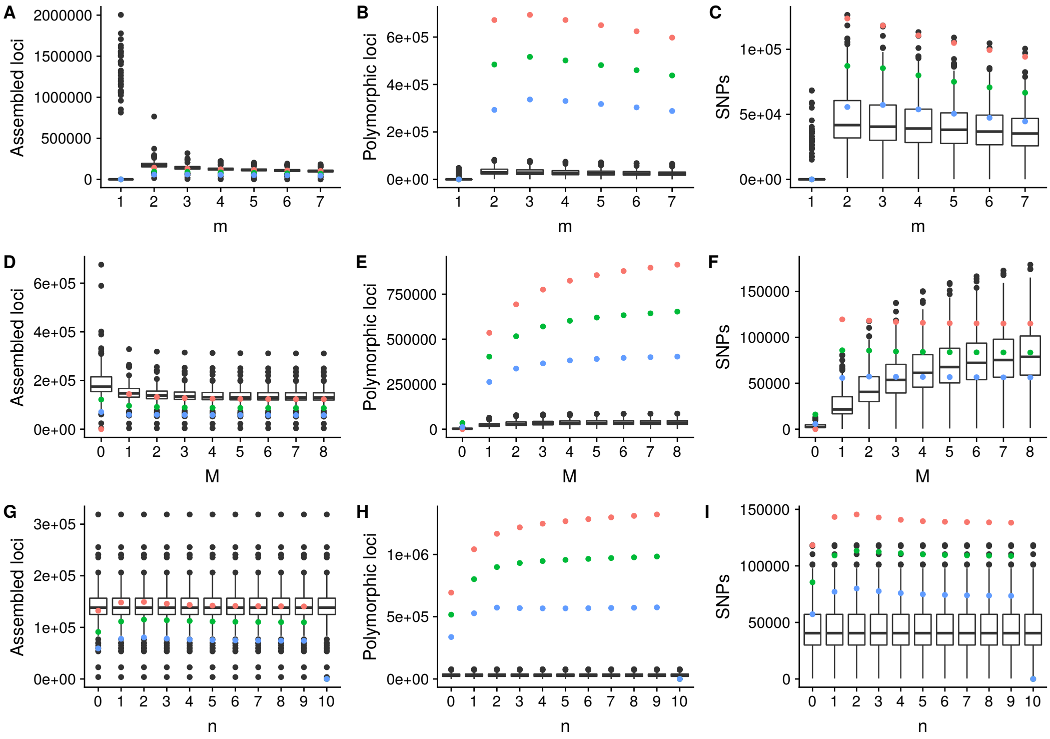


FIGURE 4.9: Variation in the number of Assembled loci, polymorphic loci and number of SNPs for each iteration of m, M and n parameters in *Apodemus sylvaticus*. Blue circles represent data found in at least 40% of the samples, green circles, in at least 60% and red circles in at least 80% of the sample. The boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range. Black dots are outliers outside the 10th and 90th percentile.

The M parameter controls the number of differences allowed between putative alleles to consider them as a locus. Increasing values of M increase the number of polymorphic loci and decrease the number of assembled loci. Even when the number of SNPs per individual increases with increasing values of M, the number of SNPs shared across the 80% of the population decreases with each iteration of M. The n parameter controls the number of differences allowed between samples to consider loci from different samples to be homologous. Increasing values of n produce an increment, at the population level, in the number of polymorphic loci. The number of SNPs increase up to n=2 and then start decreasing slowly. Due to the continuous increment in the number of polymorphic loci for both parameters and the reduction of the increment after M=4, values of 4 were selected for both parameters (Figure 4.9)(Data available in Appendix D, Section D.2.2).

A total of 981452 SNPs were called in 195 individuals. 12 individuals were removed due to missing data. Out of 981452 SNPs called, 760381 (77.47%) were removed after filtering for minor allele frequencies and further 177584 (18.09%) were removed after failing the Hardy-Weinberg equilibrium test. 8285 sites (0.84%) were removed due to mean depth values lower than the threshold of 20, leaving 35199 (3.58%) SNPs to be used in the downstream analyses. Another individual was removed at this stage due to erroneous genotypes, leaving a total of 182 individuals for further analysis (list of individuals available in Appendix D, Section D.2.3). The distribution of samples by population is shown in Figure 4.10. In total, we analysed samples from 18 different countries: Belgium (BE), Germany (DE), Denmark (DK), England (EN), Spain (ES), France (FR), Ireland (IE), Iceland (IS), Italy (IT), Norwey (NO), Poland (PL), Portugal (PT), Serbia (RS), Scotland (SC), Sweden (SE), Slovenia (Sl), Tunisia (TN) and Wales (WL).

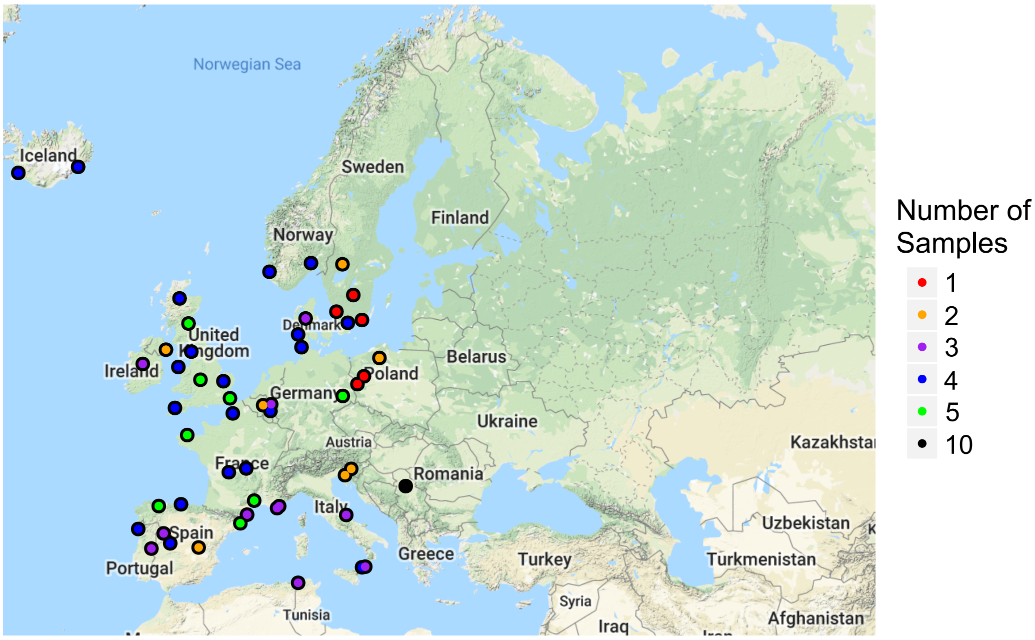


FIGURE 4.10: Distribution of the number of samples per population for *Apodemus sylvaticus* european samples

4.3.3.2 Phylogeographic history of *Apodemus sylvaticus*

The first two axes of the principal component analysis (Figure 4.11) explained cumulatively 12.95% of the variance. The two components allow for the differentiation of two groups of samples, one including samples from Serbia, Italy and Slovenia and the second one including the rest of the samples from

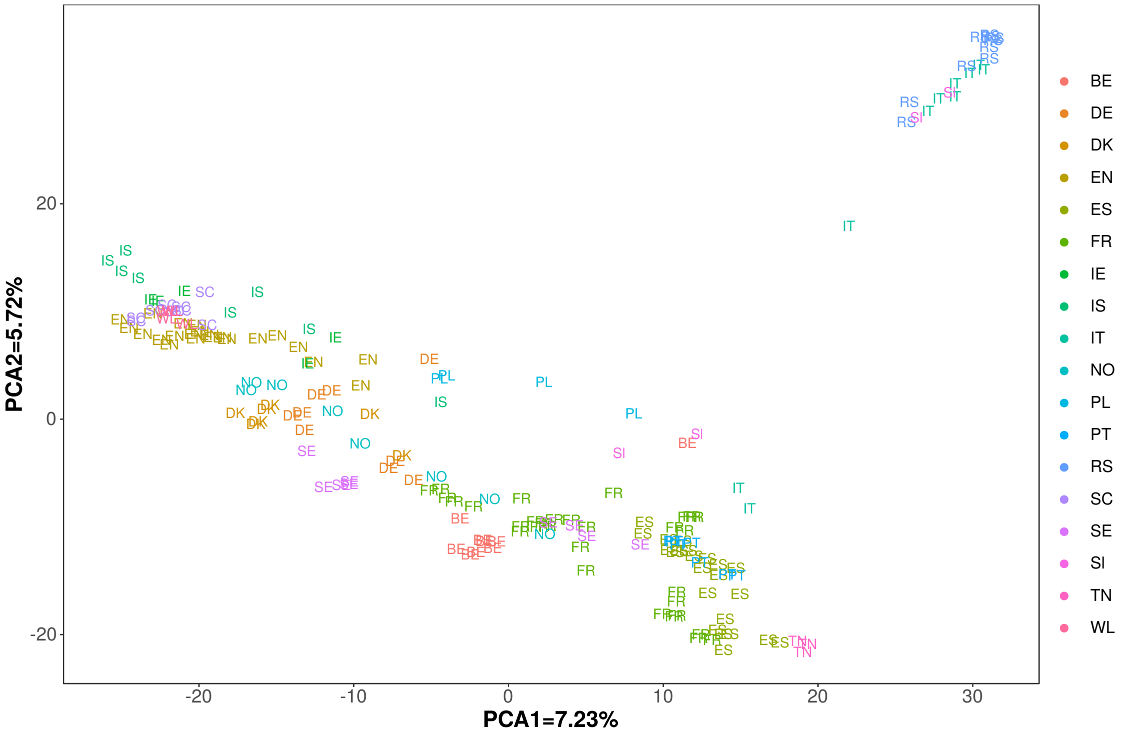


FIGURE 4.11: Principal Component Analysis for *Apodemus sylvaticus* samples from Europe. The different countries are identify with different colours and also with different characters: Belgium (BE), Germany (DE), Denmark (DK), England (EN), Spain (ES), France (FR), Ireland (IE), Iceland (IS), Italy (IT), Norwey (NO), Poland (PL), Portugal (PT), Serbia (RS), Scotland (SC), Sweden (SE), Slovenia (Sl), Tunisia (TN) and Wales (WL).

Europe. The shape of the groups on the PCA plot reflects the geography of the continent, from North Africa at the bottom right northwards (Figure 4.11). In the top left side of this group, individuals from the British Isles and Iceland appear together, also following the geography of the territory. Some individuals from Slovenia, Belgium, Italy and Poland appear at a distance from the main group of samples, towards the Italo-Balkan group.

Plotting of the other PCA axes did not reveal any other groupings of the samples (data not shown). Removal of the samples from Serbia, Italy and Slovenia increased the differentiation between continental samples and those from the British Isles and Iceland (Figure 4.12).

In general, most of the samples group together with samples from the same country and neighbouring regions, with the exception of 4 samples from Sweden that group together with samples from the Iberian peninsula and France, and 4 samples from Norway which group with French and

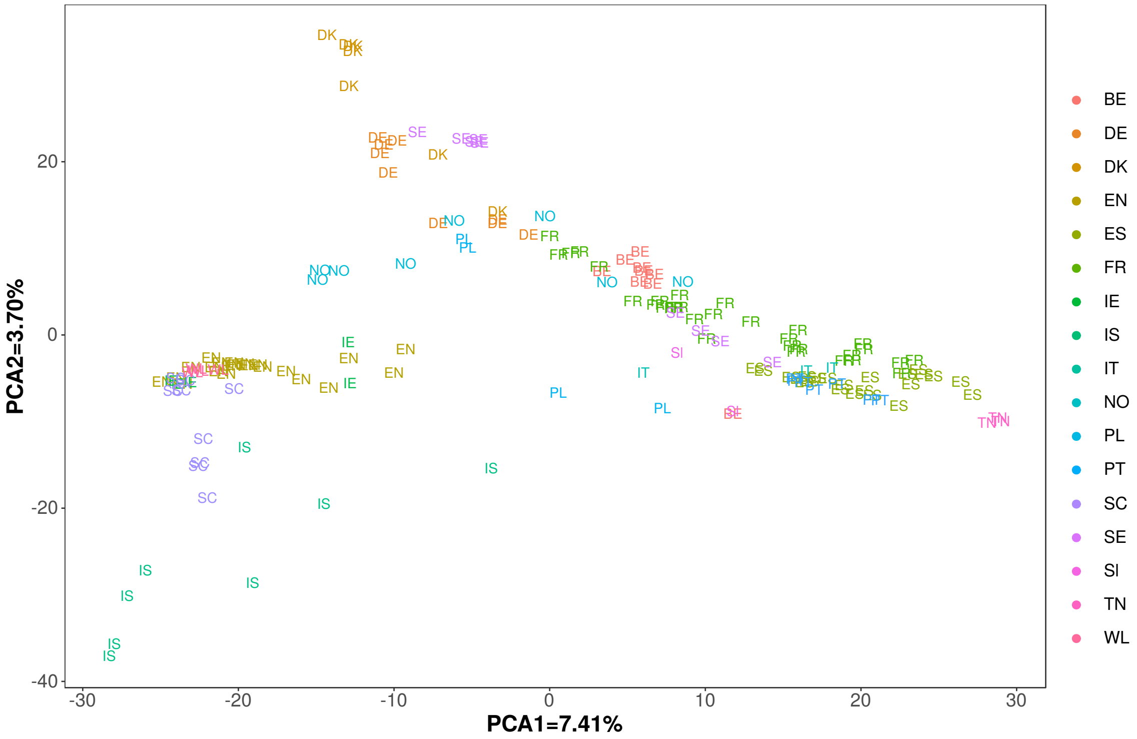


FIGURE 4.12: Principal component Analysis for the main group of *Apodemus sylvaticus* samples from Europe. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norwey, PL-Poland, PT-Portugal, SC-Scotland, SE-Sweden,TN-Tunisia and WL- Wales.

Belgian individuals.

Find.cluster function from Adegenet (Jombart, 2008) identified four different groups among the samples (Figure 4.13). The red group includes samples from Italy, Slovenia and Serbia, the same that appear on the top right side of Figure 4.11. The main group identified by PCA has been divided in three groups: the orange group includes all the samples from the British Isles and Iceland, the blue one the samples from northern Europe and the green one the samples from Tunisia, Iberia and central Europe.

Discriminant analysis of principal components performed with the previously defined groups show a closer relationship between Iberian-southern European and northern European groups than between the other groups. Despite the typical isolation in the island environments, the group from the British Isles and Iceland is closer to the southern and northern European populations than the group from the Italo-Balkan peninsulas.

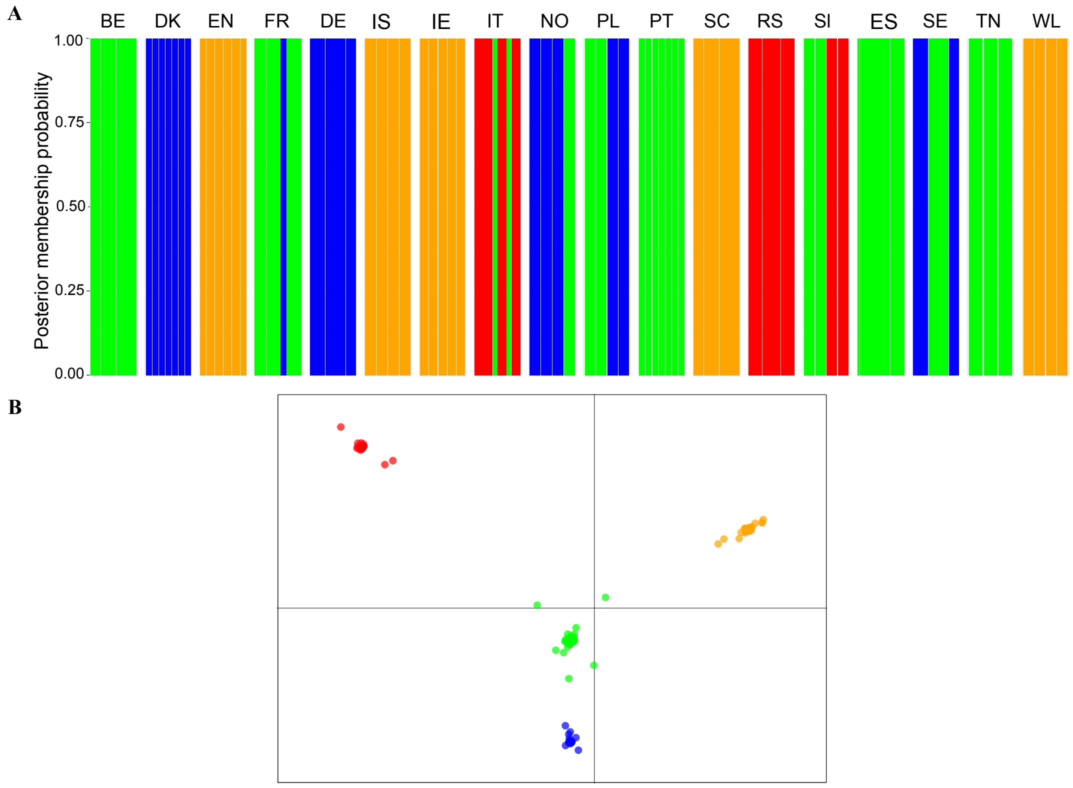


FIGURE 4.13: Compoplot and DAPC for *Apodemus sylvaticus* european samples. A: Compoplot or barplot which represents the group assignment probability of each individual to the 4 inferred groups. Each bar represents one individual and each colour represents one of the 4 inferred groups. B: Discriminant analysis of principal components scatterplot. Each dot represents one sample while the colour indicates the group to which each sample belongs to. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norwey, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

4.3.3.3 Admixture analysis

Ancestry analysis, such as ADMIXTURE or STRUCTURE, considers the genome of each individual as a mix of genomes originated from multiple hypothetical ancestral populations (Khrunin *et al.*, 2013), whose number (K) has to be specified *a priori*. Different approaches (Maximum likelihood in ADIXTURE and Bayesian Inference in Fast-STRUCTURE) have been used to parsimoniously explain the variation between the individuals included in this analysis. The methods used to estimate the optimal K for our dataset were not completely consistent, showing different results depending on the algorithm used, but congruent between them. Fast-STRUCTURE indicated that the optimal K is between K4 (model complexity that maximised marginal likelihood) and K7 (model components used to explain structure in data) in 80% of the runs performed, whereas the 20% of the runs indicated a K value between K4 (model complexity that maximizes marginal likelihood) and K6 (model components used to explain structure in data).

The optimal value of K in ADMIXTURE was estimated through the cross-validation error values of 10 repetitions of each run with different seed (Figure 4.14). The lowest cross-validation errors were found at K=5 and K=6.

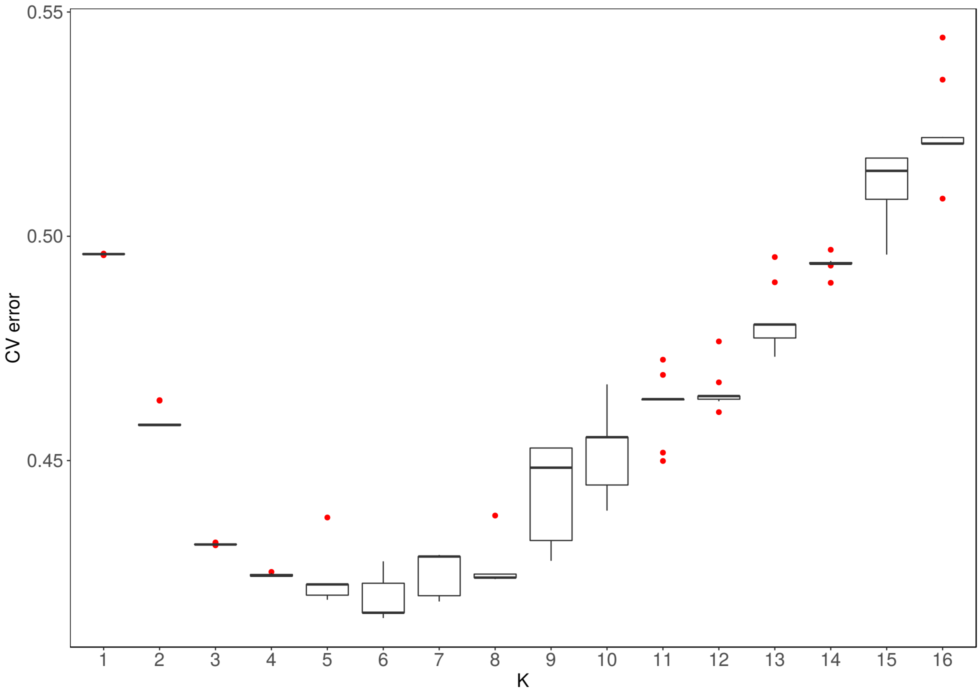


FIGURE 4.14: Distribution of cross-validation errors for 10 runs of 16 K values each one for *A. sylvaticus* dataset. Box and whisker plot were the boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range or the distance between the two hinges of the box. Red dots are outliers.

Below, ADMIXTURE results from K=2 to k=7 are presented together with the biological interpretation given to each one of the Ks. The supposition that there is a real or true value of K it is always wrong (Lawson *et al.*, 2018), therefore all the most plausible models will be explored. Through this extensive analysis we would try to identify the best approximation to explain the structure of our data. It is also imporant to notice that ADMIXTURE/STRUCTURE analysis are sensitive to different sample sizes. Groups with small sample sizes or groups that have undergone low population genetic drift can appear as a mixed of other groups rather to assign to their own ancestral group (Lawson *et al.*, 2018).

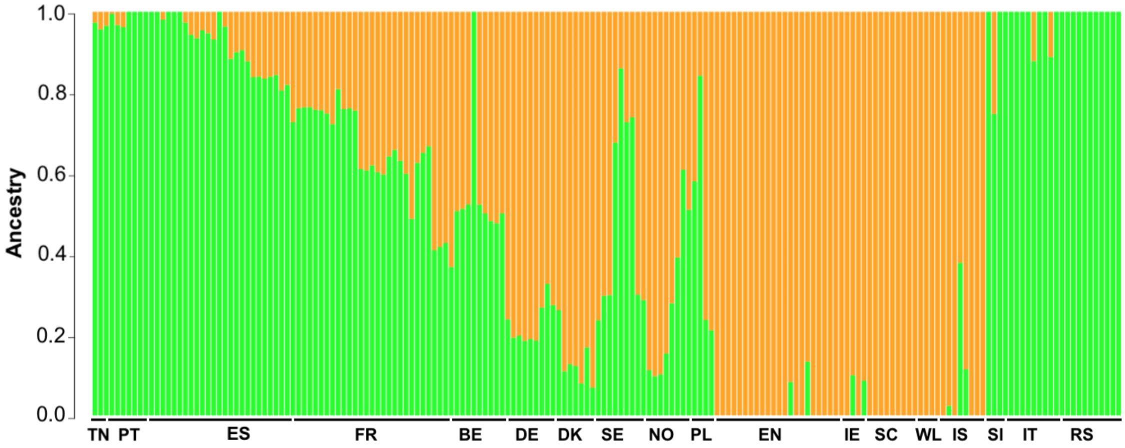


FIGURE 4.15: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=2. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norwey, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

K=2 shows the presence of two different ancestral components, been one of the them (green) the predominant one in the two traditional southern European refugia, while the second ancestral component is predominant in the Bristish Isles and Iceland. Both ancestral components admix in north-western Europe (Figure 4.15).

Increasing K to 3, a Balkan ancestral component appears, differentiating the ancestry into two traditional refugia. This third group is limited to Serbian, Italian and Slovenian samples, with a small contribution to the genomes from Iberian and African samples, among others. The major contribution of the Balkan ancestral component in Central Europe has been

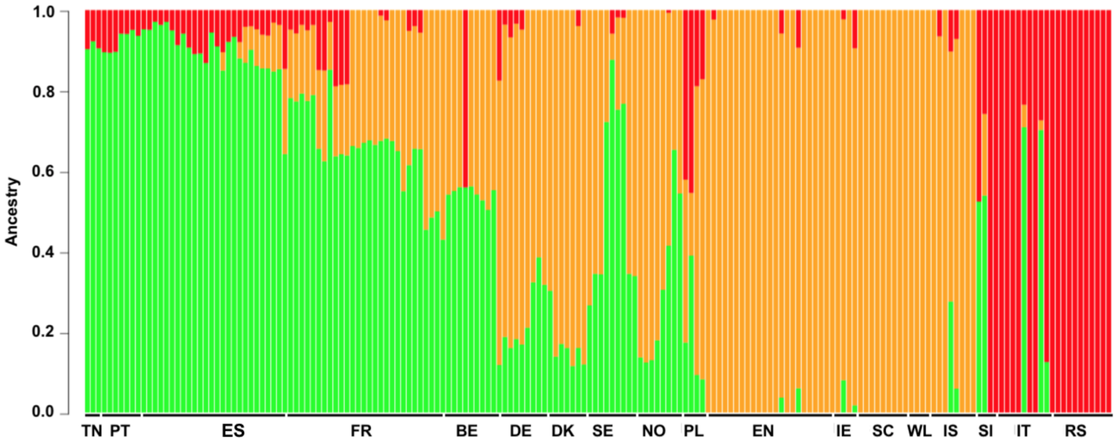


FIGURE 4.16: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=3. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norwey, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

found in samples from Poland an a sample from Belgium (Figure 4.16).

K4 shows the existence of 4 different ancestral components, two of them dominanting in the two traditional refugia used by *Apodemus sylvaticus* during the Quaternary glaciations (Figure 4.17).

The red ancestry component, that could have an Italo-Balkanic origin, appears only in samples from these locations and in a low proportion of admixture in Central European samples from France, Belgium, Germany and Poland. The green ancestry component, that could be considered as of the Iberian origin, dominates in Iberia and Tunisia and its contribution to the genetic structure of Central European population decreases towards central and eastern Europe, were it is replaced by the blue ancestral component. As previously observed in the PCA (Figure 4.11), the green ancestry component has an important contribution to the genetic variation of one of the Swedish populations. The blue ancestry component is characteristic of the northern European populations and dominates in Denmark and Germany, were admixture with other components is reduced. Finally, the orange ancestry component is widespread in the British Isles and Iceland, contributing to Norwegian and, to a smaller degree, to Central European populations.

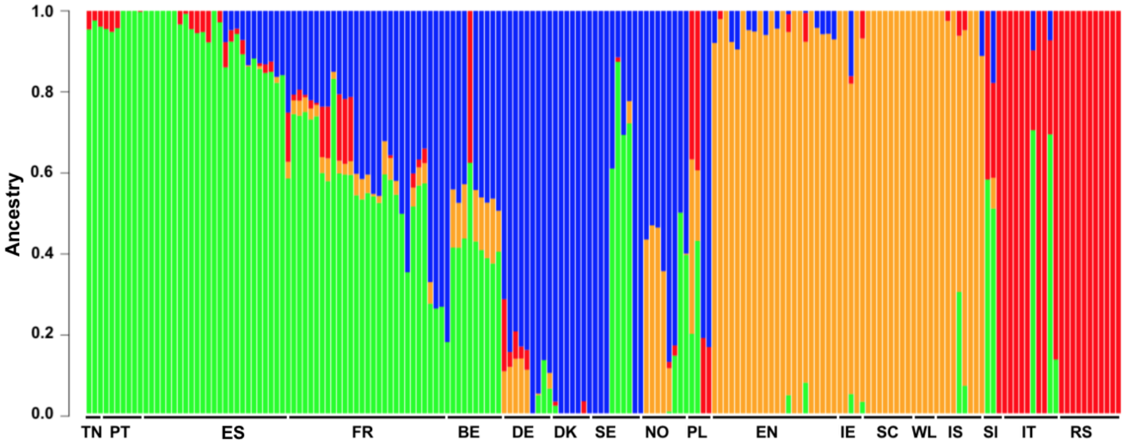


FIGURE 4.17: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=4. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norway, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

Increasing K to 5 reveals the existence of an Icelandic component, predominant in Iceland and admixed with the British component in Scotland, and, to a lesser extent, in England, Norway and Sweden (Figure 4.18)..

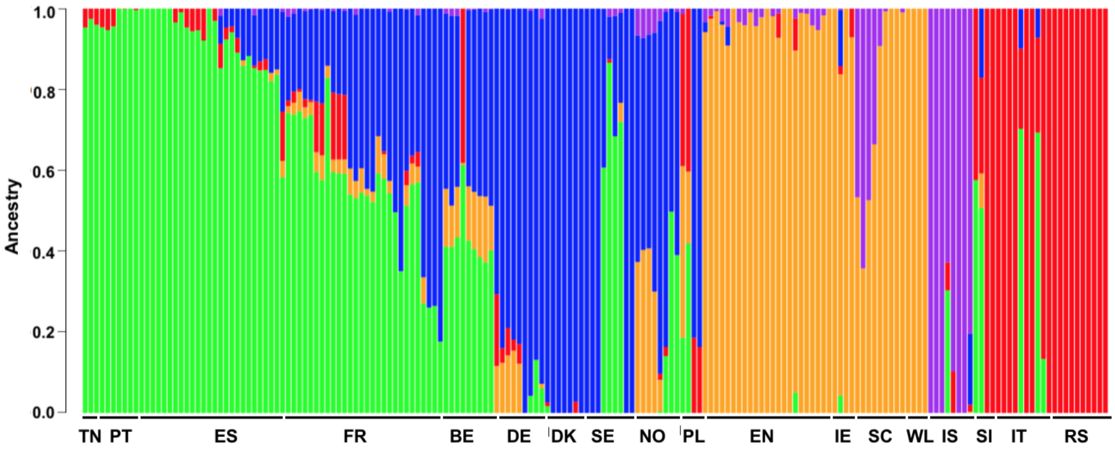


FIGURE 4.18: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=5. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norway, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

The sixth ancestral component identified in the admixture analysis only appeared in one sample from Belgium, two samples from Poland and one sample from Slovenia, and as a small contribution to the genetic background of samples from France, England, Ireland, Serbia and northern Europe (Figure 4.19).

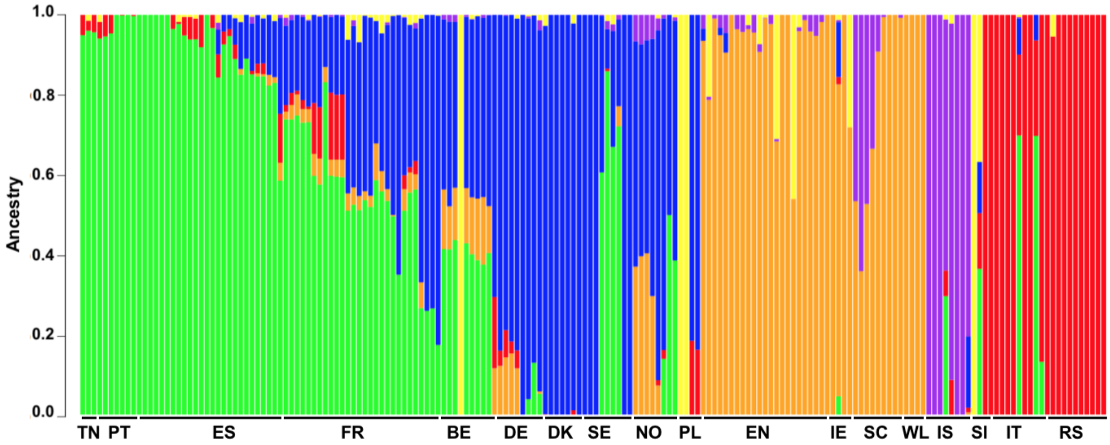


FIGURE 4.19: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=6. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norway, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL- Wales.

Increasing values of K to 7 revealed the existence of a component that is predominant in France and limited mainly to adjacent regions, particularly Spain or Belgium (Figure 4.20). The common pattern observed in all the above analyses of admixture is a close relationship of certain Swedish samples with Iberian samples. These Swedish genomes harbour a higher contribution from the Iberian component than from the other northern European samples. A similar pattern is also observed in Norway, where one of the populations appears admixed with the Islands component and the rest with the Iberian component.

From K=5, increasing values of K generates small groups with a no clear biological meaning and also separate the main groups into multiple smaller groups, separating Iceland from the rest British Isles.

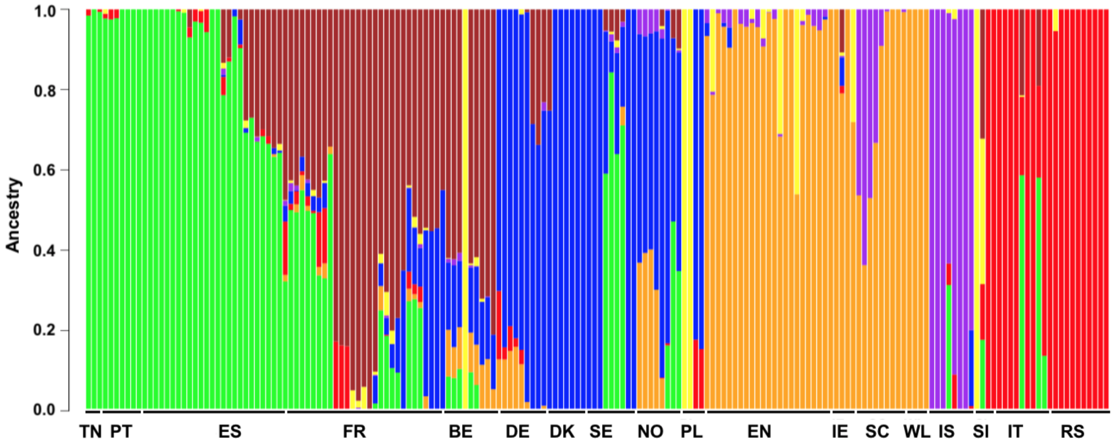


FIGURE 4.20: Admixture plot showing the ancestry proportions for each individual from *Apodemus sylvaticus* for K=7. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norwey, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, SI-Slovenia, TN-Tunisia and WL- Wales.

4.3.3.4 Population differentiation

At the population level, four different parameters were calculated to characterise the *A. sylvaticus samples*: nucleotide diversity, expected heterozygosity, percentage of polymorphic loci and F_{ST}.

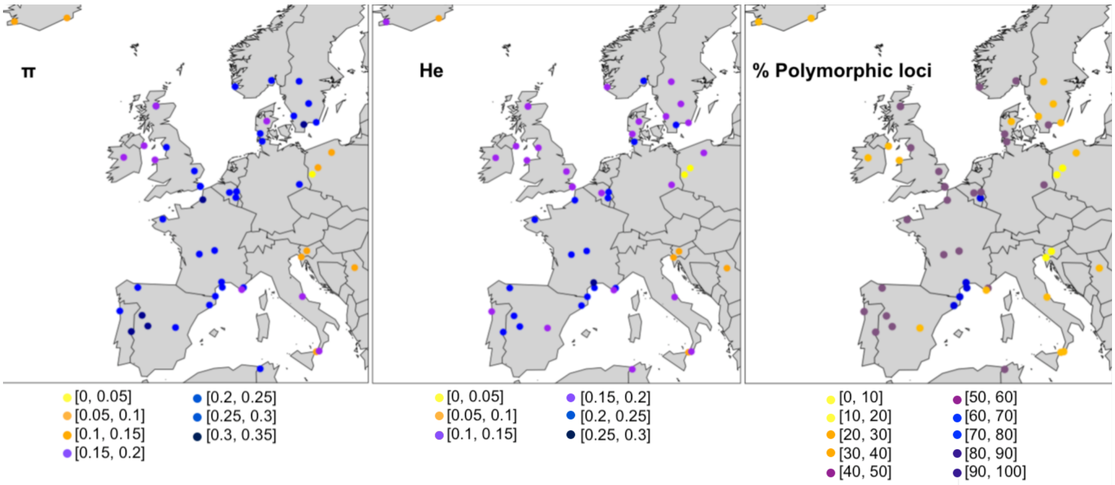


FIGURE 4.21: Distribution of different genetic diversity parameters in *Apodemus sylvaticus*: Π, He and % of polymorphic loci

TABLE 4.1: F_{ST} values between the 4 groups identified for DAPC analysis in *Apodemus sylvaticus*

	British isles	Balkans	Northern Europe
Iberian	0.05	0.05	0.03
British		0.08	0.03
Balkans			0.08

The highest nucleotide diversity was found in Iberia, France and Sweden, decreasing through the British Isles and central Europe (Figure 4.21). Minimums were found in Iceland and eastern Europe. A very similar pattern has been observed while analysing the expected heterozygosity and percentage of polymorphic loci, always showing maximums in Iberia or southern France and decreasing values through Central Europe. Minimum values of the above parameters have always been observed in the Italo-Balkan peninsulas and Iceland, but a low percentage of polymorphic loci were found in Sweden and Ireland (Data available in Appendix D, Section D.2.4).

Population differentiation shows large variation, with F_{ST} differing by an order of magnitude: from 0.06 (between populations from Montpellier (France) and Montseny (Spain)) to 0.56 (between populations from Krosno Odrzańskie (Poland) and Halmstad (Sweden))(Appendix D, Section D.2.5). The highest values have been observed in comparisons with historical samples from Poland and only one sample from each population. Exluding these two populations, the highest F_{ST} value, 0.41, between the populations of Halmstad (Sweden) and Ljubljana (Slovenia). There is a very weak positive correlation between genetic distance, expressed as $F_{ST}/(1\text{-}F_{ST})$ and geographic distance between samples (R^2=0.26, p-value < 0.01) (Figure 4.22).

Considering the groups previously identified for DAPC analysis, F_{ST} values ranged from 0.03 to 0.08 (Table 4.1). The largest differences have been found between the Italo-Balkan group (red) and the group from the Islands (orange) and northern Europe (blue). Italo-Balkan samples are more related to the Iberian group samples than to any other group, but the Iberian group is more closely related to the northern and island groups than to the Balkan one.

The tree shows two different clades, with a strong bootstrap support of 100% and 85% (Figure 4.23). These two clades correspond to the two traditional refugia used by *A sylvaticus*. The northern and islands groups appear within the same clade than the Iberian samples and they cluster together and

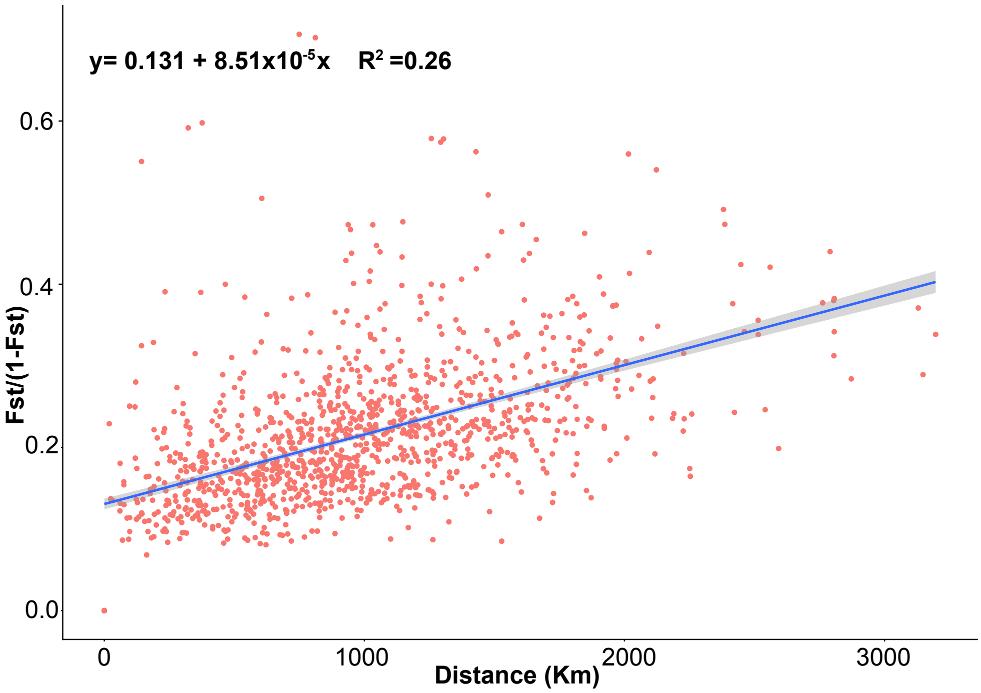


FIGURE 4.22: Correlation between genetic distance vs physical distance in *Apodemus sylvaticus*

with the admixed individuals from France an Belgium with a 99% bootstrap support. Samples from the British Isles an Iceland cluster together with a 100% bootstrap support and are close to the Scandinavian group (with 97 % bootstrap support). This group is closely related to the samples from northern Europe (Germany, Denmark and some French samples)(98% bootstrap support), but the internal relationship between the northern European and Island groups with French and Belgian samples are not well supported.

Within the Iberian group, the internal relationships are not well resolved, however we have been able to detect what might be an African lineage, in a red box, within the Iberian group. This group includes samples from Tunisia and it is well supported, with a 100% bootstrap support. Samples from southern France (in black at the top part of the tree in Figure 4.23 are basal to the rest of the Iberian, Nothern-Europe and Islands samples.

Four samples from *A. sylvaticus* cluster together with the outgroup of *A. flavicollis*. These samples are the individuals from Slovenia, Belgium and Poland that appear in PCA within the main group of *A. sylvaticus* but at a distance towards the Italo-Balkan group of *A. sylvaticus*.

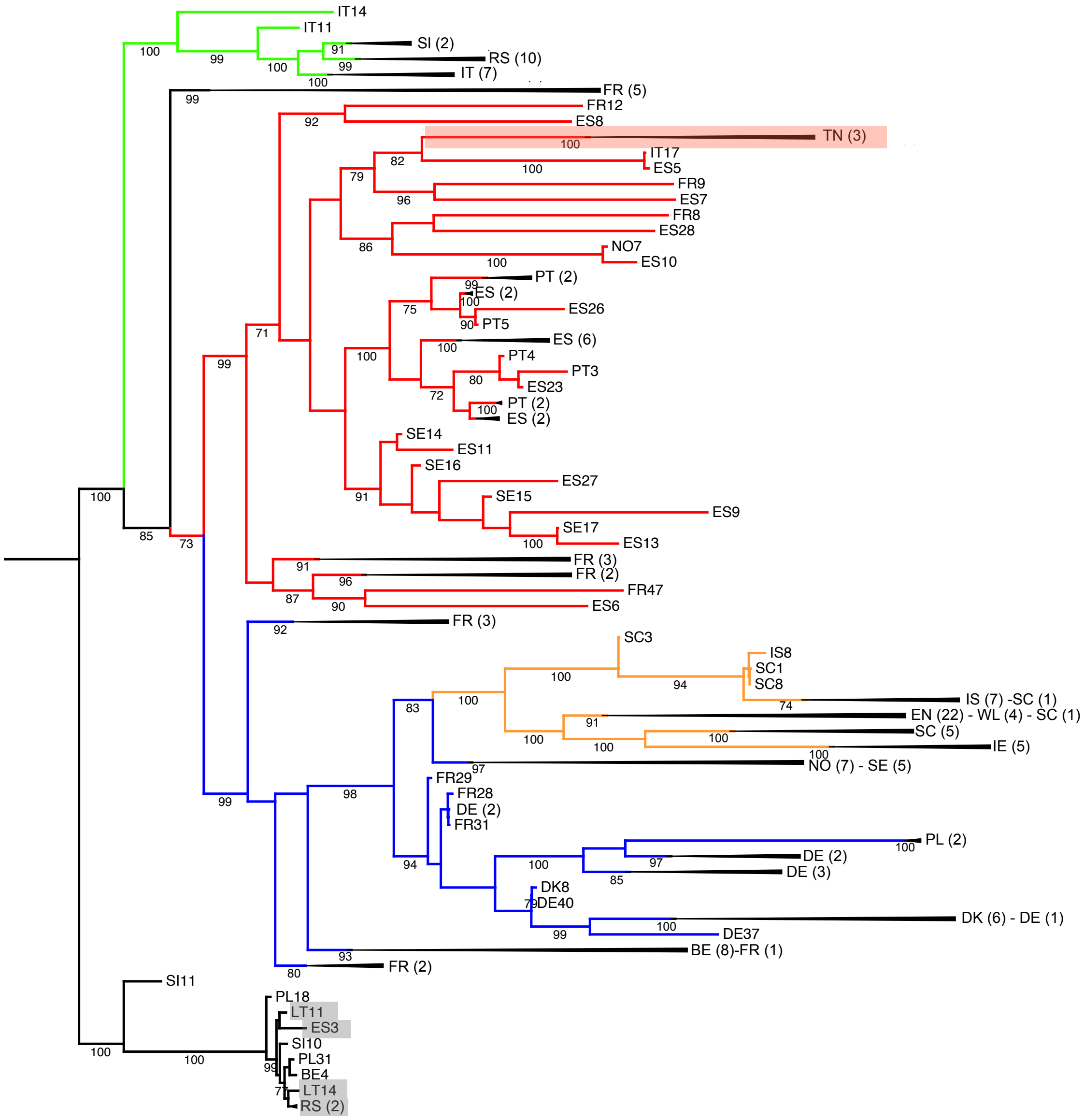


FIGURE 4.23: Maximum likelihood phylogenetic tree of *Apodemus sylvaticus* samples. Analysis performed with SNPhylo (Lee *et al.*, 2014) using samples from *Apodemus flavicollis* as outgroup (in grey boxes). The colours represent previously described refugia: Red: Italo-Balkan, green: iberian, blue: northern group and orange: the British Isles and Iceland. Bootstrap values higher than 70% are shown on the bottom of each branch. Clusters of samples from the same country or neighbourd regions have being collapsed and the bootstrap value of the cluster is shown at the beginning of the branch. The number of samples inside each collapsed branch appear between parentheses. BE-Belgium, DE-Germany, DK-Denmark, EN-England, ES-Spain, FR-France, IE-Ireland, IS-Iceland, IT-Italy, NO-Norway, PL-Poland, PT-Portugal, RS-Serbia, SC-Scotland, SE-Sweden, Sl-Slovenia, TN-Tunisia and WL-Wales

4.3.3.5 Inference of population history through DIYABC: long-distance genetic exchanges

DIYABC analysis was used to infer the most likely scenario to explain the similarities observed between long distance populations. There were run only for the swedish population resembling samples from northern Spain or southern France.

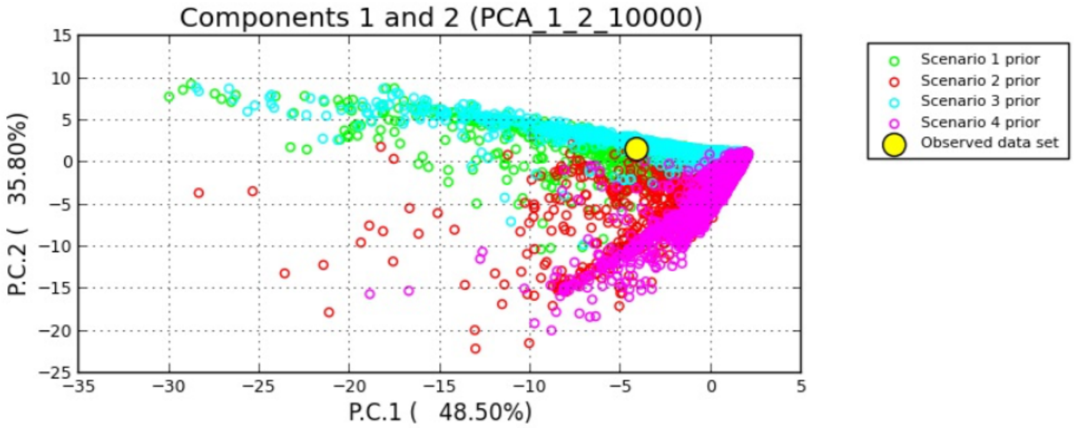


FIGURE 4.24: Principal Components Analysis for Pre-evaluation of scenarios prios combinations. Each dot represents a simulated dataset and the colour the scenario for which it has been simulated. The yellow big dot represents the observed dataset

DIYABC analysis discarded the possibility of an early split between Iberian and northern-Swedish samples, with the observed dataset appearing at the top of scenarios 1 and 3 prior distributions (Figure 4.24). From the 10000 reconstructed trees, 5920 trees supported scenario 1. This scenario indicates the arrival of the Iberian lineage to Sweden around 8 Ka, probably through the Doggerland. However the estimation of the population size and the time of the split are always on the edge of their distributions, potentially indicating poor support for the proposed scenario.

4.3.4 *Apodemus flavicollis*

4.3.4.1 Selection of parameters and variant calling

Genotype calling parameters for *A. flavicollis* were selected with the same procedure as for *A. sylvaticus*. As already seen in *Apodemus sylvaticus*, increasing values of m increases the coverage per sample (Figure 4.25)(Data available in Appendix D, Section D.3.1).

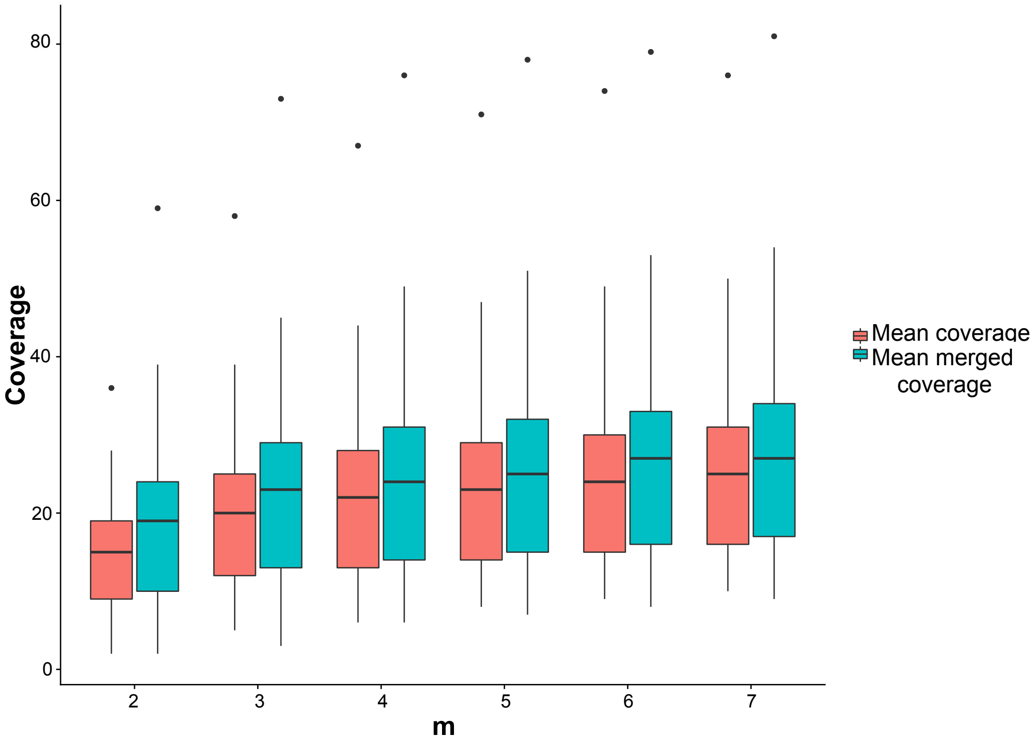


FIGURE 4.25: Distribution of mean coverage for each iteration of the m parameter for *A. flavicollis* dataset. Mean coverage, in red, is the average value obtained for each sample using only primary reads while mean merged coverage, in blue, is the average coverage value after merging alleles into loci. The boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range. Black dots are outliers outside the 10th and 90th percentile.

The selected parameters for calling the stacks and variants were: minimum number of identical, raw reads required to create a stack (m) of three, number of mismatches allowed between loci for each individual (M) of four and number of mismatches allowed between loci when building the catalogue (n) of four (Figure 4.26) (Data available in Appendix D, Section D.3.2.).

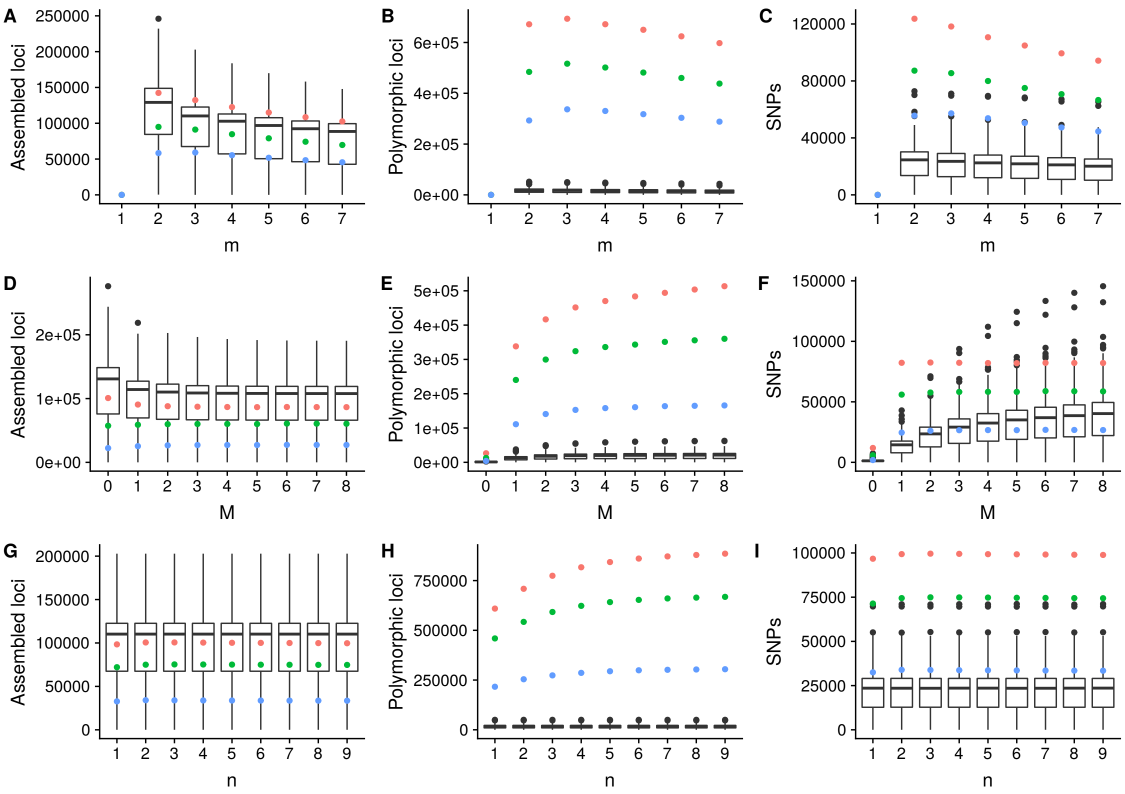


FIGURE 4.26: Variation in the number of assembled loci, polymorphic loci and SNPs for each iteration of m, M and n paramters for *Apodemus flavicollis* dataset. The boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range. Black dots are outliers outside the 10th and 90th percentile.. Blue circles represent data found in at least 40% of the samples, green circles, in at least 60% and red circles in at least 80% of the samples

A total of 710475 SNPs were called in 250 individuals. 38 individuals were removed due to missing data, leaving 212 individuals for downstream filtering. Out of 710475 SNPs called, 575254(80.96%) were removed after filtering for MAF and further 109456 (15.40%) were removed after failing the Hardy-Weinberg equilibrium test. 3893 sites (0.54%) were removed due to presenting mean depth values lower than a threshold of 20, leaving 21782 (3.06%) SNPs to be used in the downstream analyses. Another 15 individuals were removed at this point, due to the presence of erroneous genotypes, leaving a total of 197 individuals for the downstream analyses (List of individuals available in Appendix D, Section D.3.3). The distribution of samples per population can be found on Figure 4.27. In total, we analysed samples from 15 different countries: Austria (AT), Germany (DE), England (EN), Spain (ES), France (FR), Greece (GR), Italy (IT), Macedonia (MK), Poland (PL), Romania (RO), Serbia (RS), Russia (RU), Sweden (SE), Slovakia (SK) and Slovenia (Sl).

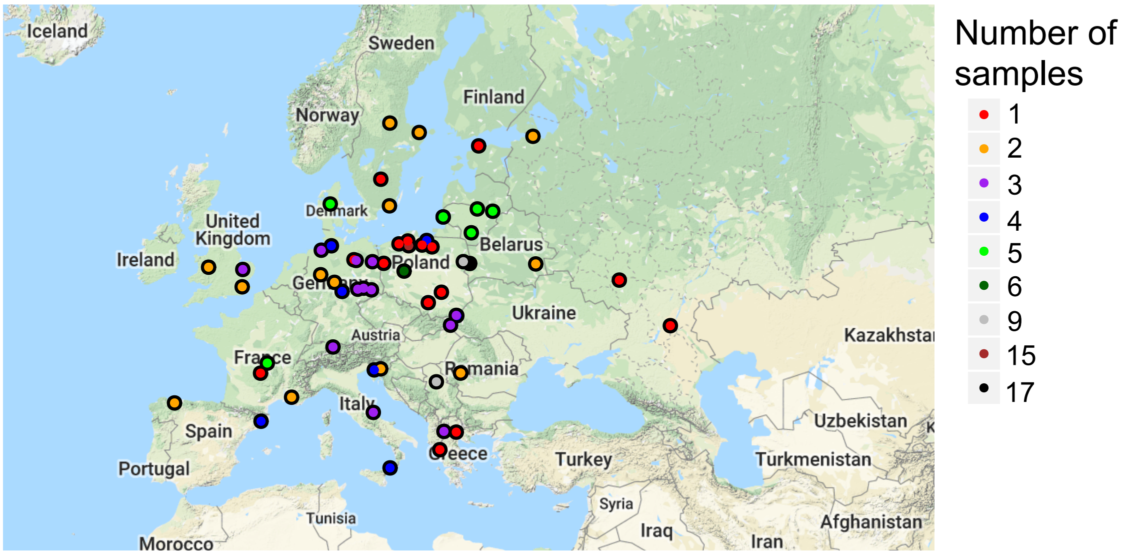


FIGURE 4.27: Distribution of the number of samples per population of *Apodemus flavicollis*

4.3.4.2 Phylogeographic history of *Apodemus flavicollis*

The first two axis of the principal component analysis (Figure 4.28) explained cumulatively 7.69% of the variance. Similarly to *A. sylvaticus*, the first two components allow the differentiation of two main groups. Most of the samples from Spain and Italy appear in the bottom left corner while the rest of the samples from the continent appear on the right side of the plot.

Starting from the bottom right side, there are samples from Greece, Macedonia and Serbia, the traditional refugia of this species. These samples appear closely related with samples from Slovenia, Lithuania, Poland , Germany and Austria, reflecting the geography of the continent. A third group of samples in the top right of the PCA plot include samples from Sweden and England.

Two small groups appear in between the continental one and the Italo-Iberian one, including samples from Spain and France. As seen before for *A. sylvaticus*, most of the samples are grouping with samples from neighbouring regions, with the exception of a single sample from Sweden, which appears close to Romanian and Russian samples and one samples from Poland that appear together with the French samples.

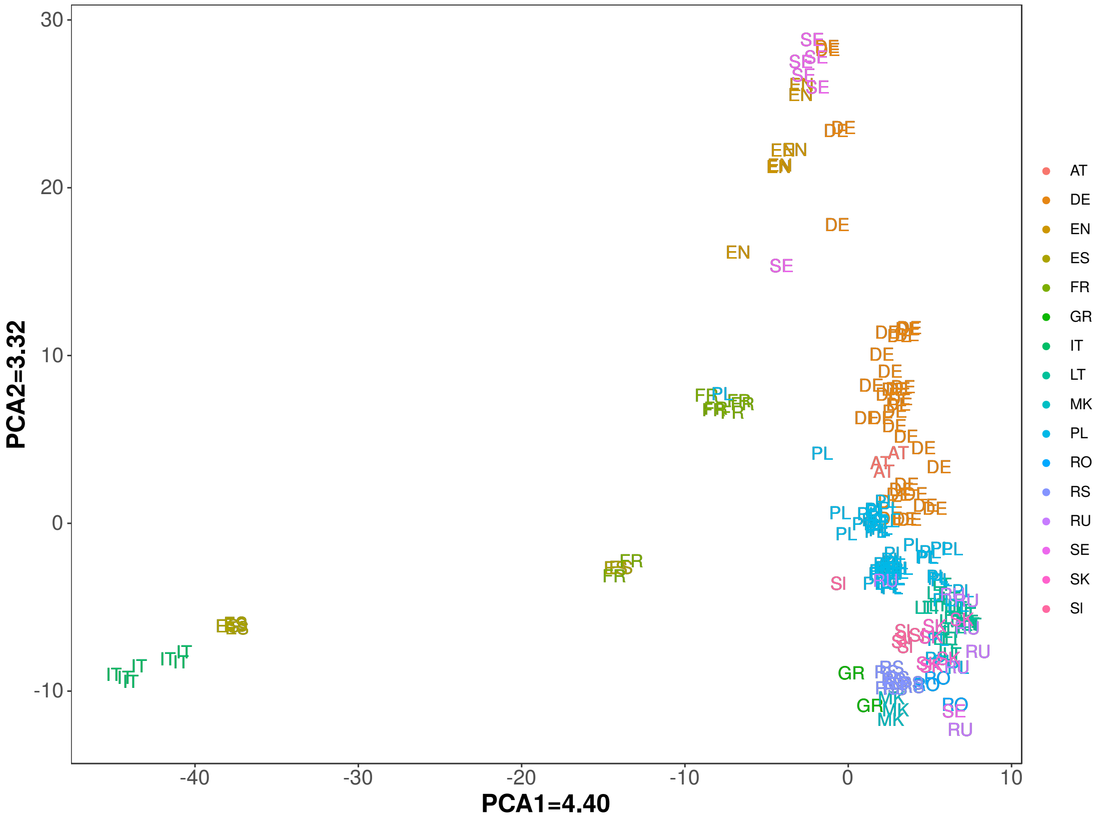


FIGURE 4.28: Principal Component Analysis for *Apodemus flavicollis* european dataset. The different countries are identify with different colours and also with different characters: AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

Find.cluster function from Adegenet identified three different groups (Figure 4.29), which are largely in accordance with the PCA results. The orange group includes the samples from the Balkans and continental Europe. The red group includes samples from England, Denmark, Sweden and a part of France, while the green group includes samples from Italy, Spain and the rest of France. Based on the distances between the groups in the DAPC analysis, the orange group (considered as the Balkan group) and the red group (northern European) appear to be more closely related than the green Italo-Iberian.

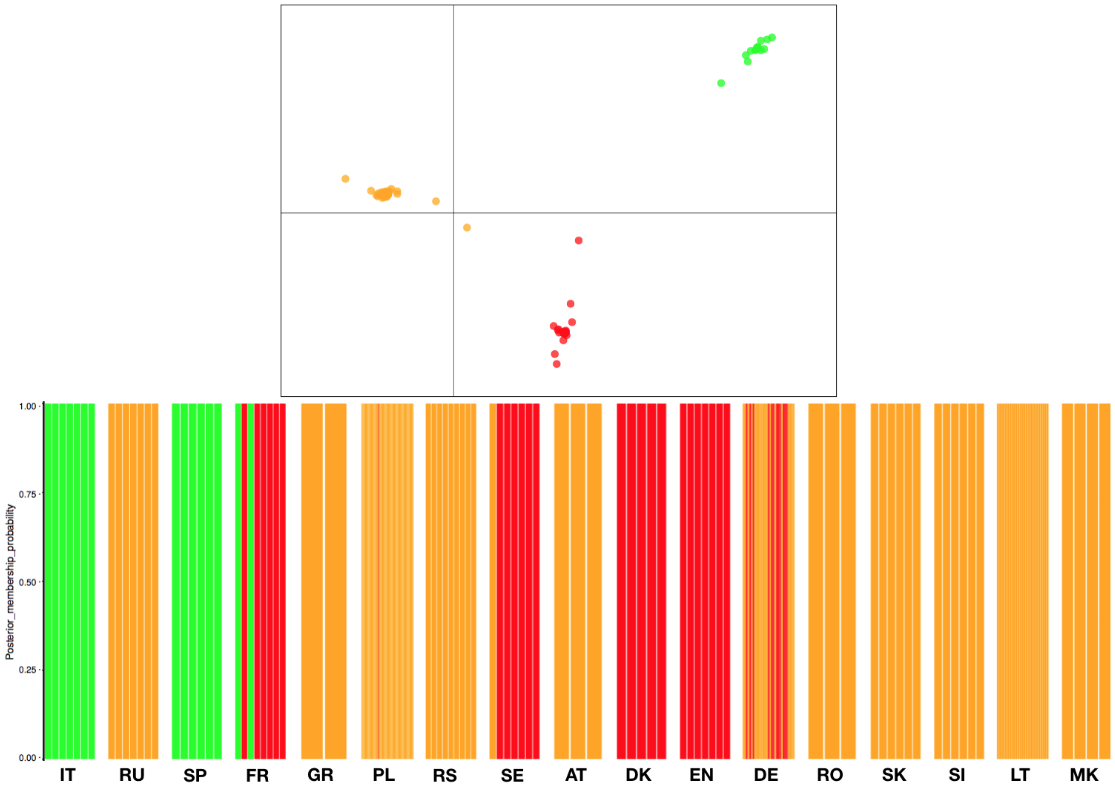


FIGURE 4.29: Compoplot and DAPC analysis for *Apodemus flavicollis* european dataset. A: Discriminant analysis of principal components scatterplot. Each dot represents one sample while the colour indicates the group to which each sample belongs to. B: Compoplot or barplot which represents the group assignment probability of each individual to the three inferred groups. Each bar represents one individual and each colour represents one of the three inferred groups. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

4.3.4.3 Admixture analysis

FastStructure analysis indicates that the best value of K is between 3 (model complexity that maximizes marginal likelihood) and 6 (Model components used to explain structure in data) in the 40% of the runs, between 3 and 5 in the 30% of the runs, between 3 and 7 in the 10 % of the runs and between 3 and 4 in the other 10 %.

The optimal value of K in ADMIXTURE was estimated through the cross-validation error values (Figure 4.30). The lowest cross-validation errors were found at K=3 or K=4.

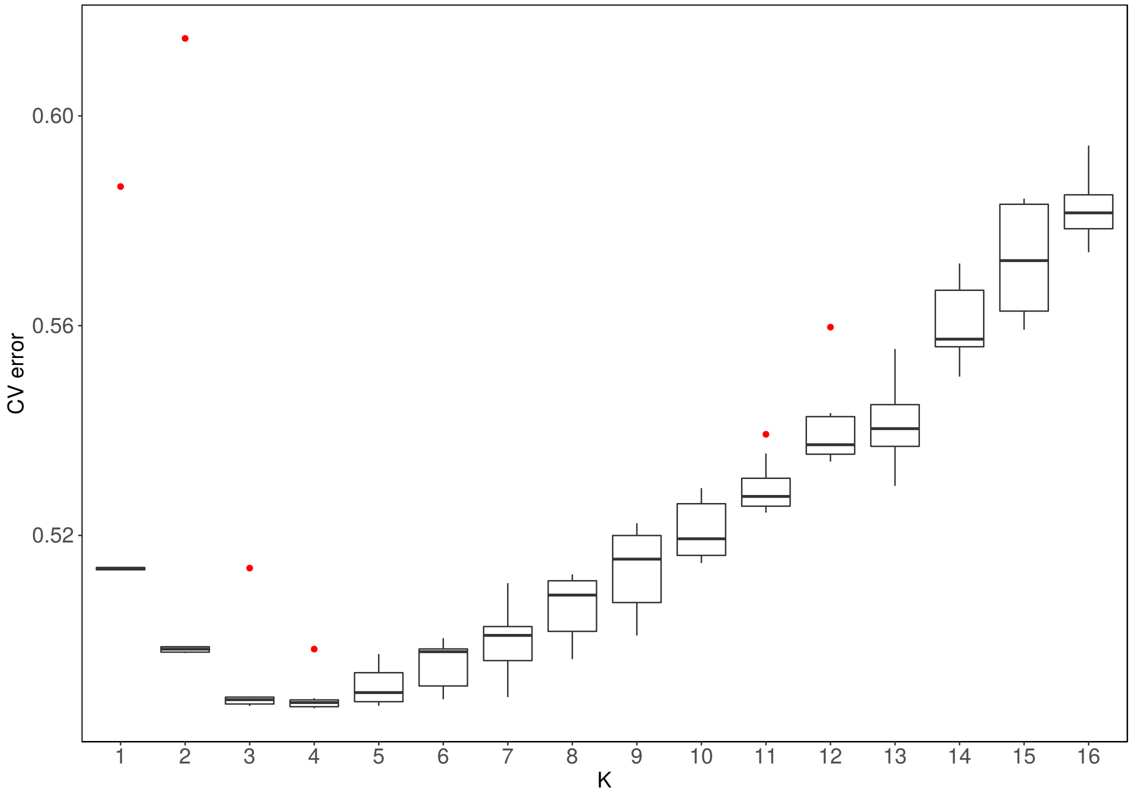


FIGURE 4.30: Distribution of cross-validation errors for 10 runs of 16 Ks each run for *A. flavicollis*. Results for *Apodemus flavicollis* european dataset. Box and whisker plot were the boxes represent the 25th-75th percentiles and the whiskers represent values higher or lower than 1.5 times the interquartile range or the distance between the two hinges of the box. Red dots are outliers.

K=2 show the presence of two different ancestral components. The orange component could be considered of Balkan origin.It is widespread through the east of Europe, getting reduce in northern and eastern European populations. The green component is the main component from Italy-Iberia, England, Sweden and Denmark populations and it is presented in different level of admixture in most of the others samples (Figure 4.31).

Increasing K to 3 does not affect the orange component, but causes the green component to divide in two. The contribution of the green component now is very limited in European samples outside Iberia and Italy, peninsulas that have traditionally been considered to not contribute to the recolonisation of Europe after the last Ice Age. In agreement with this hypothesis, the Ibero-Italian component is only present as a very low level of admixture in the rest of the samples (Figure 4.32).

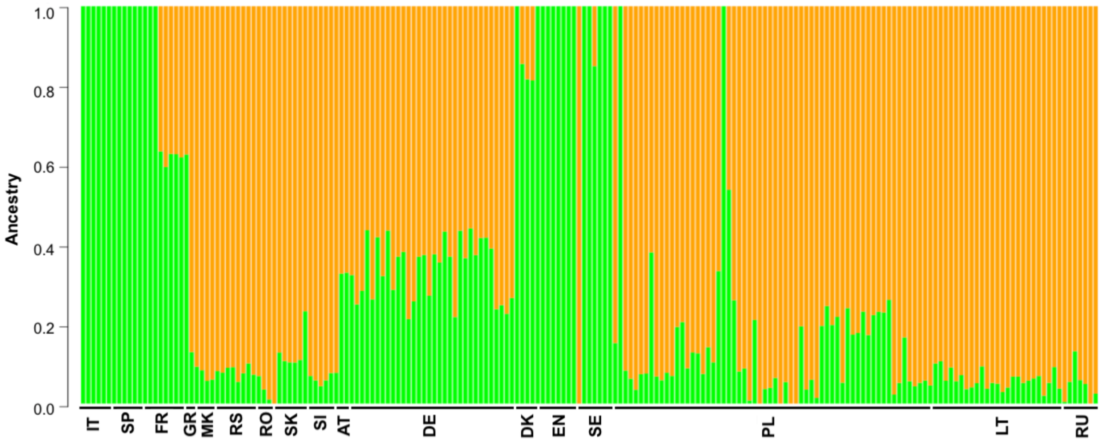


FIGURE 4.31: Admixture plot showing the ancestry proportions for each individual from *Apodemus flavicollis* for K=2. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

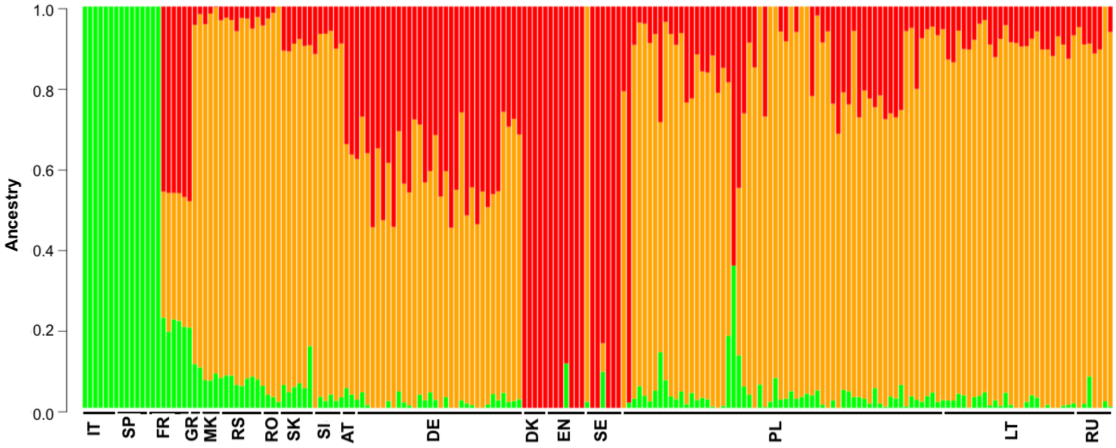


FIGURE 4.32: Admixture plot showing the ancestry proportions for each individual from *Apodemus flavicollis* for K=3. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

The new red component is the main component in Denmark, England and Sweden and it is present in different proportions of admixture in samples from Central Europe. Its presence is highly reduced in the Balkan peninsula and completely absent in the Italo-Iberian peninsulas.

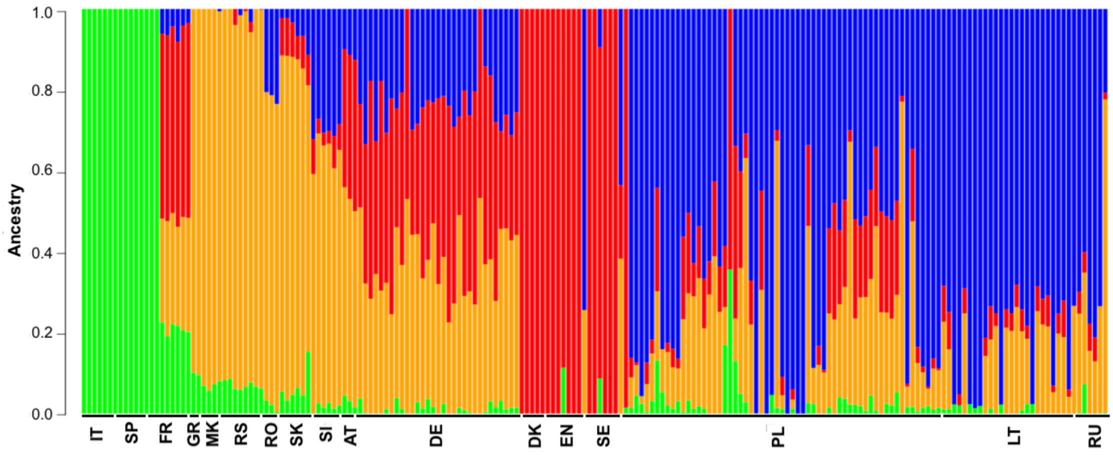


FIGURE 4.33: Admixture plot showing the ancestry proportions for each individual from *Apodemus flavicollis* for K=4. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

Increasing K to 4 causes the split of the orange component. The new blue component is predominant in Poland and Lithuania and contributes to the genetic background of central European populations. Its origin appears to be in northeastern Europe and its prevalence decreases southwestwards. The remaining orange component now represents the traditional Balkan refugia, whose contribution decreases with the distance to the peninsula (Figure 4.33).

Increasing K to 5 divides the red component in two (Figure 4.34). The red component now includes only the English samples and its contribution to other populations is very limited. The new purple component is the main component in Sweden, where it is the only component for most of the samples, and Denmark. This component contributes in a high proportion to the ancestry of the neighbouring regions, such as Germany, France, Poland, Slovenia and Austria, and its contribution decreases with the distance from northern-Central Europe. In addition, there are two samples (from Sweden and Poland) with unusual pattern of admixture that is very different to other samples from the same locations.

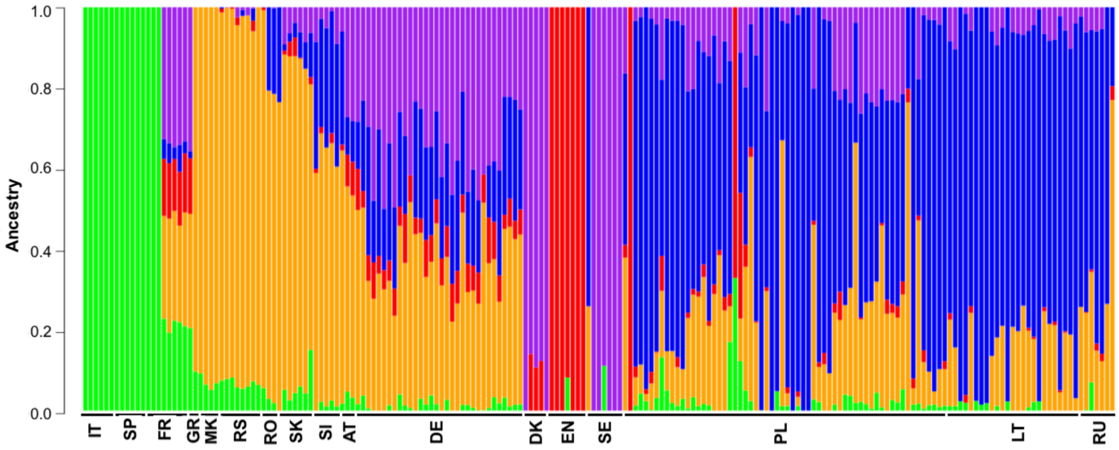


FIGURE 4.34: Admixture plot showing the ancestry proportions for each individual from *Apodemus flavicollis* for K=5. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

4.3.4.4 Population differentiation

The highest nucleotide diversity has been found in samples from Romania, Russia and Poland, decreasing through Central Europe and reaching minimum values in southern France and Iberia (Figure 4.35). Nucleotide diversity in Italy and northern Sweden was also quite low. A similar pattern has been observed for the expected heterozygosity and the percentage of polymorphic loci. The highest expected heterozygosity has been found in samples from Serbia and Slovenia, with minimums in southern France, Italy and Spain. Three samples from Poland, Italy and southern Sweden also showed relatively low values of expected heterozygosity. The highest percentage of polymorphic loci were found again in Serbia. High levels of polymorphic loci were only found in Serbia, Slovenia, Slovakia, Poland, Lithuania, three populations in Germany and one population in France (Data available in Appendix D, Section D.3.4).

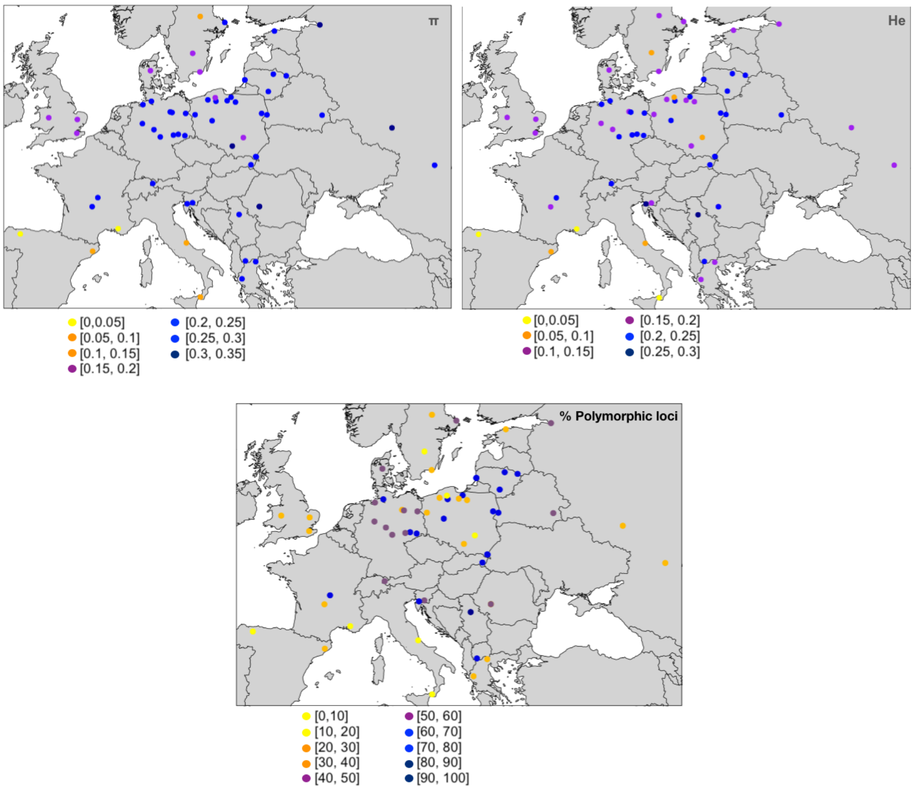


FIGURE 4.35: Distribution of different genetic diversity parameters through *Apodemus flavicollis* distribution range: Π, He and % of polymorphic loci

F_{ST} values between populations ranged from 0.04 to 0.59 (median= 0.14, average= 0.17, stdev=0.10) with the highest values found in comparisons between Iberian or French samples against distant populations from Sweden, Germany and Russia (Data available in Appendix D, Section D.3.5). There is a very weak correlation between genetic distances and the distance between the samples (R^2=0.15, p-value<0.01) (Figure 4.36), analysing them together or by groups based on the age of the samples.

TABLE 4.2: F_{ST} between the three groups identified for DAPC analysis in *Apodemus flavicollis*

	Northern Europe	**Balkan**
Italo-Iberian	0.061	0.018
Northern Europe		0.013

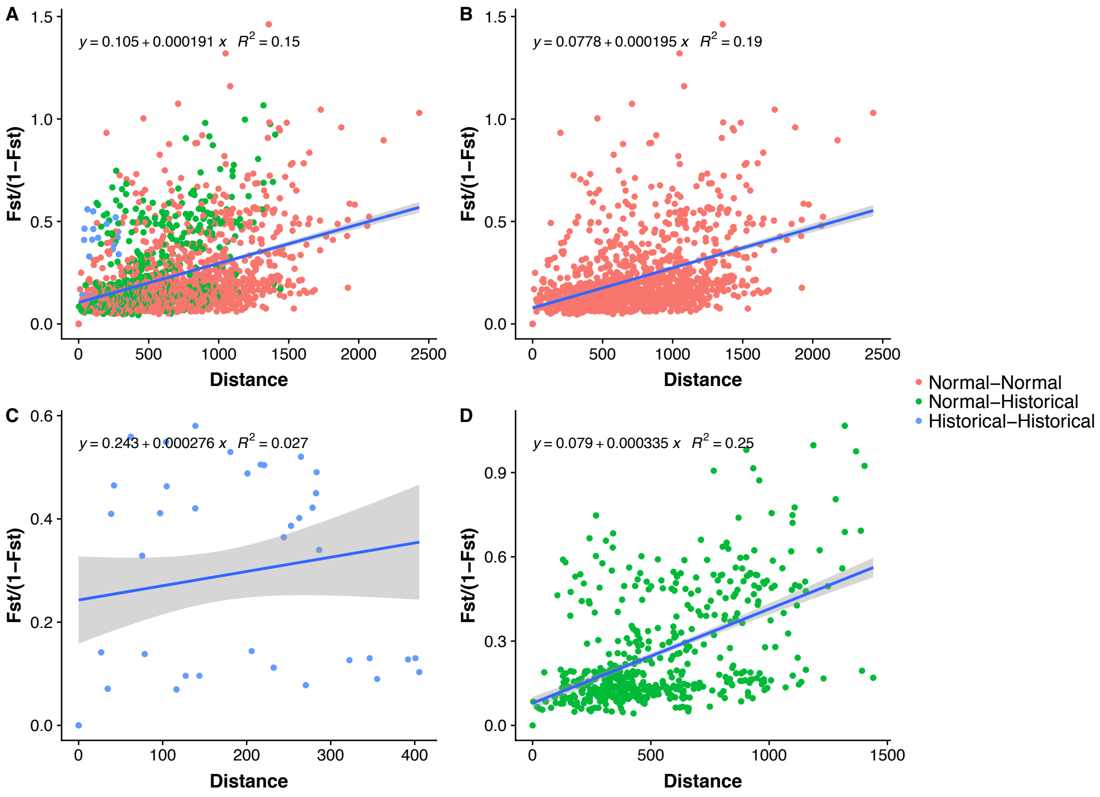


FIGURE 4.36: Correlation between genetic distances (F_{ST}/(1-F_{ST})) and physical distances in *Apodemus flavicollis*. Normal make reference to modern fresh samples and historical make reference to the dry skins collected from Museums.

F_{ST} values between the groups, considering only 3 main groups identified for DAPC analysis (Figure 4.29), show that the Balkan (orange) and the northern European (red) groups are the most similar, with the Italo-Iberian (green) group being closer to the Balkan group than to the northern one.

TABLE 4.3: F_{ST} values between groups identified for K=4 in *Apodemus flavicollis*

	Northern Europe	**Balkan**	**Eastern Europe**
Italo-Iberian	0.061	0.031	0.031
Northern Europe		0.022	0.025
Balkan			0.010

Considering the existence of 4 groups, the closest relationship has been found between the Balkan (orange) and the eastern European (blue) groups (Table 4.3) . The northern European group (red) group is closer to the Balkan one than to any other group and is close to the eastern group.

Similarly to what we have found considering only three groups, the Italo-Iberian (green) group is the most distant to all the other groups, with the highest F_{st} values found between the Italo-Iberian group and the northern one. The differences in F^{st} between the Italo-Iberian group and the Balkan or the eastern group are very similar.

A. flavicollis samples form two main groups on the phylogenetic tree (Figure 4.37). The separation between samples from the Mediterranean peninsulas is clearly visble: samples from Spain and Italy (in green) cluster together with a 100% bootstrap support, with all the other samples from Europe clustering together also with a 100 % support (Figure 4.37).

The other groups, coloured in orange (Balkan), blue (eastern Europe) and red (northern Europe) are also well supported (98 %, 97% and 96 % bootstrap support). Samples in orange correspond with the area traditionally described as the main refugia for *A. flavicollis*, the Balkans. The blue subtree includes mainly samples from Lithuania, Poland, and Russia while the red subtree includes only samples from Denmark, Sweden and England. Samples in olive, in a basal position within to the main European group, exhibit a high degree of admixture between different ancestral components. French samples showed, mainly, admixture between the Italo-Iberian, Balkan and northern European component, while Austrian samples have a lower contribution from the Italo-Iberian component and a higher contribution from the Balkan ancestral component and a small contribution from the northern European component. Samples in purple also showed a high degree of admixture between the Balkan, the northern and the the eastern European components and appear in a basal position within the northern European group.

Moreover, the outgroup, which includes 6 samples from *A. sylvaticus*, have recruited 4 samples that have been previously morphologically described as *A. flavicollis*: 2 samples from Spain and 2 from France. These are the same samples that appeared in between the two main groups on the PCA plot in *Apodemus sylvaticus analysis* (Figure 4.28). In the ADMIXTURE analysis, the Spanish samples present the contribution of a single ancestral component, the Italo-Iberian, while the French samples present admixture with other components, similarly to other samples from France.

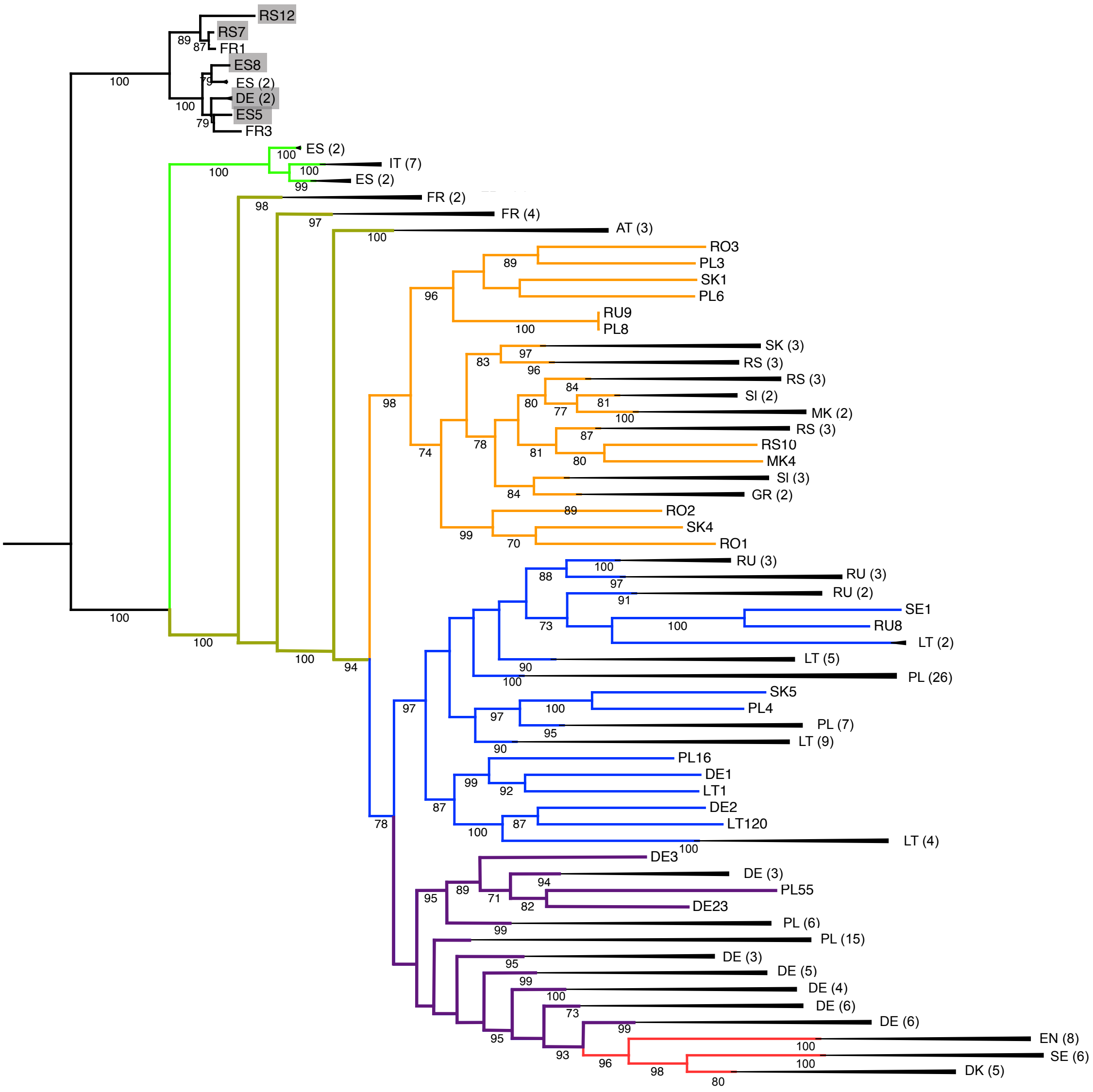


FIGURE 4.37: Maximum likelihood phylogenetic tree of *Apodemus flavicollis* samples. *Apodemus sylvaticus* samples have been used as outgroup and appear inside grey boxes.The colours indicated the main groups identified: green, Italo-Iberian group, orange, Balkan group, blue, eastern group and red, northern group. Bootstrap values higher than 70% are shown on the bottom of each branch. Clusters of samples from the same country have being collapsed and the bootstrap value of the cluster is shown at the beginning of the branch. The number of samples inside each collapsed branch appear between parentheses. AT-Austria, DE-Germany, EN-England, ES-Spain, FR-France, GR-Greece, IT-Italy, MK-Macedonia, PL-Poland, RO-Romania, RS-Serbia, RU-Russia, SE-Sweden, SK-Slovakia and Sl-Slovenia.

4.3.5 Species differentiation

Principal component analysis shows clear differentiation between the two continental groups identified for each species (Figure 4.38). Samples from the Balkans and Italy from *A. sylvaticus* appear distant from the main group and closer to *A. flavicollis*, although, this group is still clearly distinguishable. Three of the samples from *A. sylvaticus* that clustered with *A. flavicollis* on the phylogenetic tree and which position on PCA and genetic structure was ambiguous, do appear closer to *A. flavicollis*, but slightly differentiated from the main group. The fourth sample, from Slovenia, appear equidistant to both species, complicating the clear identification of this particular sample. Another sample from Slovenia, but in this case originally identified as *A. flavicollis*, also appears distant from the other samples from a similar location and close to the other Slovenian sample, complicating its species designation. One sample from France originally identified as *A. flavicollis* clearly clustered with *A. sylvaticus*, indicating a missidentification of this sample.

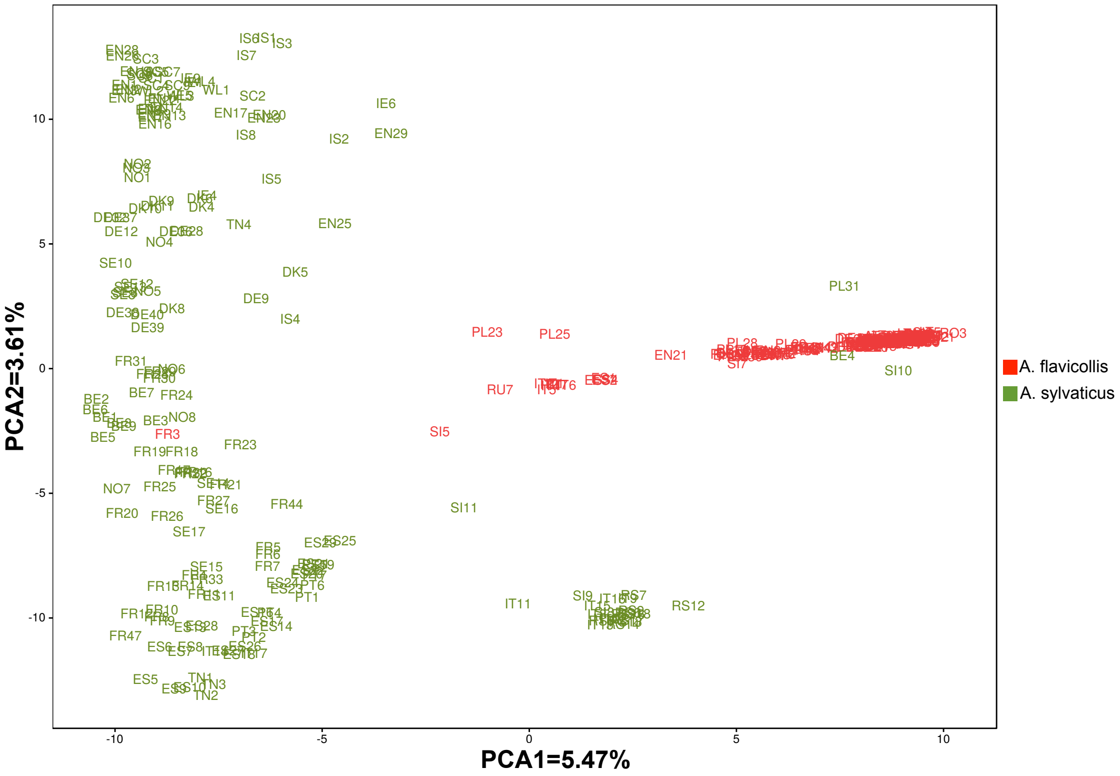


FIGURE 4.38: Principal components analysis for all the samples included on our analysis

4.4 Discussion

In this work, I have presented the results of a whole genome, high density genotyping by double digestion restriction-site-associated DNA sequencing (ddRAD-seq) on two species of mice from the genus *Apodemus*. To my knowledge, this is the first application of the whole-genome approach to study these organisms. It has allowed me to generate sequences from thousands of different genomic regions for *Apodemus flavicollis* and *Apodemys sylvaticus*, identify tens of thousands of SNPs markers and perform continental-scale analysis of the relationships between multiple populations. Ultimately, these resources will significantly contribute to the development of *Apodemus* as a model organism.

It is worth noting that this work constitutes the first genome-wide undertaking for our research group and myself and, indeed, it has been a learning experience, not only on population genomics and phylogenomics but also on bioinformatics. The work presented here is based on 364 Gb of raw data, from 428 individuals that I generated starting from tissue samples all the way up to optimisation of a basic pipeline for future analysis of RAD-seq data in our group. All the scripts used on this project have been made publicly available to support future research in the field and enable the reproducibility of my results. Furthermore, all the bioinformatics skills develop during this project are not exclusive for population genetics and phylogeographic studies, but can be applied in other organisms and related fields. It speaks to the enormous and exciting progress that happened in the field of ecological genomics in the last decade that such insights are now accessible even to small groups with limited budget.

I am preparing two manuscripts for publication based on this work. One manuscript describes the results presented in Chapter 2, the analysis of the Polish populations, with emphasis on *Apodemus flavicollis*. The second manuscript is based on the European phylogeography results described in this chapter for both species. Following successful testing on the performance of PCR duplicates detection in the modified quaddRAD adapters performed by a fellow PhD student, Rohan Raval, in our group, I anticipate publishing the adapter sequences and their application as a technical note.

4.4.1 Technical considerations

4.4.1.1 PCR duplicates

RAD-sequencing approaches have proven to be a cost-effective method for genotyping and characterising populations for a wide range of species, even in the absence of a reference genome (Rodriguez-Ezpeleta *et al.*, 2017; Pegadaraju *et al.*, 2013; Sharma *et al.*, 2012). The original RAD-seq protocols (Miller *et al.*, 2007; Baird *et al.*, 2008) use only one enzyme for digesting DNA and include a random shearing step, which produces sequences of different lengttextcolorredhs. This feature in turn allows for identification and removal of PCR duplicates. Modifications of these protocols, such as ddRAD-seq (Peterson *et al.*, 2012), which use two different enzymes for DNA digestion, increase the flexibility to select a particulate number of loci for sequencing, the reliability of recovering the same regions in independent samples and experiments and, importantly, decrease the costs associated with library preparation. However, one of the most important disadvantages of ddRAD-seq protocols (Peterson *et al.*, 2012; Poland and Rife, 2012) is that detection and removal of PCR duplicates is now impossible.

The majority of PCR duplicates originate during the library preparation, in the PCR amplification step, but some can also arise during the cluster generation step in Illumina sequencing. The duplicate sequences can be confounded with real reads and can modify allele frequencies and increase homozygosity (Pompanon *et al.*, 2005). Much has been discussed about the advantages and disadvantages of the modified RAD-seq protocols (Andrews and Luikart, 2014; Puritz *et al.*, 2014b), contributing to the development of newer methods trying to eliminate some of these issues, in particular of duplicate identification and removal (Franchini *et al.*, 2017; Schweyen *et al.*, 2014).

In this work I have only followed ddRAD-seq protocols, Poland and Rife (2012) for the library preparation in the pilot project described in Chapter 2 and a modified version of the quaddRAD protocol I developed based on Franchini *et al.* (2017) for the library preparation in the main project. As the first protocol did not allow the identification of PCR duplicates, I prepared the library using the minimum number of PCR cycles required (12) to produce a visible amplification on an agarose gel (Figure 4.39).

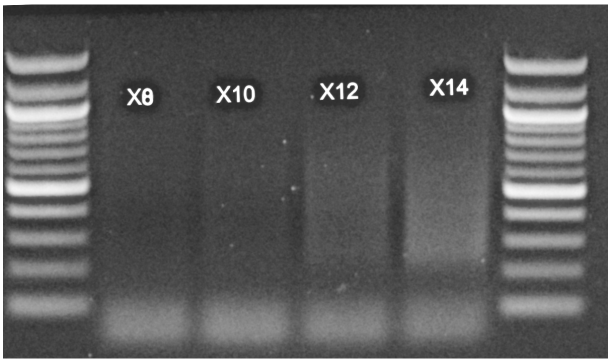


FIGURE 4.39: Test for the minimum number of cycles required to amplify the library. Agarose gel showing results of the amplification of the library performed with different number of PCR cycles. 12 cycles were finally selected for library preparation (see Methods in Chapter 2).

4.4.1.2 New adapter design to eliminate PCR duplicates

This approach does not solve the issue with PCR duplicates, but, at least, helps to limit it as much as possible. In order to eliminate them completely, we chose a new ddRAD-seq protocol for the main project presented here. The new design incorporates a degenerate base region in the adapters, effectively generating a random 4-nucleotide sequence in each adapter, enabling detection of PCR duplicates (Figure 3.6) (Franchini *et al.*, 2017; Schweyen *et al.*, 2014). The new adapters included all possible combinations of nucleotides: in total, 108 different combinations of random 4-mers were attached to each of the adapters. When two identical sequences also contained the same barcode and the same sequence in the degenerate base region, they were considered PCR duplicates and all but one of them were removed. The performance of the adapters and the selection of the number of cycles for library amplification have been proved effective on Chapter 3, producing, on average, a 2.65% of PCR duplicates (Stdev: 0.61), that were removed from further analysis.

4.4.1.3 Compatibility of the two ddRAD-seq protocols

The two protocols presented in this thesis should allow to recover the same set of loci, i.e. they should allow both datasets to be combined for the phylogeography study. However, the combination of both datasets was difficult and most of the sequences from Chapter 2 were removed in the phylogeography analysis due to having more than 50% of missing data. In particular, the 10 samples from *A. sylvaticus* and 35 samples from *A. flavicollis* from Chapter

2 were excluded from the phylogeography analysis. Samples sequenced in Chapter 2, due to the issues with degraded adapters discussed in Chapter 3, had a lower coverage (10x for m=2 and 12x for m=3, Appendix A Figure A.2) than the samples sequenced in the phylogeography project (18x for m=2 and 22x for m=3)(Chapter 4, Figures 4.8, 4.25). The settings selected for the analysis of the complete dataset (m=3) were stricter than the ones used on Chapter 2 project (m=2) and some loci could have been lost due to a lower coverage.

Depending on the aims of a given project, different filtering approaches can be followed, for example, to reduce the number of loci to consider but increase the number of samples sequenced for these loci, keeping a higher number of samples overall. In this case, and due to the sufficient number of samples available from Poland and neighbouring regions, we chose to excluded them from further analysis, while keeping the best parameters calculated for the main phylogeography dataset.

4.4.1.4 Advantages of the modified quaddRAD protocol

From a methodological point of view, the development of the modified quaddRAD protocol (Franchini *et al.*, 2017) allows, for the first time, the identification, quantification and elimination of chimeric sequences from the sequencing reads. Until now, the prevalence of these artefacts and their effect on RAD-sequencing analysis has been unknown in the literature. PCR chimeras can be produced in each PCR step, hence they can appear during the library amplification, but also during sequencing, on the flowcell. Our design has the power to detect both types of chimeras, but not simultaneously. For example, during the library preparation performed in Chapter 4, PCR amplification was performed for each sample individually, therefore eliminating possibility of producing chimeric molecules. In consequence, all the chimeras that have been identified in this project are sequencing chimeras only - those that arose on the flow cell during bridge amplification. However, current libraries prepared in our group by Rohan Raval have multiplexed samples with different inner barcodes and equal outer barcodes before PCR amplification. Therefore the chimeras observed in that output will represent the total number of chimeric reads, those produced during library preparation PCR and during bridge amplification on the flowcell. It will not be possible to differentiate between them.

The percentage of sequencing chimeras found in my libraries range from 0.90% to 1.43% (Average:1.07, Stdev:0.24). Given their low numbers, they are likely the least problematic for the analysis because they will be elimiated during analysis due to their low coverage. I expect higher number of PCR chimeras to be observed in Rohan Raval's libraries, due to the combination of sequencing chimeras and PCR chimeras. In terms of investigating their effect on the genotyping performance, the best way to test it would be by mixing the chimeras identified in Chapter 4 with sequences from another library prepared using combinations of the inner barcodes (for example: i5-1:i7-2 or i5-1:i7-3). These combinations of barcodes have not been used on our library preparation, as we only used fixed pairs of inner adapters, and are the ones that are present in chimeric sequences. Only following this approach we could have samples with the same barcode combination that chimeric sequences at the same time that we could differentiate between real reads and chimeric sequences and study their effect on the analysis.

PCR duplicate removal is arguably more important than identification and removal of chimeric sequences, as PCR duplicates, due to their higher coverage, will pass the constrains set during the RAD-sequencing analysis and can inflate the proportion of homozygotes loci, a problem known as allele dropout. PCR duplicates, therefore, can lead to a significant bias in population genetic analysis (Schweyen *et al.*, 2014). The percentage of PCR duplicates found in my project, even when I reduced the number of cycles used to the minimum (Figure B.1), is higher than the percentage of sequencing chimeras (Average: 2.65, Stdev: 0.61). The percentage of PCR duplicates found on this study is much lower than elsewhere, where up to 33.48% of the reads were classified as PCR duplicates (Schweyen *et al.*, 2014). Similar results to Schweyen *et al.* (2014) were also found by Franchini *et al.* (2017) when using low input DNA (0.01 ng) and a high number of PCR cycles (x26). However, Franchini *et al.* (2017) found similar percentage of retained reads, to those obtained in our analysis, while using 10 ng of DNA and only 12 PCR cyles (3,27%). Therefore, our results are comparable to previously published data obtained under similar conditions.

Overall, this work represents a development and successful validation of a library preparation and sequencing protocol for a modified quaddRAD protocol, which includes higher multiplexing and ability to identify PCR or sequencing duplicates and chimeras. This upgraded quaddRAD protocol is now being used in our group in another study on 300+ samples of *Apodemus* and forms a basis of novel pull-down protocol to reliably genotype even highly degraded samples, developed in collaboration with Prof. Nadir Alvarez from the University of Geneva, Switzerland.

The analysis performed in this work attempted to characterise the major patterns of genetic diversity, through whole genome genotyping, in two species of mammals on a continental scale, illuminating likely routes of post-glacial colonisation of the European continent.

4.4.2 *Apodemus sylvaticus*

The number of different genetic lineages found for *A. sylvaticus* differs between different approaches (STRUCTURE and ADMIXTURE) and also between runs of the same method with different seeds, but it is always between 4 and 7.

PCA analysis differentiated two main groups of samples: from the Italo-Balkan and the Iberian refugia (Figure 4.11). This grouping was consistent with the pattern observed on the phylogenetic tree (Figure 4.23). The highest levels of nucleotide diversity, heterozygosity and polymorphic loci found in Iberia and southern France indicate that Iberia acted as a refugium, retaining the highest diversity (first recolonisation scenario: south-north movements (Randi, 2007))(Figure 4.40). Paleontological data (Aguilar and Michaux, unpublished but cited in Michaux *et al.* (2003)) corroborates the presence of *A. sylvaticus* in the Balkans during the Quaternary glaciations.

The low levels of these parameters found on the Italo-Balkan peninsulas, therefore, could reflect, as previously suggested by Michaux *et al.* (2003), a strong bottleneck suffered by the populations there during the last ice ages. Furthermore, multiple species of *Apodemus*, including *A. flavicollis* (Michaux *et al.*, 2004), survived in the Balkan peninsula as well. When both species co-occur in the same environment, *A. sylvaticus* is usually dominated by *A. flavicollis* (Michaux *et al.*, 2005). The interspecific competition could have contributed to small population size and low diversity in *A. sylvaticus* in the Balkans and Italy.

These groups (Italo-Balkan and Iberian) split 1.5–1.6 Ma, as suggested by Michaux *et al.* (2003) based on mtDNA. Since then, and due to the presence of important geographical barriers, such as the Alps, the Italo-Balkan group has been isolated from the other main refugia used by *A. sylvaticus*, in

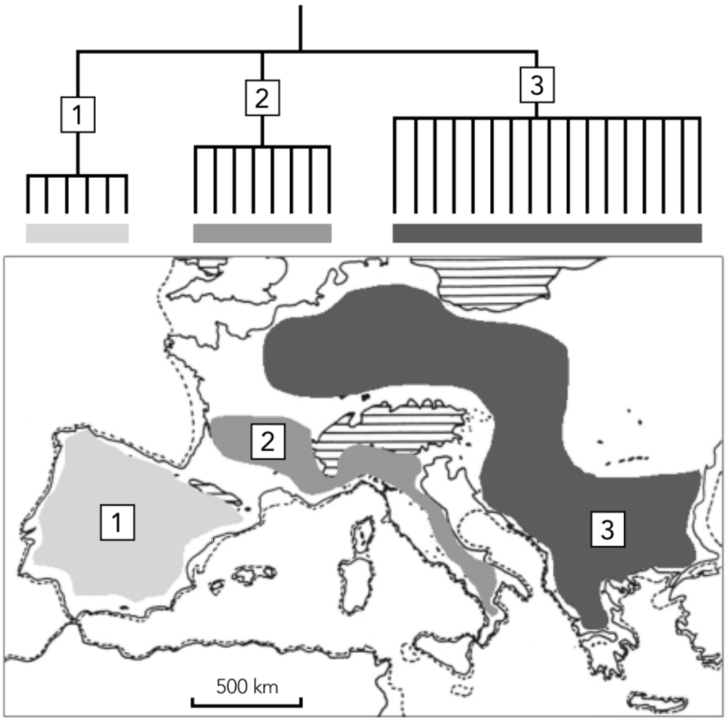


FIGURE 4.40: First recolonization scenario: Southern European peninsulas. Randi (2007)

Iberia. However, genetic exchange between both Italian and Balkan peninsulas could have occurred due to reductions on the Adriatic Sea levels during glacier periods (Michaux *et al.*, 2003). These exchange could explain the genetic similarities found in this and previous studies between the populations (Michaux *et al.*, 2003; Herman *et al.*, 2017) of both peninsulas. Overall, the pattern observed here is in agreement with two main refugia in the southern European peninsulas and the recolonisation of the European continent from Iberia (Michaux *et al.*, 2003).

This pattern, species surviving in more than one European refugia, has also been observed in other European mammals, as for example: Brown bear, *Ursus arctos* (Taberlet and Bouvet, 1994), lesser white-toothed shrew, *Crocidura suaveolens*) (Dubey *et al.*, 2006) and the red deer, *Cervus elaphus* (Zachos and Hartl, 2011)

Similarly, the distribution of *A. sylvaticus* lineages in Europe presented by Herman *et al.* (2017) is broadly similar to the lineages described here for K=3 (Figure 4.41 and Figure 4.32), although my analysis does not support K=3 as the best description of the data (the value of K that have been supported by

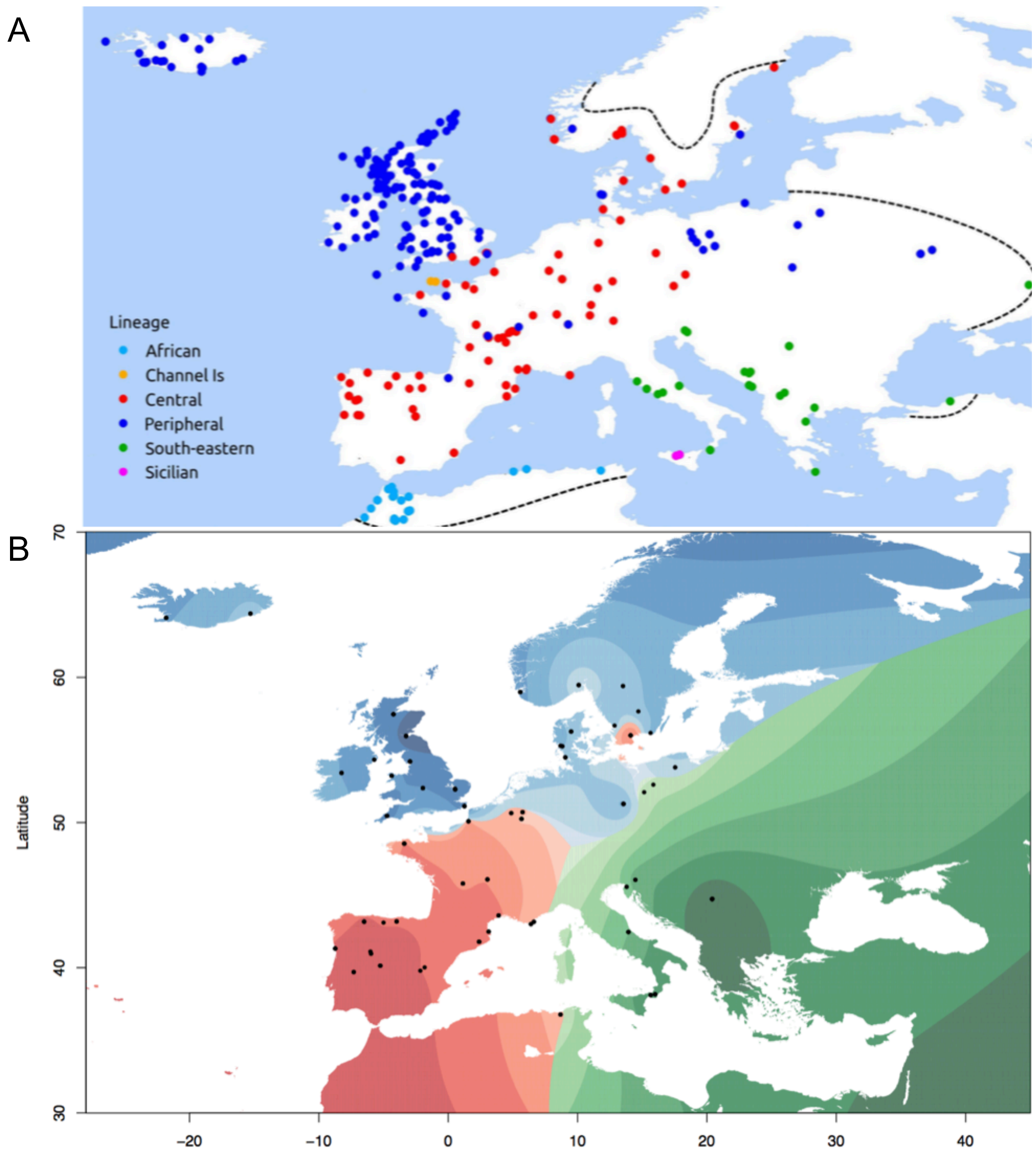


FIGURE 4.41: Comparison of groups found by Herman *et al.* (2017) using cytochrome B sequences (A) and the groups found in our analysis for K=3 (B)

ADMIXTURE, STRUCTURE and find.clusters function from Adegenet (Jombart, 2008) is 4). The main differences between both studies come from the genetic background of Scandinavian populations, that show, mainly, a mitochondrial Iberian ancestry (Herman *et al.*, 2017), whereas the nuclear genetic background belongs mainly to the northern group (Figure 4.41 and Figure 4.32). The lack of samples from Central Europe makes it difficult to observe the extent of the Iberian lineage in central Europe, but admixture between the different groups have been clearly observed in the ancestry analysis (Figures 4.17, 4.18, 4.19). A deeper sampling from Germany, Austria, Czechia, Switzerland and Netherlands would help determine the extent of the Iberian lineage and the potential origin of the northern lineage.

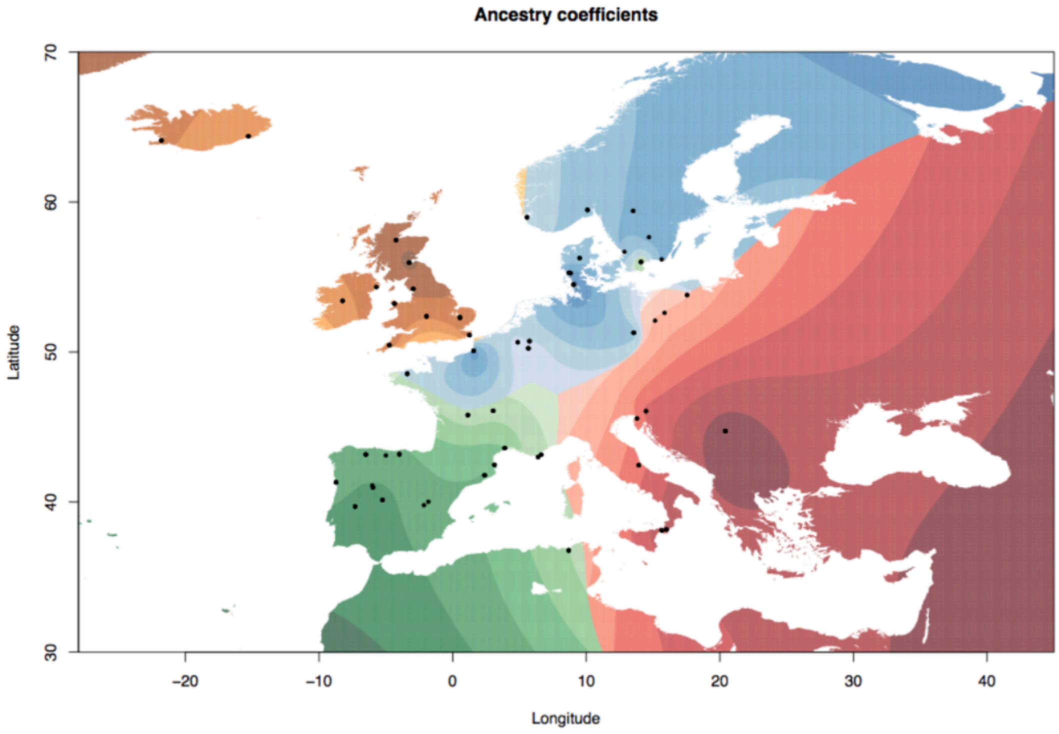


FIGURE 4.42: Distribution of *Apodemus sylvaticus* groups for K=4

K values of 2 and 3 show a clear differentiation of a group including the British Isles, Iceland and widespread through central Europe, that is not so clearly observed in PCA analysis and on the tree. The split of this group in a northern European and British isles lineages, or even in an Icelandic lineage is what it is revealed by increasing values of K. Values of K higher than 4 show new groups that do not seem to have a clear biological or phylogeographic interpretation and likely are not related to glacial refugia.

The presence of the component of the British Isles, Iceland and the western part of Norway, could be the consequence of isolation by distance processes. Populations in these regions could have been isolated after the drowning of Doggerland, 8 Ka, and could have evolved independently since then. A glacial refugium in southern England have been previously suggested for red deer, *Cervus elaphus* (Lister, 1984), however, it is unlikely for *Apodemus*, as climate simulations do not support England as a suitable environment for them during the LGM (Fløjgaard *et al.*, 2009).

4.4.2.1 The (non)existence of northern refugia and long distance movements

There are two main findings in my work that amend and clarify previous views of the phylogeographic history of *Apodemus* ((Michaux *et al.*, 2003; Michaux *et al.*, 2005; Herman *et al.*, 2017)). Firstly, new insights into the possible existence of a northern glacial refugium for *Apodemus*, postulated by Herman *et al.* (2017) in their analysis of 981 mitochondrial cytochrome b sequences.

The phylogenetic tree built in this work shows the "peripheral" or combined northern European-British Isles group as a clade within the Iberian one. The time calculated by Herman *et al.* (2017) for the most recent common ancestor of the peripheral lineage has a median value of 16,363 Ka, lower than the ones obtained for the well known refugia in Iberia (median value of 22.254 Ka) and the Italo-Balkan peninsulas (median value of 19.868 Ka). This timing situates the most recent common ancestor of this lineage after the last glacial maximum. This calibrations are not accurate enough to exclude the possibility of a northern refugia, however, my results suggest that both the northern European and the British Isles group are derived from the Iberian refugia. Alternatively, it is possible that they have originated from different rounds of colonisations, which have been isolated from the rest of the refugium for the duration of the last glaciations. Herman *et al.* (2017) suggested two different locations for this potential northern refugia: the Dordogne, in southern France, or the Carpathian region. My data could still support the idea suggested by Herman *et al.* (2017) of a northern refugia in Dordogne but definitely reject the possibility of a Carpathian refugium due to the close relationship between the Iberian and the northern lineages. A extensive sampling in Iberia and southern France could help us to understand if Iberian acted as a single refugia or as a refugium within refugia that could potentially explain the origin of the northern groups (Gomez and Lunt, 2007). Additionally, an intensive sampling of Central Europe could help us to understand the extent of the northern groups.

A calibration of the phylogenetic tree built in this project is needed in order to confirm the time of the split between the different groups found (Italo-Balkan, Iberian, British Isles-Iceland and northern Europe) and clarify if the split occurred before or after the last glaciation. This data is not yet available, but we are moving forward with this analysys, which will be completed in time for the submission of the publication. We have already used thousands of loci to estimate the divergence between the two species of *Apodemus* included in the Polish study (Chapter 2), in order to estimate the molecular clock of these species. Furthermore, *in silico* digestion of the genomes of *Peromyscus maniculatus* and *Mus musculus* have been performed and will be used to calculate the time of the split between *Mus-Apodemus*, *Mus-Peromyscus*, *Peromyscus-Apodemus* and *A. sylvaticus-A. flavicollis* as input for SNAPP programme (a package of BEAST2 (Bouckaert *et al.*, 2014)). Once the times of the split are calculated, it will be possible to more confidently reject the existence of the Carpathian refugium.

A second significant finding in my analysis is providing evidence for a long-distance movement of individuals. I have identified a southern Swedish population of *A. sylvaticus* that is more closely related to Spanish-French populations than to other Swedish populations (Figures 4.11 and 4.17). This is unexpected due to the distance between both countries and the high level of admixture between the northern and the iberian components on Central Europe. I have been considered two scenarios to explain the origin of this population.

It is possible that individuals from southern-eastern Europe spread more than 8000 years ago through Europe and crossed to Scandinavia through the emerged Doggerland. When Doggerland drowned, some individuals from the southern European clade could have been isolated in Scandinavia. After this event, the southern European group could have been replaced in northern Europe, breaking the contact between these two groups. Another possibility is that this exchange occurred later in time, with the help of humans. It is known that *Mus musculus domesticus* mice from western Europe presents a mtDNA lineage that it is restricted to Norway and northern and western areas of the British Isles. The distribution of this group has been previously linked to the influence of Norwegian vikings (Searle *et al.*, 2008). In this case, it is hypothesised that the Vikings took mice from their home region to the colonised areas, or that such movement took place in the opposite direction. Therefore, the hypothesis that the Vikings, or any other human group, could bring southern European wood mice to Scandinavia, could also apply to *Apodemus*.

DIYABC is an approach to Approximate Bayesian Computation for inference of population history that uses molecular markers to test the likelihood of different phylogeographic scenarios. When the four different scenarios explained in this chapter were tested on my data, the two models supporting a closer relationship between the Swedish population and the northern group were clearly rejected. On the other hand, DIYABC analysis appears to support the hypothesis of a natural (i.e. non-human related) movement of individuals through Doggerland. However, I note the poor estimation of the population size and timings, based on which DIYABC supports its model, and therefore these results should be treated as very preliminary. I have been performing more simulations of the different scenarios and parameters, which are expected to be completed in time for the publication. It is worth noting, however, that all the results obtained so far in the modelling are consistent with the genetic pattern observed in PCA (Figure 4.11) and ADMIXTURE analysis (Figure 4.17), showing that the Swedish population is more closely related to Iberian samples than to northern European samples.

A similarly interesting pattern detected in my data is the admixture between the northern European and the British components in a western Norwegian population. Whether this population can be considered as a source for the British populations or if British individuals have contributed, likely through human-assisted movements, to the Norwegian populations, will be modelled after an accurate estimation of the population size and timings for the Swedish model. A similar case - a connection between Scottish and Icelandic samples - has been observed by Herman *et al.* (2017).

4.4.3 *Apodemus flavicollis*

The number of genetically different lineages found for *Apodemus flavicollis* also differ between different approaches (STRUCTURE and ADMIXTURE) and also between runs of the same method with different seeds, but the range of possible K value is lower than for *Apodemus sylvaticus*: between 3 and 4. Similarly to *Apodemus sylvaticus*, two different groups have been clearly identified by PCA analysis (Figure 4.28): one in the Italo-Iberian peninsulas and another in eastern Europe.

A K value of 3 (Figure 4.43) shows how *A. flavicollis* could have colonised most of Europe from a Balkan refugium. It suggests that samples from Iberia

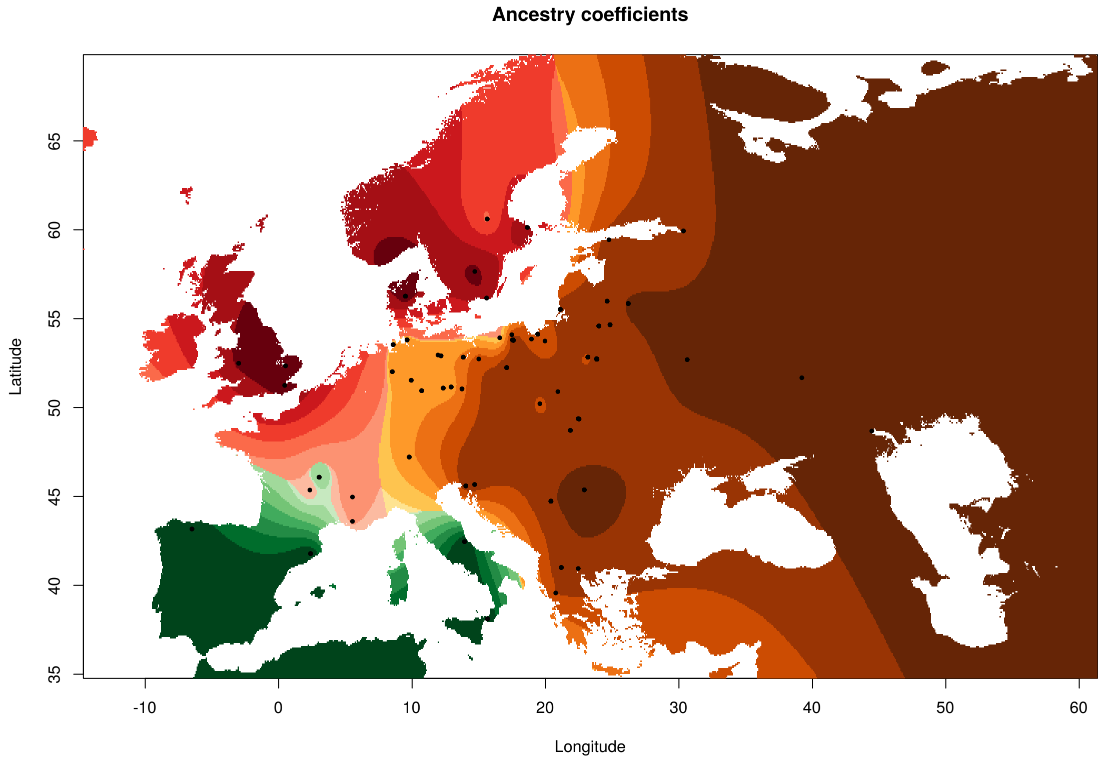


FIGURE 4.43: Distribution of *Apodemus flavicollis* groups for K=3

and Italy shared a common ancestry that was restricted to the southern European peninsulas and southern France. They are characterized by low nucleotide diversity, heterozygosity and percentage of polymorphic loci (Figure 4.35). Studies by Michaux *et al.* (2004) postulated that Iberia did not act as a refugium for *A. flavicollis*, but that the population arrived there from a Balkan refugium through rapid expansion that involved multiple bottlenecks, considerably reducing their genetic diversity. In contrast, the phylogenetic tree in this study (Figure 2.7) shows a basal split between the Italo-Iberian group and the rest of the samples, suggesting that the separation between this group and the Balkan refugia could have occurred before the end of the glaciations. These results are in agreement with the fossil record, which dated *A. flavicollis* fossils from southern Spain to 25 Ky, immediately before the Last Glacial Maximum (Fløjgaard *et al.*, 2009). However, it is important to notice that our sampling in western Europe is quite poor and could have influenced the results by, e.g., showing a higher differentiation of Iberian and Italian samples than they are in reality.

The third group, with a northern distribution, however, has not been previously identified in other studies. This group includes samples from Scandinavia and the British Isles, and reaches south to cover most of France. Previous studies on the phylogeographic structure of *A. flavicollis* (Michaux *et al.*, 2004) were focused on the southern European peninsulas as the source of current European populations and its sampling on the northern part of *A. flavicollis* distribution range was very limited. Nevertheless, most of the northern samples included in our analysis cluster within the Balkan group (K=3, Figure 4.43).

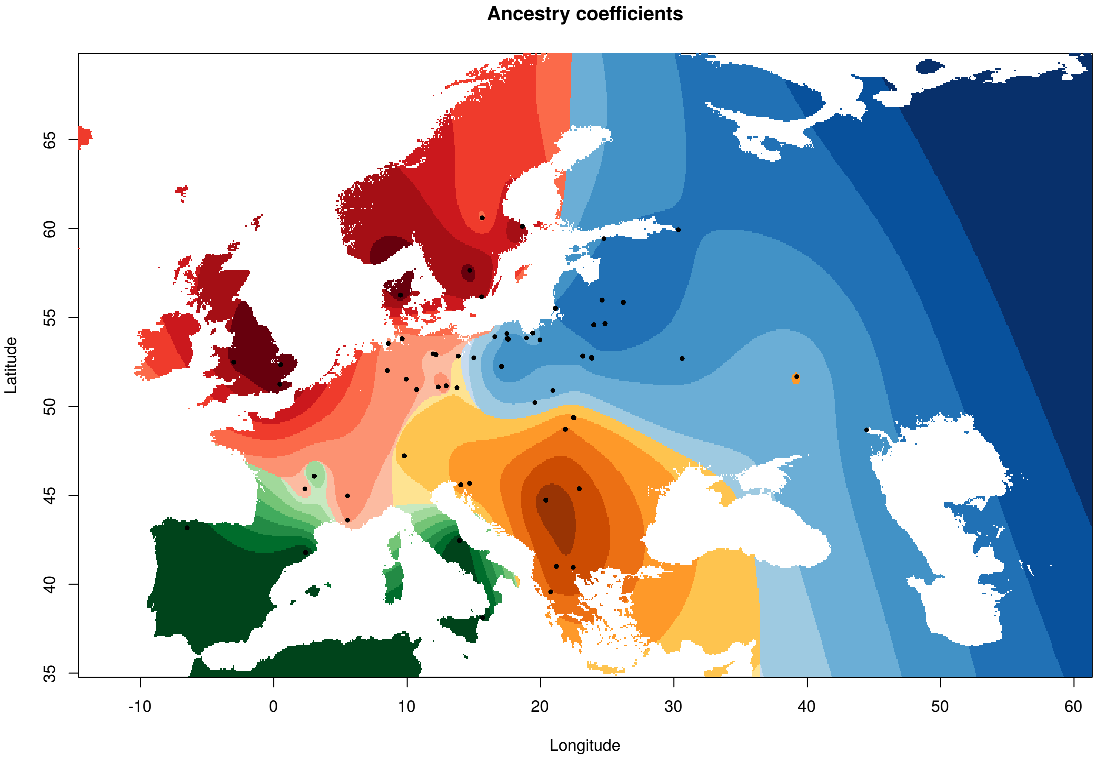


FIGURE 4.44: Distribution of *A. flavicollis* groups for K=4

A K value of 4 revealed the existence of a new lineage, distributed in central-eastern Europe, from Poland to Russia (Figure 4.44). One sample in southern Russia, however, belongs to the Balkan group, but the method used to build the map (maps function from TESS3 (Caye *et al.*, 2016) did not linked this population with the rest of the group.

The small F_{ST} differences observed between the northern, the Balkan and the eastern groups as well as the pattern observed on PCA (Figure 4.28) and on the phylogenetic tree (Figure 2.7) do indicate a similar origin for the three groups, in broad agreement with the hypothesis of Michaux *et al.* (2004), who postulated the existence of three genetic groups or subclades of *A. flavicollis* surviving in the Balkan region and recolonising most of the European continent from there. However, a Carphatian or Caucasian origin of the eastern group can not be excluded yet. A more extensive sampling of the Balkan and Carphatian region will be needed in order to clarify this possibility. Indeed, we expect to receive more samples from those regions thanks to a collaboration with Dr Barbara Tschirren from the University of Exeter.

It is also noticeable, in contrast to what we have seen in *A. sylvaticus*, that in this study we do not see any evidence supporting long-distance movement of *A. flavicollis* populations, despite covering similar area and similar number of samples. This difference can be due to their different ecological habits. *A. flavicollis* usually inhabits inside forests while *A. sylvaticus*, a more generalist species, also inhabits man-made habitats, such as urban parks, gardens, arable fields or pastures.

4.4.4 Effect of the unequal number of samples on ADMIXTURE analysis

My previous analyses (Chapter 2) have shown the effect of unequal number of samples on the estimation of population genetics parameters and ADMIXTURE results. Results obtained through the analysis of 100 permutations, including equal number of individuals, with randomisation of the individuals selected for each run, have shown similar estimation of population genetic parameters, except for the number of private alleles (Tables 2.2 and 2.4). I observed that the main differences between ADMIXTURE analyses performed either with equal or unequal number of samples affected more the selection of the optimal value of K, rather than the component distribution for the same K values (data not shown). Groups containing fewer samples or that have experienced little genetic drift are more likely to appear as a mix of other groups (Lawson *et al.*, 2018). Even when we tried to include an equal number of samples per population, it has not always been possible (Figures 4.10, 4.27) due to differences on the number of samples obtained, the quality of their DNA extraction or the quality of the sequenced data. However, as the European phylogeography analysis is not performed at the population level, but on a more wider scale, with much smaller differences between the groups compared, we expect our analysis to be immune to these effects. Additionally, in both the Polish and the European studies, I rely on the biologically relevant parameters of the data (e.g. our *a priori* knowledge of how many populations we sampled from). The less numerous groups on our phylogeographic analysis have been the groups for refugia that did not contribute to the continental populations (samples from the Italo-Balkan peninsulas in *A. sylvaticus* and Iberian samples for *A. flavicollis*. In both cases, from K=3, these groups have been clearly identified.

4.4.5 Comparative phylogeography of *A. sylvaticus* and *A. flavicollis*

4.4.5.1 SNP catalogue for species identification

A. flavicollis and *A. sylvaticus* are two sibling species that live in sympatry in the forests and fields of the European Plain. They display similar behaviour and morphology. The main differences between these two species are slightly different ecological habits. They are so similar that their identification in the field can be problematic, at least in the southern part of their distribution range. In such cases, morphometrics (Barčiová and Macholán, 2009) and cytB (Michaux *et al.*, 2001), have been used to identify the species. However, neither the morphometrics nor the cytB sequencing is straightforward, the latter due to a nuclear pseudogene of cytB that can mislead the analyses of genetic diversity and relatedness(Dubey *et al.*, 2009) .

Here, I have constructed a catalogue of 1471404 loci based on all samples in the phylogeography study that allow the differentiation of the two species. Application of this catalogue to our samples identified eight misidentified samples (four samples morphologicaly identified as *A. flavicollis* that belong to *A.sylvaticus* and four samples classified as *A. sylvatiucs* that seem to be *A flavicollis*. The position of those samples in a species-specific Principal Component Analysis (Figures 4.11 and 4.28) was unexpected, as a higher differentiation between species than between different refugia was expected. Comparing this catalogue with the catalogue built in Chapter 2, an increased power to identify species is observed. The previous catalogue has a clear bias towards *A. flavicollis*, due to the higher number of samples included from this species to build it. Now, that the number of samples from both species is larger and samples are more equally distributed between the two species, only a couple of samples from Slovenia still appeared in a central position between both species which can lead to miss-identification problems. The reasons behind this unclear position are still unknown and could suggest hybridisation between both species. In order to clearly identify the species of each sample, it would be beneficial to compare the distribution of SNPs according to their genomic position to potentially identify haplotype blocks shared between the two species. Multiple comparison of pooled samples from both species, as performed by, for example, popooluation2 (Kofler *et al.*, 2011), could reveal genomic regions that are fixed in each species. These regions can potentially be used to detect new markers that would allow an unambiguous differentiation between the two species.

Very little has been previously said about the possibility of hybridisation between the two species, but the reported similarities between *A. sylvaticus* and *A .flavicollis* in the southern part of their distribution could be interpreted as a signal of hybridisation. Hybridisation has been considered unlikely to occur, given a long divergence time - 4 Ma of independent evolution for each species. However, in the light of the results obtained for two Slovenian samples, it would be worth to analyse it formally. Firstly , I would resequence both samples, to completely eliminate any possibility of contamination between multiple samples. Only if the same results would be the same, after resequencing of the samples, an ABBA-BABA test could be performed.

The two species seem to have follow similar strategies to survive to the Pleistocene glaciations, but their main refugia have not overlapped, with *A. flavicollis* surviving mainly in the Balkan region while *A. sylvaticus* did it in Iberia. Low genetic diversity has been found for both species in the other Mediterranean peninsulas, which has been previously interpreted, along with other results, as a fast postglacial arrival, with successive strong bottlenecks, of *A. flavicollis* to the Italo-Balkan region (Michaux *et al.*, 2005), and as the consequence of a strong bottleneck during one of the last ice ages for *A. sylvaticus* (Michaux *et al.*, 2005). In order for my analysis to fully relate to those findings, it will be necessary to calculate the times of the splits between the different groups using a more complete tree (including *Mus musculus* and *Peromyscus maniculatus*) than the ones presented in this project.

Nevertheless, the position of both "alternative" refugia for both species is similar, in what could be a sign of an early separation of the populations from the traditional Mediterranean refugia, rather than a rapid expansion for *A. flavicollis*. The presence of an stable population of *A. sylvaticus* in Iberia could have limited the survival of *A. flavicollis* in the region, in the same way that an stable population of *A. flavicollis* in the Italo-Balkan region could have limited the survival of *A. sylvaticus* there, through interspecific competition for similar resources.

Northern groups have been found both in *A. sylvaticus* and *A. flavicollis*. A more extensive sampling of Europe, including multiple populations from each potential refugium would be needed to clarify the origin on those groups. Some species have shown strong population substructure within the Iberian glacial refugium itself, which could mean that Iberia was not a single refugium, but a group of them, a refugia within refugia (Gomez and Lunt, 2007). Gomez and Lunt (2007) indicates that a poor sampling of the southern Mediterranean peninsulas could fail to identify their real phylogeographic history and could lead to an erroneous inference of northern refugia. For example, a group can be inferred to be derived from the Dordogne French refugium, while it could come from an unsampled Iberian refugium, whose haplotypes are firstly sampled in southern France.

4.4.5.2 Sampling for phylogeography: present and future

Our sampling, even though it is quite extensive and includes more than 50 populations from around Europe, is still limited to fully resolve the finer details of population movements, such as the origin of the northern groups. A deeper sampling of Iberia, France, Italy, the Balkan region (including samples from Croatia, Bosnia, Albania, Bulgaria and Romania) and the Caucasus is needed in order to clarify the genetic structure of their populations before beeing able to confirm the existence of the northern refugia. Furthermore, our sampling also contain important gaps in Central Europe. More samples from France, Germany, Czechia, Hungary, Romania, Ukrania and Russia would be needed to clarify the geographical range on each group, their possible origins and admixture zones.

Nevertheless, even though I could not confirm or reject the hypothesis of a northern refugium for *A. sylvaticus*, this work clarifies the existence of this group and eliminates the possibility of the eastern origin of the northern group. In contrast, all the northern groups of *A. flavicollis* apper to have a Balkan or even eastern origin.

4.4.5.3 Long-distance genetic exchange in *A. sylvaticus* but not in *A. flavicollis*

Analysis of phylogeographic patterns in *A. sylvaticus*, but notably not in *A. flavicollis*, has shown evidence for long-term genetic exchanges between populations across Europe (Figure 4.42). At least three long distance genetic exchanges for this species have been identified: Iberia-Sweden, Norway-France and Norway-British Isles. The propensity of *A. sylvaticus* to be introduced on islands through human mediated transport has previously been suggested by Herman *et al.* (2017), who found 12 mtDNA haplotypes that were shared between areas separated by sea, even during the time that Doggerland was emerged (Figure 4.45).

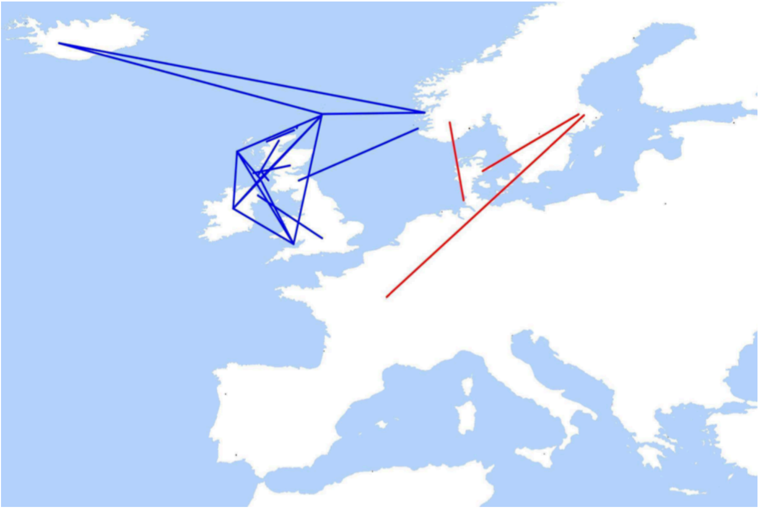


FIGURE 4.45: Connections between regions with shared haplotypes between populations of *A. sylvaticus* based on cytB data. Image from Herman *et al.* (2017)

Our findings support those of Herman *et al.* (2017) using only mitochondrial character. Human-assisted migration is the likely explanation (as in the case of Mus musculus migrations from Europe to the British Isles (Searle *et al.*, 2008), but my dataset does not allow to speculate on the cause of these connections yet. However, what is interesting is the lack of such signal for any population of *Apodemus flavicollis*.

4.5 Future directions

Even though these results are broadly in agreement with findings based on mitochondrial genes (Michaux *et al.*, 2003; Michaux *et al.*, 2004; Michaux *et al.*, 2005; Herman *et al.*, 2017), the number and genome-wide markers available in my study allowed not only a much higher reliability of the findings but also revealed previously hidden patterns of relationships between distantly related populations. While human-assisted movement is likely an explanation of these patterns, my models do not yet fully resolve the question of timings and direction that could explain the observations. The key directions for this research to enable much fuller resolution of phylogeographic patterns across the continent, are, roughly in order of decreasing importance:

- Timing of events

Timing the split between the different groups determined by my analysis is needed to determine if the different groups survived in different refugia during the Pleistocene glaciation or if they split after the end of the last ice age. My ongoing work concentrates on obtaining these calibrations using SNAPP (Bryant *et al.*, 2012) from BEAST (Suchard *et al.*, 2018) to estimate the age of the split between the main groups, using both species as well as *Apodemus speciosus*, *Peromyscus maniculatus* and *Mus musculus* sequences. *In silico* digestion of the three genomes have already been performed and the sequences are ready to be matched against the catalogue generated during the analysis. Another possibility is using Fastsimcoal2 (Excoffier *et al.*, 2013) for the same purpose. The divergence between species calculated in Chapter 2 allows us to have a more accurate molecular clock for these species and helps to properly calibrate those and therefore clarify if *Apodemus* species have survived in a northern refugium.

- Modelling long-distance genetic exchanges

The direction and the time of the movements of *Apodemus sylvaticus* have not been resolved yet. DIYABC (Cornuet *et al.*, 2014) analysis performed to determined the origin of the Swedish population have clearly determined the Southern Mediterranean origin of this population, but further analyses are needed in order to identify better estimations of the population size needed for this analysis.

- Increase sampling

Our group keeps searching and establishing new collaborations with researchers around Europe, who could provide us samples from the critical regions highlighted in my analyses. Given the long history of ecological observations of *Apodemus* in Europe, we are positive about the prospect of improving our sampling. Our recent collaborator, Dr Barabara Tschirren from the University of Exeter, has already sent us samples and contacts to researchers with samples from central Europe.

- Whole high-quality genome sequence for both species

During the development of this project, two draft genomes have been made publicly available: *A. sylvaticus*, unpublished but made available from Dr. Andrew Turner in 2015 (University of Liverpool) and *A. speciosus* (Matsunami *et al.*, 2018) (a Japanese species). We are collaborating with Dr. Steve Paterson from the University of Liverpool on the improvement of *A. sylvaticus* genome by combining the existing data with long read data obtained on the 10x Genomics platform (10x Genomics, Pleasanton, CA). Furthermore, we are also collaborating with him to sequence and annotate the *A.flavicollis* genome on the same platform.

4.5.1 Ongoing work in the BrykLab

This project and resources that our group has been working on recently will contribute to the development of *Apodemus* as a model organism for ecological and evolutionary genomic. For example, our group is currently using the protocol developed in Chapter 3 to investigate the heritability of the basal metabolic rates and torpor in a wild population of *A. flavicollis* in Białowieża, Poland and combine it with long-term observations of survival. Such studies will demonstrate the power of having a widespread, wild mammal with fully developed genomic resources available for research. In addition to these projects, the modification of the quaddRAD protocol I developed in this thesis is being further modified to allow recovery of homologous regions in highly degraded samples, allowing to incorporate museum samples into current studies on *Apodemus*.

Bibliography

Abdul-Muneer, P.M. (2014). “Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies”. In: *Genetics research international* 2014.

Adkins, R.M., Gelke, E.L., Rowe, D., and Honeycutt, R.L. (2001). “Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes”. In: *Molecular Biology and Evolution* 18.5, pp. 777–791.

Adnađević, T., Bugarski-Stanojević, V., Blagojević, J., Stamenković, G., and Vujošević, M. (2012). “Genetic differentiation in populations of the yellow-necked mouse, Apodemus flavicollis, harbouring B chromosomes in different frequencies”. In: *Population ecology* 54.4, pp. 537–548.

Alexander, D.H. and Lange, K. (2011). “Enhancements to the ADMIXTURE algorithm for individual ancestry estimation”. In: *BMC bioinformatics* 12.1, p. 246.

Alexander, D.H., Novembre, J., and Lange, K. (2009). “Fast model-based estimation of ancestry in unrelated individuals”. In: *Genome research* 19.9, pp. 1655–1664.

Alexander, M., Ho, S.Y., Molak, M., Barnett, R., Carlborg, Ö., Dorshorst, B., Honaker, C., Besnier, F.s, Wahlberg, P., Dobney, K., and Siegel, P. (2015). “Mitogenomic analysis of a 50-generation chicken pedigree reveals a rapid rate of mitochondrial evolution and evidence for paternal mtDNA inheritance”. In: *Biology letters* 11.10, p. 20150561.

Allio, R., Donega, S., Galtier, N., and Nabholz, B. (2017). “Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker". In: *Molecular biology and evolution* 34.11, pp. 2762–2772.

Alter, S.E, Munshi-South, J., and Stiassny, M.L. (2017). "Genome wide SNP data reveal cryptic phylogeographic structure and microallopatric divergence in a rapids-adapted clade of cichlids from the Congo River". In: *Molecular ecology* 26.5, pp. 1401–1419.

Andrews, K.R. and Luikart, G. (2014). "Recent novel approaches for population genomics data analysis". In: *Molecular ecology* 23.7, pp. 1661–1667.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. (2016). "Harnessing the power of RADseq for ecological and evolutionary genomics". In: *Nature Reviews Genetics* 17.2, p. 81.

Andrews, S. (2010). "Babraham bioinformatics—FastQC A quality control tool for high throughput sequence data". In: $https://www.bioinformatics.$ *babraham.ac.uk/projects/*$fastqc/$. Accessed:2018-07-15.

Avise, J.C. (2000). *Phylogeography: the history and formation of species*. Harvard university press.

Avise, J.C., Arnold, J., Ball, R.M, Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., and Saunders, N.C. (1987). "Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics". In: *Annual review of ecology and systematics* 18.1, pp. 489–522.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). "Rapid SNP discovery and genetic mapping using sequenced RAD markers". In: *PloS one* 3.10, e3376.

Barčiová, L. and Macholán, M. (2009). "Morphometric key for the discrimination of two wood mice species, Apodemus sylvaticus and A. flavicollis". In: *Acta Zoologica Academiae Scientiarum Hungaricae* 55.1, pp. 31–38.

Bartmann, S. and Gerlach, G. (2001). "Multiple paternity and similar variance in reproductive success of male and female wood mice (Apodemus sylvaticus) housed in an enclosure". In: *Ethology* 107.10, pp. 889–899.

Bartram, J., Mountjoy, E., Brooks, T., Hancock, J., Williamson, H., Wright, G., Moppett, J., Goulden, N., and Hubank, M. (2016). "Accurate sample assignment in a multiplexed, ultrasensitive, high-throughput sequencing assay for minimal residual disease". In: *The Journal of Molecular Diagnostics* 18.4, pp. 494–506.

Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D., and Blaxter, M.L. (2011). "Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism". In: *PloS one* 6.4, e19315.

Berckmoes, V., Scheirs, J., Jordaens, K., Blust, R., Backeljau, T., and Verhagen, R. (2005). "Effects of environmental pollution on microsatellite DNA diversity in wood mouse (Apodemus sylvaticus) populations". In: *Environmental toxicology and chemistry* 24.11, pp. 2898–2907.

Bernatchez, L. (2001). "The evolutionary history of brown trout (Salmo trutta L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation". In: *Evolution* 55.2, pp. 351–379.

Bhagwat, S.A. and Willis, K.J. (2008). "Species persistence in northerly glacial refugia of Europe: a matter of chance or biogeographical traits?" In: *Journal of Biogeography* 35.3, pp. 464–482.

Bilton, D.T., Mirol, P.M., Mascheretti, S., Fredga, K., Zima, J., and Searle, J.B. (1998). "Mediterranean Europe as an area of endemism for small mammals rather than a source for northwards postglacial colonization". In: *Proceedings of the Royal Society of London B: Biological Sciences* 265.1402, pp. 1219–1226.

Blanco-Bercial, L. and Bucklin, A. (2016). "New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus". In: *Molecular ecology* 25.7, pp. 1566–1580.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, pp. 2114–2120.

Boore, J.L. (1999). "Animal mitochondrial genomes". In: *Nucleic acids research* 27.8, pp. 1767–1780.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 10.4, e1003537.

Britten, R.J. (2002). "Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels". In: *Proceedings of the National Academy of Sciences* 99.21, pp. 13633–13635.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., and RoyChoudhury, A. (2012). "Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis". In: *Molecular biology and evolution* 29.8, pp. 1917–1932.

Bugarski-Stanojević, V., Blagojević, J., Adnađević, T., Jojić, V., and Vujošević, M. (2008). "Molecular phylogeny and distribution of three Apodemus species (Muridae, Rodentia) in Serbia". In: *Journal of Zoological Systematics and Evolutionary Research* 46.3, pp. 278–286.

Bugarski-Stanojević, V., Blagojević, J., Adnađević, T., Jovanović, V., and Vujošević, M. (2013). "Identification of the sibling species Apodemus sylvaticus and Apodemus flavicollis (Rodentia, Muridae)—Comparison of molecular methods". In: *Zoologischer Anzeiger-A Journal of Comparative Zoology* 252.4, pp. 579–587.

Bugarski-Stanojević, V., Stamenković, G., Blagojević, J., Liehr, T., Kosyakova, N., Rajičić, M., and Vujošević, M. (2016). "Exploring supernumeraries-a new marker for screening of B-chromosomes presence in the yellow necked mouse Apodemus flavicollis". In: *PloS one* 11.8, e0160946.

Butet, A. and Delettre, Y.R. (2011). "Diet differentiation between European arvicoline and murine rodents". In: *Acta theriologica* 56.4, p. 297.

Cao, Y.N., Wang, I.J., Chen, L.Y., Ding, Y.Q., Liu, L.X., and Qiu, Y.X (2018). "Inferring spatial patterns and drivers of population divergence of Neolitsea sericea (Lauraceae), based on molecular phylogeography and landscape genomics". In: *Molecular phylogenetics and evolution* 126, pp. 162–172.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). "Stacks: building and genotyping loci de novo from short-read sequences". In: *G3: Genes, genomes, genetics* 1.3, pp. 171–182.

Caye, K., Deist, T.M., Martins, H., Michel, O., and François, O. (2016). "TESS3: fast inference of spatial population structure and genome scans for selection". In: *Molecular Ecology Resources* 16.2, pp. 540–548.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1, p. 7.

Chatterjee, A., Rodger, E.J., Stockwell, P.A., Weeks, R.J., and Morison, I.M. (2012). "Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries". In: *BioMed Research International* 2012.

Cheddadi, R., Vendramin, G.G., Litt, T.s, François, L., Kageyama, M., Lorentz, S., Laurent, J.M., De Beaulieu, J.L., Sadori, L., Jost, A., and Lunt, D. (2006).

"Imprints of glacial refugia in the modern genetic diversity of Pinus sylvestris". In: *Global Ecology and Biogeography* 15.3, pp. 271–282.

Cherepanov, A., Yildirim, E., and Vries, S. de (2001). "Joining of short DNA oligonucleotides with base pair mismatches by T4 DNA ligase". In: *The Journal of Biochemistry* 129.1, pp. 61–68.

Chevret, P., Veryrunes, F., and Britton-Davidian, J. (2005). "Molecular phylogeny of the genus Mus (Rodentia: Murinae) based on mitochondrial and nuclear data". In: *Biological Journal of the Linnean Society* 84.3, pp. 417–427.

Chial, H. (2008). "DNA Fingerprinting Using Amplified Fragment Length Polymorphisms (AFLP)". In:

Chistiakov, D.A., Hellemans, B., and Volckaert, F.A. (2006). "Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics". In: *Aquaculture* 255.1-4, pp. 1–29.

Chong, Z., Ruan, J., and Wu, C.I. (2012). "Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads". In: *Bioinformatics* 28.21, pp. 2732–2737.

Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., and McCabe, A.M. (2009). "The last glacial maximum". In: *science* 325.5941, pp. 710–714.

Coles, B.J. (2000). "Doggerland: the cultural dynamics of a shifting coastline". In: *Geological Society, London, Special Publications* 175.1, pp. 393–401.

Combosch, D.J. and Vollmer, S.V. (2015). "Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific Pocillopora corals". In: *Molecular phylogenetics and evolution* 88, pp. 154–162.

Consuegra, S., De Leániz, C.G., Serdio, A., Morales, M. G., Straus, L.G., Knox, D., and Verspoor, E (2002). "Mitochondrial DNA variation in Pleistocene and modern Atlantic salmon from the Iberian glacial refugium". In: *Molecular ecology* 11.10, pp. 2037–2048.

Cornuet, J.M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.M., and Estoup, A. (2014). "DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data". In: *Bioinformatics* 30.8, pp. 1187–1189.

Costa, R., Pereira, G., Garrido, I., Sousa, M.M. Tavares-de, and Espinosa, F. (2016). "Comparison of RAPD, ISSR, and AFLP molecular markers to reveal and classify orchardgrass (Dactylis glomerata L.) germplasm variations". In: *PloS one* 11.4, e0152972.

Creer, S., Thorpe, R.S., Malhotra, A., Chou, W.H., and Stenson, A.G. (2004). "The utility of AFLPs for supporting mitochondrial DNA phylogeographical analyses in the Taiwanese bamboo viper, Trimeresurus stejnegeri". In: *Journal of Evolutionary Biology* 17.1, pp. 100–107.

Cromie, G.A., Hyma, K.E., Ludlow, C.L., Garmendia-Torres, C., Gilbert, T.L., May, P., Huang, A.A., Dudley, A.M., and Fay, J.C. (2013). "Genomic sequence diversity and population structure of Saccharomyces cerevisiae assessed by RAD-seq". In: *G3: Genes, Genomes, Genetics* 3.12, pp. 2163–2171.

Cull, B., Vaux, A.G., Ottowell, L.J., Gillingham, E.L., and Medlock, J.M. (2017). "Tick infestation of small mammals in an English woodland". In: *Journal of Vector Ecology* 42.1, pp. 74–83.

Czarnomska, S.D., Niedziałkowska, M., Borowik, T., and Jędrzejewska, B. (2018). "Regional and local patterns of genetic variation and structure in yellow-necked mice - the roles of geographic distance, population abundance, and winter severity". In: *Ecology and Evolution*, 00:1–16.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., and McVean, G. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158.

DeConto, R.M. and Pollard, D. (2003). "Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO 2". In: *Nature* 421.6920, p. 245.

Dubey, S., Zaitsev, M., Cosson, J.F: and Abdukadier A.and Vogel, P. (2006). "Pliocene and Pleistocene diversification and multiple refugia in a Eurasian shrew (Crocidura suaveolens group)". In: *Molecular Phylogenetics and Evolution* 38.3, pp. 635–647.

Dubey, S., Michaux, J., Brünner, H., Hutterer, R., and Vogel, P. (2009). "False phylogenies on wood mice due to cryptic cytochrome-b pseudogene". In: *Molecular Phylogenetics and Evolution* 50.3, pp. 633–641.

Eaton, D.A. (2014). "PyRAD: assembly of de novo RADseq loci for phylogenetic analyses". In: *Bioinformatics* 30.13, pp. 1844–1849.

Eaton, D.A. and Ree, R.H. (2013). "Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae)". In: *Systematic biology* 62.5, pp. 689–706.

Edwards, S.V., Shultz, A.J., and Campbell-Staton, S.C. (2015). "Next-generation sequencing and the expanding domain of phylogeography". In: *Folia Zoologica* 64.3, pp. 187–206.

Eggert, L.S., Rasner, C.A., and Woodruff, D.S. (2002). "The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers". In: *Proceedings of the Royal Society of London B: Biological Sciences* 269.1504, pp. 1993–2006.

Ellegren, H. (2004). "Microsatellites: simple sequences with complex evolution". In: *Nature reviews genetics* 5.6, p. 435.

Excoffier, L., Dupanloup, I.e, Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). "Robust demographic inference from genomic and SNP data". In: *PLoS genetics* 9.10, e1003905.

Eyison, H.M. and Kıvanç, E. (2016). "Karyotype of Apodemus flavicollis in Giresun, Turkey". In: *Journal of Entomology and Zoology Studies* 4.2, pp. 497–499.

Fabre, P.H., Hautier, L., Dimitrov, D., and Douzery, E.J. (2012). "A glimpse on the pattern of rodent diversification: a phylogenetic approach". In: *BMC evolutionary biology* 12.1, p. 88.

Faircloth, B.C. and Glenn, T.C. (2012). "Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels". In: *PloS one* 7.8, e42543.

Feliner, G.N. (2011). "Southern European glacial refugia: a tale of tales". In: *Taxon* 60.2, pp. 365–372.

Fernández, M.E., Goszczynski, D.E., Lirón, J.P., Villegas-Castagnasso, E.E., Carino, M.H., Ripoli, M.V., Rogberg-Muñoz, A., Posik, D.M., Peral-García, P., and Giovambattista, G. (2013). "Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd". In: *Genetics and molecular biology* 36.2, pp. 185–191.

Fernández, R., Schubert, M., Vargas-Velázquez, A.M., Brownlow, A., Víkingsson, G.A., Siebert, U., Jensen, L.F., Øien, N., Wall, D., Rogan, E., and Mikkelsen, B. (2016). "A genomewide catalogue of single nucleotide polymorphisms in white-beaked and Atlantic white-sided dolphins". In: *Molecular ecology resources* 16.1, pp. 266–276.

Fischer, M.C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K.K., Holderegger, R., and Widmer, A. (2017). "Estimating genomic diversity and population differentiation–an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri". In: *BMC genomics* 18.1, p. 69.

Fleischmann, R.D., Adams, M. D, White, O., Clayton, R. A, Kirkness, E. F, Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223, pp. 496–512.

Fløjgaard, C., Normand, S., Skov, F., and Svenning, J.C. (2009). "Ice age distributions of European small mammals: insights from species distribution modelling". In: *Journal of Biogeography* 36.6, pp. 1152–1163.

Franchini, P., Monné Parera, D., Kautt, A.F., and Meyer, A. (2017). "quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage". In: *Molecular Ecology* 26.10, pp. 2783–2795.

Gandolfi, A., Crestanello, B., Fagotti, A., Simoncelli, F., Chiesa, S., Girardi, M., Giovagnoli, E., Marangoni, C., Di Rosa, I., and Lucentini, L. (2017). "New evidences of mitochondrial DNA heteroplasmy by putative paternal leakage between the rock partridge (Alectoris graeca) and the chukar partridge (Alectoris chukar)". In: *PloS one* 12.1, e0170507.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C.e, Pudlo, P., Cornuet, J.M, and Estoup, A. (2013). "The effect of RAD allele dropout on the estimation of genetic variation within and between populations". In: *Molecular Ecology* 22.11, pp. 3165–3178.

Gibson, A., Gowri-Shankar, V., Higgs, P.G., and Rattray, M. (2004). "A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods". In: *Molecular Biology and Evolution* 22.2, pp. 251–264.

Gomez, A. and Lunt, D.H. (2007). "Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula". In: *Phylogeography of southern European refugia*. Springer, pp. 155–188.

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., and Houston, R.D. (2014). "Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing". In: *Bmc Genomics* 15.1, p. 166.

Gortat, T., Gryczyńska-Siemiątkowska, A., Rutkowski, R., Kozakiewicz, A., Mikoszewski, A., and Kozakiewicz, M. (2010). "Landscape pattern and genetic structure of a yellow-necked mouse Apodemus flavicollis population in north-eastern Poland". In: *Acta Theriologica* 55.2, pp. 109–121.

Gustafsson, C.M., Falkenberg, M., and Larsson, N.G. (2016). "Maintenance and expression of mammalian mitochondrial DNA". In: *Annual review of biochemistry* 85, pp. 133–160.

Hammerman, N.M., Rivera-Vicens, R.E., Galaska, M.P., Weil, E., Appledoorn, R.S., Alfaro, M., and Schizas, N.V. (2018). "Population connectivity of the plating coral Agaricia lamarcki from southwest Puerto Rico". In: *Coral Reefs* 37.1, pp. 183–191.

Harff, J., Bailey, G.N., and Lüth, F. (2015). "Geology and archaeology: submerged landscapes of the continental shelf: an introduction". In: *Geological Society, London, Special Publications* 411, SP411–13.

Hauser, L., Baird, M., Hilborn, R.A.Y., Seeb, L.W., and Seeb, J.E. (2011). "An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population". In: *Molecular ecology resources* 11, pp. 150–161.

Hayden, E.C. (2014). "Is the $1,000 genome for real?" In: *Nature News*.

Henning, F., Lee, H.J, Franchini, P., and Meyer, A. (2014). "Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping". In: *Molecular ecology* 23.21, pp. 5224–5240.

Herman, J.S., Jóhannesdóttir, F., Jones, E.P., McDevitt, A.D., Michaux, J.R., White, T.A., Wójcik, J.M., and Searle, J.B. (2017). "Post-glacial colonization of Europe by the wood mouse, Apodemus sylvaticus: evidence of a northern refugium and dispersal with humans". In: *Biological Journal of the Linnean Society* 120.2, pp. 313–332.

Hewitt, G.M. (1999). "Post-glacial re-colonization of European biota". In: *Biological journal of the Linnean Society* 68.1-2, pp. 87–112.

Hewitt, G.M. (2000). "The genetic legacy of the Quaternary ice ages". In: *Nature* 405.6789, p. 907.

Hewitt, G.M. (2011). "Quaternary phylogeography: the roots of hybrid zones". In: *Genetica* 139.5, pp. 617–638.

Heyman, P., Mele, R.V., Smajlovic, L., Dobly, A., Cochez, C., and Vandenvelde, C. (2009). "Association between habitat and prevalence of hantavirus infections in bank voles (Myodes glareolus) and wood mice (Apodemus sylvaticus)". In: *Vector-Borne and Zoonotic Diseases* 9.2, pp. 141–146.

Hipp, A.L., Eaton, D.A., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P.S. (2014). "A framework phylogeny of the American oak clade based on sequenced RAD data". In: *PLoS One* 9.4, e93975.

Hodel, R.G., Chen, S., Payton, A.C., McDaniel, S.F., Soltis, P., and Soltis, D.E. (2017). "Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering". In: *Scientific reports* 7.1, p. 17598.

Hoffberg, S.L., Kieran, T.J., Catchen, J.M., Devault, A., Faircloth, B.C., Mauricio, R., and Glenn, T.C. (2016). "RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data". In: *Molecular ecology resources* 16.5, pp. 1264–1278.

Hohenlohe, P.A., Day, M.D., Amish, S.J., Miller, M.R., Kamps-Hughes, N., Boyer, M.C., Muhlfeld, C.C., Allendorf, F.W., Johnson, E.A., and Luikart, G. (2013). "Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing". In: *Molecular ecology* 22.11, pp. 3002–3013.

Hou, Y., Nowak, M.D., Mirré, V., Bjorå, C.S., Brochmann, C., and Popp, M. (2015). "Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus Diapensia (Diapensiaceae)". In: *PloS one* 10.10, e0140175.

Hughes, J. (2011). "Sequence-manipulation". In: *https://github.com/josephhughes/Sequence-manipulation*.

Illumina (2018a). "Diagnosing and preventing flow cell overclustering on the Miseq systemp". In: *https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf*. Accessed: 20th October 2018.

Illumina (2018b). "Effects of Index Misassignment on Multiplexing and Downstream Analysis". In: *https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf*. Accessed: 20th October 2018.

IUCN (2017). "The IUCN red list of threatened species". In: *http://www.iucnredlist.org*. Accessed: 23th July 2018.

Jaarola, M. and Searle, J.B. (2002). "Phylogeography of field voles (Microtus agrestis) in Eurasia inferred from mitochondrial DNA sequences". In: *Molecular ecology* 11.12, pp. 2613–2621.

Jeffries, D.L., Copp, G.H., Lawson Handley, L., Olsén, K.H., Sayer, C.D., and Hänfling, B. (2016). "Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, Carassius carassius, L." In: *Molecular ecology* 25.13, pp. 2997–3018.

Jojić, V., Blagojević, J., and Vujošević, M. (2011). "B chromosomes and cranial variability in yellow-necked field mice (Apodemus flavicollis)". In: *Journal of Mammalogy* 92.2, pp. 396–406.

Jojić, V., Bugarski-Stanojević, V., Blagojević, J., and Vujošević, M. (2014). "Discrimination of the sibling species Apodemus flavicollis and A. sylvaticus (Rodentia, Muridae)". In: *Zoologischer Anzeiger-A Journal of Comparative Zoology* 253.4, pp. 261–269.

Jombart, T. (2008). "adegenet: a R package for the multivariate analysis of genetic markers". In: *Bioinformatics* 24.11, pp. 1403–1405.

Juškaitis, R. (2002). "Spatial distribution of the yellow-necked mouse (Apodemus flavicollis) in large forest areas and its relation with seed crop of forest trees". In: *Mammalian Biology-Zeitschrift für Säugetierkunde* 67.4, pp. 206–211.

Khrunin, A.V., Khokhrin, D.V., Filippova, I.N., Esko T.and Nelis, M., Bebyakova, N.A., Bolotova, N.L., Klovins, J., Nikitina-Zake, L., Rehnström, K., and Ripatti, S. (2013). "A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe". In: *PloS one* 8.3, e58552.

Kleef, H.L. and Wijsman, H.J. (2015). "Mast, mice and pine marten (Martes martes): the pine marten's reproductive response to wood mouse (Apodemus sylvaticus) fluctuations in the Netherlands". In: *Lutra* 58, pp. 23–33.

Klempa, B., Schmidt, H.A., Ulrich, R., Kaluz, S., Labuda, M., Meisel, H., Hjelle, B., and Krüger, D.H. (2003). "Genetic interaction between distinct Dobrava hantavirus subtypes in Apodemus agrarius and A. flavicollis in nature". In: *Journal of virology* 77.1, pp. 804–809.

Knowles, L.L., Massatti, R., He, Q., Olson, L.E., and Lanier, H.C. (2016). "Quantifying the similarity between genes and geography across Alaska's alpine small mammals". In: *Journal of Biogeography* 43.7, pp. 1464–1476.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., and Schlötterer, C. (2011). "PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals". In: *PloS one* 6.1, e15925.

Kolodziej, M., Melgies, A., Joniec-Wiechetek, J., Michalski, A., Nowakowska, A., Pitucha, G., and Niemcewicz, M. (2018). "First molecular characterization of Dobrava-Belgrade virus found in Apodemus flavicollis in Poland". In: *Annals of Agricultural and Environmental Medicine* 25.2, pp. 368–373.

Korn, H. (1986). "Changes in home range size during growth and maturation of the wood mouse (Apodemus sylvaticus) and the bank vole (Clethrionomys glareolus)". In: *Oecologia* 68.4, pp. 623–628.

Koskinen, M.T., Nilsson, J., Veselov, A.J., Potutkin, A.G., Ranta, E., and Primmer, C.R. (2002). "Microsatellite data resolve phylogeographic patterns in European grayling, Thymallus thymallus, Salmonidae". In: *Heredity* 88.5, p. 391.

Kotlík, P., Deffontaine, V., Mascheretti, S., Zima, J., Michaux, J.R., and Searle, J.B. (2006). "A northern glacial refugium for bank voles (Clethrionomys glareolus)". In: *Proceedings of the National Academy of Sciences* 103.40, pp. 14860–14864.

Kozłowski, J., Konarzewski, M., and Gawelczyk, A.T. (2003). "Cell size as a link between noncoding DNA and metabolic rate scaling". In: *Proceedings of the National Academy of Sciences* 100.24, pp. 14080–14085.

Ladoukakis, E.D. and Zouros, E. (2017). "Evolution and inheritance of animal mitochondrial DNA: rules and exceptions". In: *Journal of Biological Research-Thessaloniki* 24.1, p. 2.

Lambertini, C., Gustafsson, M.H.G., Frydenberg, J., Lissner, J., Speranza, M., and Brix, H. (2006). "A phylogeographic study of the cosmopolitan genus Phragmites (Poaceae) based on AFLPs". In: *Plant Systematics and Evolution* 258.3-4, pp. 161–182.

Lanier, H.C., Massatti, R., He, Q., Olson, L.E., and Knowles, L.L. (2015). "Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genomic variation in Collared Pikas (Ochotona collaris)". In: *Molecular Ecology* 24.14, pp. 3688–3705.

Lawson, D.J., Van Dorp, L., and Falush, D. (2018). "A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots". In: *Nature communications* 9.1, p. 3258.

Leach, M. (1990). *Mice of the British Isles*. Shire Publications.

Lee, T.H., Guo, H., Wang, X., Kim, C., and Paterson, A.H. (2014). "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data". In: *BMC genomics* 15.1, p. 162.

Li, W.H., Tanimura, M., and Sharp, P.M. (1987). “An evaluation of the molecular clock hypothesis using mammalian DNA sequences”. In: *Journal of molecular evolution* 25.4, pp. 330–342.

Lippold, S., Matzke, N.J., Reissmann, M., and Hofreiter, M. (2011). “Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication”. In: *BMC evolutionary biology* 11.1, p. 328.

Lister, AM (1984). “Evolutionary and ecological origins of British deer”. In: *Proceedings of the Royal Society of Edinburgh, Section B: Biological Sciences* 82.4, pp. 205–229.

Liu, X., Wei, F., Li, M., Jiang, X., Feng, Z., and Hu, J. (2004). “Molecular phylogeny and taxonomy of wood mice (genus Apodemus Kaup, 1829) based on complete mtDNA cytochrome b sequences, with emphasis on Chinese species”. In: *Molecular Phylogenetics and Evolution* 33.1, pp. 1–15.

Llewellyn, M.S., Miles, M.A., Carrasco, H.J., Lewis, M.D., Yeo, M., Vargas, J., Torrico, F., Diosque, P., Valente, V., Valente, S.A., and Gaunt, M.W. (2009). “Genome-scale multilocus microsatellite typing of Trypanosoma cruzi discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection”. In: *PLoS pathogens* 5.5, e1000410.

Luka, V. and Riegert, J. (2018). “Apodemus mice as the main prey that determines reproductive output of tawny owl (Strix aluco) in Central Europe”. In: *Population Ecology* 60.3, pp. 237–249.

Lunt, D.H., Ibrahim, K.M., and Hewitt, G.M. (1998). “mtDNA phylogeography and postglacial patterns of subdivision in the meadow grasshopper Chorthippus parallelus”. In: *Heredity* 80.5, p. 633.

Luo, S., Valencia, C.A., Zhang, J., Lee, N.C., Slone J.and Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., Chen, S.M., Chien, Y.H., Hwu, W.L., Fan, P.C., Wong, L.J., Atwal, P. S., and Huang, T. (2018). “Biparental Inheritance of Mitochondrial DNA in Humans”. In: *Proceedings of the National Academy of Sciences*. ISSN: 0027-8424.

Macdonald, D.W. and Tattersall, F. (2001). *Britain’s Mammals: The Challenge for Conservation*. People’s Trust for Endangered Species.

Machordom, A., Araujo, R., Toledo, C., Zouros, E., and Ladoukakis, E.D. (2015). “Female-dependent transmission of paternal mtDNA is a shared feature of bivalve species with doubly uniparental inheritance (DUI) of mitochondrial DNA”. In: *Journal of Zoological Systematics and Evolutionary Research* 53.3, pp. 200–204.

MacHugh, D.E., Shriver, M.D., Loftus, R.T., Cunningham, P., and Bradley, D.G. (1997). "Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus)". In: *Genetics* 146.3, pp. 1071–1086.

Maciak, S., Bonda-Ostaszewska, E., Czarnołęski, M., Konarzewski, M., and Kozłowski, J. (2014). "Mice divergently selected for high and low basal metabolic rates evolved different cell size and organ mass". In: *Journal of evolutionary biology* 27.3, pp. 478–487.

Makova, K.D., Patton, J.C., Krysanov, E.Y, Chesser, R.K., and Baker, R.J. (1998). "Microsatellite markers in wood mouse and striped field mouse (genus Apodemus)". In: *Molecular Ecology* 7.2, pp. 247–248.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet. journal* 17.1, pp–10.

Martiniaková, M., Omelka, R., Jancová, ., Stawarz, R., and Formicki, G. (2010). "Heavy metal content in the femora of yellow-necked mouse (Apodemus flavicollis) and wood mouse (Apodemus sylvaticus) from different types of polluted environment in Slovakia". In: *Environmental monitoring and assessment* 171.1-4, pp. 651–660.

Martínez-Nieto, M.I., Segarra-Moragues, J.G., Merlo, E., Martínez-Hernández, F., and Mota, J.F. (2013). "Genetic diversity, genetic structure and phylogeography of the Iberian endemic Gypsophila struthium (Caryophyllaceae) as revealed by AFLP and plastid DNA sequences: connecting habitat fragmentation and diversification". In: *Botanical Journal of the Linnean Society* 173.4, pp. 654–675.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D., and Emerson, B.C. (2015). "Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference". In: *Molecular Ecology Resources* 15.1, pp. 28–41.

Matić, R., Vukićević-Radić, O., Stamenković, S., and Kataranovski, D. (2007). "Can large home ranges be due to social dominance in Apodemus flavicollis?" In: *Archives of Biological Sciences* 59.4, pp. 61–62.

Matsunami, M., Endo, D., Saitou, N., Suzuki, H., and Onuma, M. (2018). "Draft genome sequence of Japanese wood mouse, Apodemus speciosus". In: *Data in Brief* 16, pp. 43–46.

Maxam, A.M. and Gilbert, W. (1977). "A new method for sequencing DNA". In: *Proceedings of the National Academy of Sciences* 74.2, pp. 560–564.

Meyer, M. and Kircher, M. (2010). "Illumina sequencing library preparation for highly multiplexed target capture and sequencing". In: 2010.6, pdb–prot5448.

Miah, G., Rafii, M., Ismail, M., Puteh, A., Rahim, H., Islam, K., and Latif, M. (2013). "A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance". In: *International journal of molecular sciences* 14.11, pp. 22499–22528.

Michaux, J.R., Kinet, S., Filippucci, M.G., Libois, R., Besnard, A., and Catzeflis, F. (2001). "Molecular identification of three sympatric species of wood mice (Apodemus sylvaticus, A. flavicollis, A. alpicola) in western Europe (Muridae: Rodentia)". In: *Molecular Ecology Notes* 1.4, pp. 260–263.

Michaux, J.R., Chevret, P., Filippucci, M.G., and Macholan, M. (2002). "Phylogeny of the genus Apodemus with a special emphasis on the subgenus Sylvaemus using the nuclear IRBP gene and two mitochondrial markers: cytochrome b and 12S rRNA". In: *Molecular phylogenetics and evolution* 23.2, pp. 123–136.

Michaux, J.R., Magnanou, E., Paradis, E., Nieberding, C., and Libois, R. (2003). "Mitochondrial phylogeography of the woodmouse (Apodemus sylvaticus) in the Western Palearctic region". In: *Molecular Ecology* 12.3, pp. 685–697.

Michaux, J.R., Libois, R., Paradis, E., and Filippucci, M.G. (2004). "Phylogeographic history of the yellow-necked fieldmouse (Apodemus flavicollis) in Europe and in the Near and Middle East". In: *Molecular phylogenetics and evolution* 32.3, pp. 788–798.

Michaux, J.R., Libois, R., and Filippucci, M.G. (2005). "So close and so different: comparative phylogeography of two small mammal species, the Yellow-necked fieldmouse (Apodemus flavicollis) and the Woodmouse (Apodemus sylvaticus) in the Western Palearctic region". In: *Heredity* 94.1, p. 52.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. (2007). "Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers". In: *Genome research* 17.2, pp. 240–248.

Mlera, L. and Bloom, M.E. (2018). "The role of mammalian reservoir hosts in tick-borne flavivirus biology". In: *Frontiers in Cellular and Infection Microbiology* 8, p. 298.

Montgomery, S.S.J. and Montgomery, W.I. (1990). “Intrapopulation variation in the diet of the wood mouse Apodemus sylvaticus”. In: *Journal of Zoology* 222.4, pp. 641–651.

Montgomery, W.I, Provan, J., McCabe, A.M., and Yalden, D.W. (2014). “Origin of British and Irish mammals: disparate post-glacial colonisation and species introductions”. In: *Quaternary Science Reviews* 98, pp. 144–165.

Moraes-Barros, N. and Morgante, J.S. (2007). “A simple protocol for the extraction and sequence analysis of DNA from study skin of museum collections”. In: *Genetics and Molecular Biology* 30.4, pp. 1181–1185.

Moura, A.E., Kenny, J.G., Chaudhuri, R., Hughes, M.A., J. Welch, A., Reisinger, R.R., Bruyn, P.N. de, Dahlheim, M.E., Hall, N., and Hoelzel, A.R. (2014). “Population genomics of the killer whale indicates ecotype evolution in sympatry involving both selection and drift”. In: *Molecular Ecology* 23.21, pp. 5179–5192.

Netušil, J., Žákovská, A., Vostal, K., Norek, A., and Stanko, M. (2013). “The occurrence of Borrelia burgdorferi sensu lato in certain ectoparasites (Mesostigmata, Siphonaptera) of Apodemus flavicollis and Myodes glareolus in chosen localities in the Czech Republic”. In: *Acta parasitologica* 58.3, pp. 337–341.

Nieberding, C., Libois, R., Douady, C.J., Morand, S., and Michaux, J.R. (2005). “Phylogeography of a nematode (Heligmosomoides polygyrus) in the western Palearctic region: persistence of northern cryptic populations during ice ages?” In: *Molecular Ecology* 14.3, pp. 765–779.

Nunes, M.D., Dolezal, M., and Schlötterer, C. (2013). “Extensive paternal mt DNA leakage in natural populations of Drosophila melanogaster”. In: *Molecular ecology* 22.8, pp. 2106–2117.

Pala, M., Olivieri, A., Achilli, A., Accetturo, M., Metspalu, E., Reidla, M., Tamm, E., Karmin, M., Reisberg, T., Kashani, B.H., Perego, U.A., Carossa, V., Gandini, F., Pereira, J.B., Soares, P., Angerhofer, N., Rychkov, S., Al-Zahery, N., Carelli, V, Sanatti, M.H., Houshmand, M., Hatina, J., Macaulay, V., Pereira, L., Woodward, S.R., Davies, W., Gamble, C., Baird, D., Semino, O., Villems, R., Torroni, A., and Richards, M.B. (2012). “Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia”. In: *The American journal of human genetics* 90.5, pp. 915–924.

Panculescu-Gatej, R.I., Sirbu, A., Dinu, S., Waldstrom, M., Heyman, P., Murariu, D., Petrescu, A., Szmal, C., Oprisan, G., Lundkvist, Å., and Ceianu, C.S. (2014). “Dobrava virus carried by the yellow-necked field mouse Apodemus flavicollis, causing hemorrhagic fever with renal syndrome in Romania”. In: *Vector-Borne and Zoonotic Diseases* 14.5, pp. 358–364.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S.C., Boisselier, M.C., and Samadi, S. (2015). “Use of RAD sequencing for delimiting species”. In: *Heredity* 114.5, p. 450.

Papa, A., Rogozi, E., Velo, E., Papadimitriou, E., and Bino, S. (2016). “Genetic detection of hantaviruses in rodents, Albania”. In: *Journal of medical virology* 88.8, pp. 1309–1313.

Paris, J.R., Stevens, J.R., and Catchen, J.M. (2017). “Lost in parameter space: a road map for stacks”. In: *Methods in Ecology and Evolution*.

Patton, H., Hubbard, A., Andreassen, K., Auriac, A., Whitehouse, P.L., Stroeven, A.P., Shackleton, C., Winsborrow, M., Heyman, J., and Hall, A.M. (2017). “Deglaciation of the Eurasian ice sheet complex”. In: *Quaternary Science Reviews* 169, pp. 148–172.

Pegadaraju, V., Nipper, R., Hulke B.and Qi, L., and Schultz, Q. (2013). “De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach”. In: *BMC genomics* 14.1, p. 556.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). “Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species”. In: *PloS one* 7.5, e37135.

Piñeiro, A. and Barja, I. (2011). “Trophic strategy of the wildcat Felis silvestris in relation to seasonal variation in the availability and vulnerability to capture of Apodemus mice”. In: *Mammalian Biology-Zeitschrift für Säugetierkunde* 76.3, pp. 302–307.

Pleines, T. and Blattner, F.R. (2008). “Phylogeographic implications of an AFLP phylogeny of the American diploid Hordeum species (Poaceae: Triticeae)”. In: *Taxon* 57.3, pp. 875–881.

Poland, J.A. and Rife, T.W. (2012). “Genotyping-by-sequencing for plant breeding and genetics”. In: *The Plant Genome* 5.3, pp. 92–102.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.L. (2012). “Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach”. In: *PloS one* 7.2, e32253.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). "Genotyping errors: causes, consequences and solutions". In: *Nature Reviews Genetics* 6.11, p. 847.

Pucek, Z., Jędrzejewski, W., Jędrzejewska, B., and Pucek, M. (1993). "Rodent population dynamics in a primeval deciduous forest (Białowieża National Park) in relation to weather, seed crop, and predation". In: *Acta Theriologica* 38.2, pp. 199–232.

Pudlo, P., Marin, J.M, Estoup, A., Cornuet, J.M, Gautier, M., and Robert, C.P. (2015). "Reliable ABC model choice via random forests". In: *Bioinformatics* 32.6, pp. 859–866.

Puechmaille, S.J (2016). "The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem". In: *Molecular Ecology Resources* 16.3, pp. 608–627.

Purcell, S. and Chang, C.C. (2018). "PLINK 1.9 package". In: *https://www.cog-genomics.org/plink/1.9/*. Accessed: 17th July 2018.

Puritz, J.B., Hollenbeck, C.M., and Gold, J.R. (2014a). "dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms". In: *PeerJ* 2, e431.

Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D.I., and Bird, C.E. (2014b). "Demystifying the RAD fad". In: *Molecular Ecology* 23.24, pp. 5937–5942.

Raj, A., Stephens, M., and Pritchard, J.K. (2014). "Variational inference of population structure in large SNP datasets". In: *Genetics*, genetics–114.

Rajičić, M., Romanenko, S.A., Karamysheva, T.V., Blagojević, J., Adnađević, T., Budinski, I., Bogdanov, A.S., Trifonov, V.A., Rubtsov, N.B., and Vujošević, M. (2017). "The origin of B chromosomes in yellow-necked mice (Apodemus flavicollis)—Break rules but keep playing the game". In: *PloS one* 12.3, e0172704.

Randi, E. (2007). "Phylogeography of south European mammals". In: *Phylogeography of southern European refugia*. Springer, pp. 101–126.

Randolph, S.E., Miklisova, D., Lysy, J., Rogers, D.J., and Labuda, M. (1999). "Incidence from coincidence: patterns of tick infestations on rodents facilitate transmission of tick-borne encephalitis virus". In: *Parasitology* 118.2, pp. 177–186.

Rašić, G., Filipović, I., Weeks, A.R., and Hoffmann, A.A. (2014). "Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, Aedes aegypti". In: *BMC genomics* 15.1, p. 275.

Ratkiewicz, M. and Borkowska, A. (2006). "Genetic structure is influenced by environmental barriers: empirical evidence from the common voleMicrotus arvalis populations". In: *Acta theriologica* 51.4, pp. 337–344.

Raymo, M.E. and Ruddiman, W.F. (1992). "Tectonic forcing of late Cenozoic climate". In: *Nature* 359.6391, p. 117.

Razkin, O., Sonet, G., Breugelmans, K., Madeira, M.J., Gómez-Moliner, B.J, and Backeljau, T. (2016). "Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex Pyramidula: the power of RADseq data". In: *Molecular phylogenetics and evolution* 101, pp. 267–278.

Reitzel, A.M., Herrera, S., Layden, M.J., Martindale, M.Q., and Shank, T.M. (2013). "Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics". In: *Molecular ecology* 22.11, pp. 2953–2970.

Richter, D., Schlee, D.B., and Matuschka, F.R. (2011). "Reservoir competence of various rodents for the Lyme disease spirochete Borrelia spielmanii". In: *Applied and environmental microbiology*.

Rico, A., Kindlmann, P., and Sedláček, F. (2009). "Can the barrier effect of highways cause genetic subdivision in small mammals?" In: *Acta theriologica* 54.4, pp. 297–310.

Rodriguez-Ezpeleta, N., Álvarez, P., and Irigoien, X. (2017). "Genetic Diversity and Connectivity in Maurolicus muelleri in the Bay of Biscay Inferred from Thousands of SNP Markers". In: *Frontiers in genetics* 8, p. 195.

Sanger, F. and Coulson, A.R. (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of molecular biology* 94.3, pp. 441–448.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12, pp. 5463–5467.

Santucci, F., Emerson, B.C., and Hewitt, G.M. (1998). "Mitochondrial DNA phylogeography of European hedgehogs". In: *Molecular Ecology* 7.9, pp. 1163–1172.

Schlitter, D., van der Straeten, E., Amori, G., Hutterer, R., Krystufek, B., Yigit, N., and Mitsain, G. (2010). "Apodemus sylvaticus:The IUCN Red List of Threatened Species". In:

Schmitt, T. (2007). “Molecular biogeography of Europe: Pleistocene cycles and postglacial trends”. In: *Frontiers in zoology* 4.1, p. 11.

Schweyen, H., Rozenberg, A., and Leese, F. (2014). “Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters”. In: *The Biological Bulletin* 227.2, pp. 146–160.

Schönswetter, P., Suda, J., Popp, M., Weiss-Schneeweiss, H., and Brochmann, C. (2007). “Circumpolar phylogeography of Juncus biglumis (Juncaceae) inferred from AFLP fingerprints, cpDNA sequences, nuclear DNA content and chromosome numbers”. In: *Molecular Phylogenetics and Evolution* 42.1, pp. 92–103.

Searle, J.B., Jones, C.S., Gündüz, İ., Scascitelli, M., Jones, E.P., Herman, J.S., Rambau, R.V., Noble, L.R., Berry, R.J., Giménez, M.D., and Jóhannesdóttir, F (2008). “Of mice and (Viking?) men: phylogeography of British and Irish house mice”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 276.1655, pp. 201–207.

Seddon, J.M., Santucci, F., Reeve, N.J., and Hewitt, G.M. (2001). “DNA footprints of European hedgehogs, Erinaceus europaeus and E. concolor: Pleistocene refugia, postglacial expansion and colonization routes”. In: *Molecular Ecology* 10.9, pp. 2187–2198.

Shafer, A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W., and Wolf, J.B. (2017). “Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference”. In: *Methods in Ecology and Evolution* 8.8, pp. 907–917.

Sharma, R., Goossens, B., Kun-Rodrigues, C., Teixeira, T., Othman, N., Boone, J.Q., Jue, N.K., Obergfell, C., O’Neill, R.J., and Chikhi, L. (2012). “Two different high throughput sequencing approaches identify thousands of de novo genomic markers for the genetically depleted Bornean elephant”. In: *PLoS One* 7.11, e49533.

Shitara, H.i, Hayashi, J.I, Takahama, S., Kaneda, H., and Yonekawa, H. (1998). “Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage”. In: *Genetics* 148.2, pp. 851–857.

Shultz, A.J., Baker, A.J., Hill, G.E., Nolan, P.M., and Edwards, S.V. (2016). “SNP s across time and space: population genomic signatures of founder events and epizootics in the House Finch (Haemorhous mexicanus)”. In: *Ecology and evolution* 6.20, pp. 7475–7489.

Sommer, R.S. and Nadachowski, A. (2006). "Glacial refugia of mammals in Europe: evidence from fossil records". In: *Mammal Review* 36.4, pp. 251–265.

Sovic, M.G., Fries, A.C., and Gibbs, H.L. (2015). "AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data". In: *Molecular ecology resources* 15.5, pp. 1163–1171.

Stapley J.and Reger, J., Feulner, P.G., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., and Slate, J. (2010). "Adaptation genomics: the next generation". In: *Trends in ecology & evolution* 25.12, pp. 705–712.

Štefka, J. and Hypša, V. (2008). "Host specificity and genealogy of the louse Polyplax serrata on field mice, Apodemus species: A case of parasite duplication or colonisation?" In: *International journal for parasitology* 38.6, pp. 731–741.

Stewart, J.R. and Lister, A.M. (2001). "Cryptic northern refugia and the origins of the modern biota". In: *Trends in Ecology & Evolution* 16.11, pp. 608–613.

Stojak, J., McDevitt, A.D., Herman, J.S., Searle, J.B., and Wójcik, J.M. (2015). "Post-glacial colonization of eastern Europe from the Carpathian refugium: evidence from mitochondrial DNA of the common vole Microtus arvalis". In: *Biological Journal of the Linnean Society* 115.4, pp. 927–939.

Stojak, J., McDevitt, A.D., Herman, J.S., Kryštufek, B., Uhlíková, J., Purger, J.J., Lavrenchenko, L.A., Searle, J.B., and Wójcik, J.M. (2016). "Between the Balkans and the Baltic: phylogeography of a common vole mitochondrial DNA lineage limited to Central Europe". In: *PloS one* 11.12, e0168621.

Stone, K.D. and Cook, J.A. (2000). "Phylogeography of black bears (Ursus americanus) of the Pacific Northwest". In: *Canadian Journal of Zoology* 78.7, pp. 1218–1223.

Stuart, J.A. and Brown, M.F. (2006). "Mitochondrial DNA maintenance and bioenergetics". In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1757.2, pp. 79–89.

Suchan, T., Pitteloud, C., Gerasimova, N.S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M., Ronikier, M., and Alvarez, N. (2016). "Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens". In: *PLoS One* 11.3, e0151651.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). “Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10”. In: *Virus Evolution* 4.1, vey016.

Sunyer, P., Muñoz, A., Bonal, R., and Espelta, J.M. (2013). “The ecology of seed dispersal by small rodents: a role for predator and conspecific scents”. In: *Functional ecology* 27.6, pp. 1313–1321.

Suzuki, H., Filippucci, M.G., Chelomina, G.N., Sato, J.J., Serizawa, K., and Nevo, E. (2008). “A biogeographic view of Apodemus in Asia and Europe inferred from nuclear and mitochondrial gene sequences”. In: *Biochemical Genetics* 46.5-6, p. 329.

Taberlet, P. and Bouvet, J. (1994). “Mitochondrial DNA polymorphism, phylogeography, and conservation genetics of the brown bear Ursus arctos in Europe”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255.1344, pp. 195–200.

Taberlet, P., Fumagalli, L., Wust-Saucy, A.G., and Cosson, J.F. (1998). “Comparative phylogeography and postglacial colonization routes in Europe”. In: *Molecular ecology* 7.4, pp. 453–464.

Theologidis, I., Fodelianakis, S., Gaspar, M.B., and Zouros, E. (2008). “Doubly uniparental inheritance (DUI) of mitochondrial DNA in Donax trunculus (Bivalvia: Donacidae) and the problem of its sporadic detection in Bivalvia”. In: *Evolution: International Journal of Organic Evolution* 62.4, pp. 959–970.

Tin, M.M.Y., Rheindt, F.E., Cros, E., and Mikheyev, A.S. (2015). “Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy”. In: *Molecular Ecology Resources* 15.2, pp. 329–336.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). “Ten years of next-generation sequencing technology”. In: *Trends in genetics* 30.9, pp. 418–426.

Vandenberghe, J., Lowe, J., Coope, R., Litt, T., and Züller, L. (2004). “Climatic and environmental variability in the mid-latitude Europe sector during the last interglacial-glacial cycle”. In: *Past Climate Variability through Europe and Africa*. Springer, pp. 393–416.

Velickovic, M. (2007). “Measures of the developmental stability, body size and body condition in the black-striped mouse (Apodemus agrarius) as indicators of a disturbed environment in northern Serbia”. In: *Belgian Journal of Zoology* 137.2, p. 147.

Vujošević, M., Jojić, V., Bugarski-Stanojević, V., and Blagojević, J. (2007). "Habitat quality and B chromosomes in the yellow-necked mouse Apodemus flavicollis". In: *Italian Journal of Zoology* 74.4, pp. 313–316.

Vujošvić, M. (1992). "B-chromosome polymorphism in Apodemus flavicollis (Rodentia, Mammalia) during five years". In: *Caryologia* 45.3-4, pp. 347–352.

Wang, Z.n, Baker, A.J., Hill, G.E., and Edwards, S.V. (2003). "Reconciling actual and inferred population histories in the house finch (Carpodacus mexicanus) by AFLP analysis". In: *Evolution* 57.12, pp. 2852–2864.

Web of Science (2018). In: *https://wcs.webofknowledge.com/RA/analyze.do?product=UA&SID=C1t7XGK66QWnUGfvEx6&field=PY_PublicationYear_PublicationYear_en&yearSort=true*. Accessed:2018-07-23.

Welc-falęciak, R., Bajer, A., Behnke, J.M., and Siński, E. (2010). "The ecology of Bartonella spp. infections in two rodent communities in the Mazury Lake District region of Poland". In: *Parasitology* 137.7, pp. 1069–1077.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wolff, J.N., Nafisinia, M., Sutovsky, P., and Ballard, J.W. (2013). "Paternal transmission of mitochondrial DNA as an integral part of mitochondrial inheritance in metapopulations of Drosophila simulans". In: *Heredity* 110.1, p. 57.

Wu, D.Y. and Wallace, R.B. (1989). "Specificity of the nick-closing activity of bacteriophage T4 DNA ligase". In: *Gene* 76.2, pp. 245–254.

Wu, S., Wu, W., Zhang, F., Ye, J., Ni, X., Sun, J., Edwards, S.V., Meng, J., and Organ, C.L. (2012). "Molecular and paleontological evidence for a post-Cretaceous origin of rodents". In: *PloS one* 7.10, e46445.

Wójcik, J.M., Kawałko, A., Marková, S., Searle, J.B., and Kotlík, P. (2010). "Phylogeographic signatures of northward post-glacial colonization from high-latitude refugia: a case study of bank voles using museum specimens". In: *Journal of Zoology* 281.4, pp. 249–262.

Yu, Z., Wei, Z., Kong, X., and Shi, W. (2008). "Complete mitochondrial DNA sequence of oyster Crassostrea hongkongensis-a case of Tandem duplication-random loss for genome rearrangement in Crassostrea?" In: *Bmc Genomics* 9.1, p. 477.

Zachos, F.E. and Hartl, G.B. (2011). "Phylogeography, population genetics and conservation of the European red deer Cervus elaphus". In: *Mammal Review* 41.2, pp. 138–150.

Zima, J., Piálek, J., and Macholán, M. (2003). “Possible heterotic effects of B chromosomes on body mass in a population of Apodemus flavicollis”. In: *Canadian Journal of Zoology* 81.8, pp. 1312–1317.

Chapter A

Appendix Chapter 2

A.1 Samples information:

TABLE A.1: Sample ID, coordinates and environmental information for each one of the samples. Bory = Bory Tucholskie, Bial= Białowieża, Hack= Haćki, Kadz=Kadzidło

ID	Species	Location	Latitude	Longitud	Environment
D04	A. flavicollis	Bory	17.58	53.77	mesic pine forest
E04	A. flavicollis	Bory	17.56	53.81	mesic pine forest
F04	A. flavicollis	Bory	17.55	53.79	mesic pine forest
G04	A. flavicollis	Bory	17.58	53.78	dry pine forest
H04	A. flavicollis	Bory	17.58	53.77	mesic pine forest
A05	A. flavicollis	Bory	17.58	53.77	mesic pine forest
B05	A. flavicollis	Bory	17.58	53.77	sedge meadow
C05	A. flavicollis	Bory	17.58	53.77	mesic pine forest
D05	A. flavicollis	Bory	17.58	53.77	mesic pine forest
E05	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
F05	A. flavicollis	Bory	17.58	53.77	mesic pine forest
G05	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
H05	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
G12	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
H12	A. flavicollis	Bory	17.56	53.81	alder foorest at lake
C06	A. flavicollis	Bory	17.58	53.78	reeds at lake
D06	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
E06	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest

Sample	Species	Location	Long	Lat	Environment
F06	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
H06	A. flavicollis	Bory	17.56	53.81	mesic pine forest
A07	A. flavicollis	Bory	17.56	53.81	alder foorest at lake
B07	A. flavicollis	Bory	17.58	53.78	mesic pine forest
C07	A. flavicollis	Bory	17.51	53.80	oak-lime-hornbeam forest
A04	A. flavicollis	Bial	23.85	52.71	cultivated meadow
B04	A. flavicollis	Bial	23.85	52.71	cultivated meadow
C04	A. flavicollis	Bial	23.85	52.71	cultivated meadow
F08	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
G08	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
H08	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
A09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
B09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
C09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
D09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
E09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
F09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
G09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
H09	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
A10	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
B10	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
C10	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
D10	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
E10	A. flavicollis	Bial	23.83	52.72	oak-lime-hornbeam forest
F10	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
G10	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
H10	A. flavicollis	Bial	23.82	52.74	oak-lime-hornbeam forest
A11	A. flavicollis	Bial	23.82	52.75	oak-lime-hornbeam forest
B11	A. flavicollis	Bial	23.82	52.70	sedge meadow
C11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
D11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
E11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
F11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
G11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
H11	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
A12	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest

Sample	Species	Location	Long	Lat	Environment
B12	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
C12	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
D12	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
E12	A. flavicollis	Bial	23.85	52.72	oak-lime-hornbeam forest
C02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
D02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
E02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
F02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
G02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
H02	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
A03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
B03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
C03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
D03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
E03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
F03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
G03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
H03	A. flavicollis	Hack	23.17	52.83	xerothermic meadow
D07	A. sylvaticus	Bory	17.54	53.79	dry pine forest
E07	A. sylvaticus	Bory	17.56	53.79	reeds at lake
F07	A. sylvaticus	Bory	17.54	53.79	dry pine forest
G07	A. sylvaticus	Bory	17.55	53.79	mesic pine forest
H07	A. sylvaticus	Bory	17.54	53.79	dry pine forest
A08	A. sylvaticus	Kadz	21.37	53.20	dry pine forest
B08	A. sylvaticus	Kadz	21.37	53.20	dry pine forest
C08	A. sylvaticus	Kadz	21.37	53.20	dry pine forest
D08	A. sylvaticus	Kadz	21.37	53.20	dry pine forest
E08	A. sylvaticus	Kadz	21.37	53.20	dry pine forest

A.2 Barcodes used and demultiplexing results

TABLE A.2: Barcodes used and demutiplexing results.

Barcode	Filename	Total	NoRadTag	LowQuality	Retained
TATTCGCAT	D01	1754624	823059	322	802448
CCTTGCCATT	B02	5616138	2656224	1136	2524650
GGTATA	C02	2497516	1178292	412	1112807
TCTTGG	D02	1925168	926463	359	862961
GGTGT	E02	1772988	836727	301	802310
GGATA	F02	2103232	990036	342	950445
CTAAGCA	G02	2396754	1118652	420	1107411
ATTAT	H02	3592492	1672167	603	1654087
GCGCTCA	A03	1701066	796055	316	766110
ACTGCGAT	B03	2859122	1379563	513	1244319
TTCGTT	C03	2522570	1203310	467	1127262
ATATAA	D03	1448256	675350	261	664835
TGGCAACAGA	E03	1907170	896741	415	854373
CTCGTCG	F03	1424136	661282	253	647835
GCCTACCT	G03	1316424	631751	267	579215
CACCA	H03	4119158	1904665	717	1918252
AATTAG	A04	3353928	1576668	531	1528167
GGAACGA	B04	2714032	1268460	499	1237675
ACTGCT	C04	1519814	732180	279	676595
TGCTT	D04	3337318	1598963	538	1516482
GCAAGCCAT	E04	2272530	1077974	436	1028556
CGCACCAATT	F04	1328064	629634	257	597209
CTCGCGG	G04	2843128	1352618	497	1300936
AACTGG	H04	1773388	851270	311	799274
ATGAGCAA	A05	3543298	1701957	692	1580365
CTTGA	B05	2280988	1099255	413	1016552
GCGTCCT	C05	3835930	1834408	674	1724309
ACCAGGA	D05	3081248	1488008	581	1378175
CCACTCA	E05	2003682	940201	332	919846
TCACGGAAG	F05	889424	420138	187	407176
TATCA	G05	1212906	593550	171	545872
TAGCCAA	H05	1794800	838457	312	836413

Barcode	Filename	Total	NoRadTag	LowQuality	Retained
GGTGCACATT	C06	1784198	845365	349	798706
CTCTCGCAT	D06	1495486	710272	290	675749
CAGAGGT	E06	1827948	891153	317	810051
GCGTACAAT	F06	1083614	509870	219	494520
ACGCGCG	G06	1490100	697737	247	686381
GTCGCCT	H06	2562952	1219598	434	1168312
AATAACCAA	A07	2750168	1290585	509	1254852
AATGAACGA	B07	2023934	970073	414	904065
ATGGCAA	C07	2897680	1386103	501	1307711
GAAGCA	D07	4523918	2130884	804	2088074
AACGTGCCT	E07	3561580	1678483	705	1636000
CCTCG	F07	4775646	2243939	795	2212819
CTCAT	G07	2816492	1342690	453	1290660
ACGGTACT	H07	1538254	721448	267	710639
GCGCCG	A08	1581588	751487	306	715924
CAAGT	B08	2362354	1126323	386	1076036
GGAGTCAAG	C08	1931910	921128	340	866712
TGAAT	D08	2004632	977907	329	903723
CATAT	E08	2845620	1348639	463	1306750
GTGACACAT	F08	1793344	840977	320	798641
TATGT	G08	1912488	889366	326	892469
TGCAGA	H08	1587072	744356	247	728265
CATCTGCCG	A09	1927106	894811	400	865134
GGACAG	B09	2391890	1139966	395	1084539
ATCTGT	C09	4006790	1882256	717	1829175
AAGACGCT	D09	2083594	1008686	376	912137
GAATGCAATA	E09	1673516	809802	316	720740
TAGCAG	F09	1611016	772260	268	720206
CTTAG	G09	1236082	639672	195	503807
TTATTACAT	H09	903066	480838	156	346810
GCCAACAAGA	A10	2280156	1095405	452	1002024
TGCCGCAT	B10	4328430	2092480	779	1906904
CGTGTCA	C10	2174200	1069375	366	944630
CAACCACACA	D10	1994002	989286	367	844802
GCTCCGA	E10	2269544	1072489	435	1027809
CGTTCA	F10	2396728	1134028	402	1063780

Barcode	Filename	Total	NoRadTag	LowQuality	Retained
CATCACAAG	G10	1130460	547376	213	482972
TCCAG	H10	1134466	543467	172	500967
AACTGAAG	A11	2060310	972421	383	912133
GATTCA	B11	1559246	726131	226	712556
CAAGCCAATT	C11	2759210	1374653	491	1161521
TTGCGCT	D11	2013912	958443	341	907508
CGCAGACACT	E11	1773052	884978	334	742494
TGTGGA	F11	1638142	778419	287	738748
TGGATA	G11	2001520	982084	343	878640
ATAGCGT	H11	1929208	896385	353	888463
CCATAGA	A12	4910032	2303281	844	2177377
GGCACGCAT	B12	5959610	2968959	1184	2473214
ATTAACAATT	C12	1040872	496149	188	452788
CAATA	D12	2431454	1185404	393	1053547
TAGTCCAT	E12	1326952	615885	252	615685
CGTGACCT	F12	1389748	657705	274	614681
CTTCAGA	G12	9249428	4445105	1627	4157586
ATCTGCAACA	H12	3268002	1557924	623	1457762
	total	**206744014**	**98568584**	**36987**	**92741120**
	min	**889424**	**420138**	**156**	**346810**
	max	**9249428**	**4445105**	**1627**	**4157586**
	average	**2404000.16**	**1146146.32**	**430.08**	**1078385.11**
	stdev	**1282551.13**	**613923.28**	**234.71**	**575871.53**
	median	**2009272**	**975164**	**366.5**	**905786.5**

A.3 Estimation of the best parameters for the combined dataset.

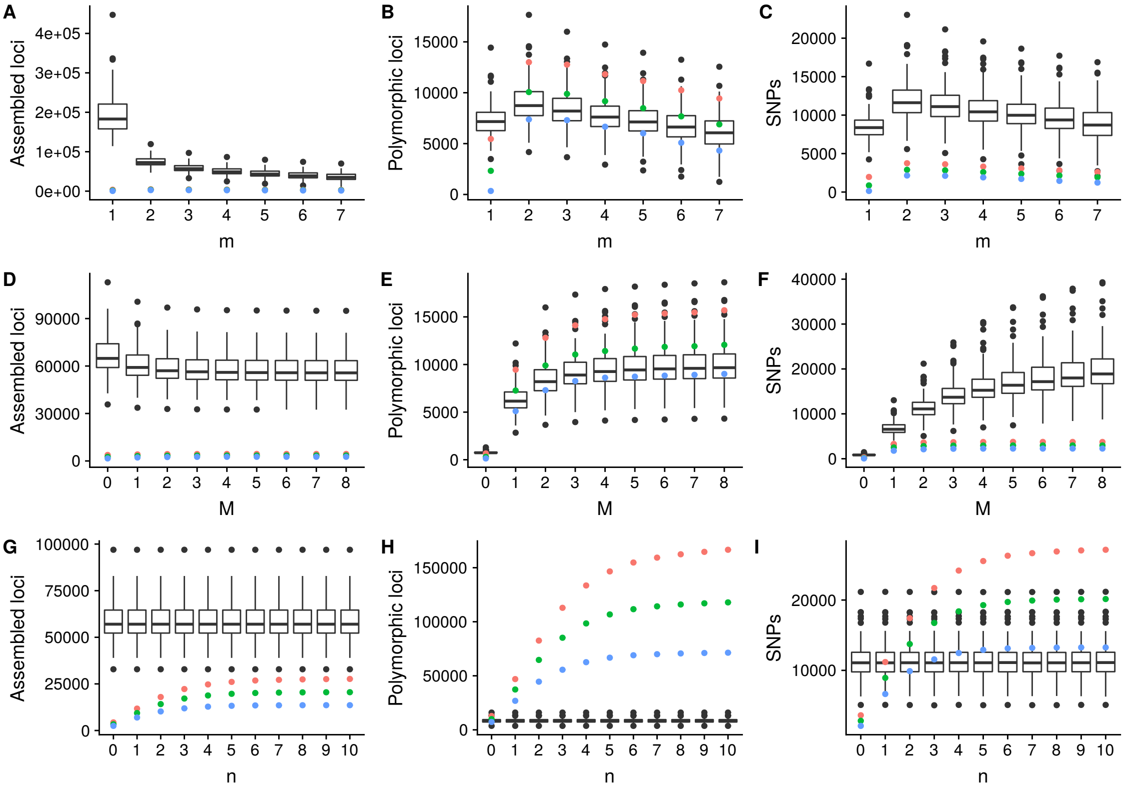


FIGURE A.1: Selection of the optimal parameters for the combined dataset. Number of assembled loci, polymorphic loci and SNPs for iterating values of m, M and n parameters. Blue circles represent data found in at least 40% of the population, green circles in the 60% and red circles in the 80%.Results for the combined dataset including samples from *Apodemus flavicollis* and *Apodemus sylvaticus*. Data used to build the graph can be found on github: *https://github.com/Marisa89/ddRADseq_poland/blob/master/Tables/Apodemus/Table_selection_best_parameters_Apodemus.xlsx*

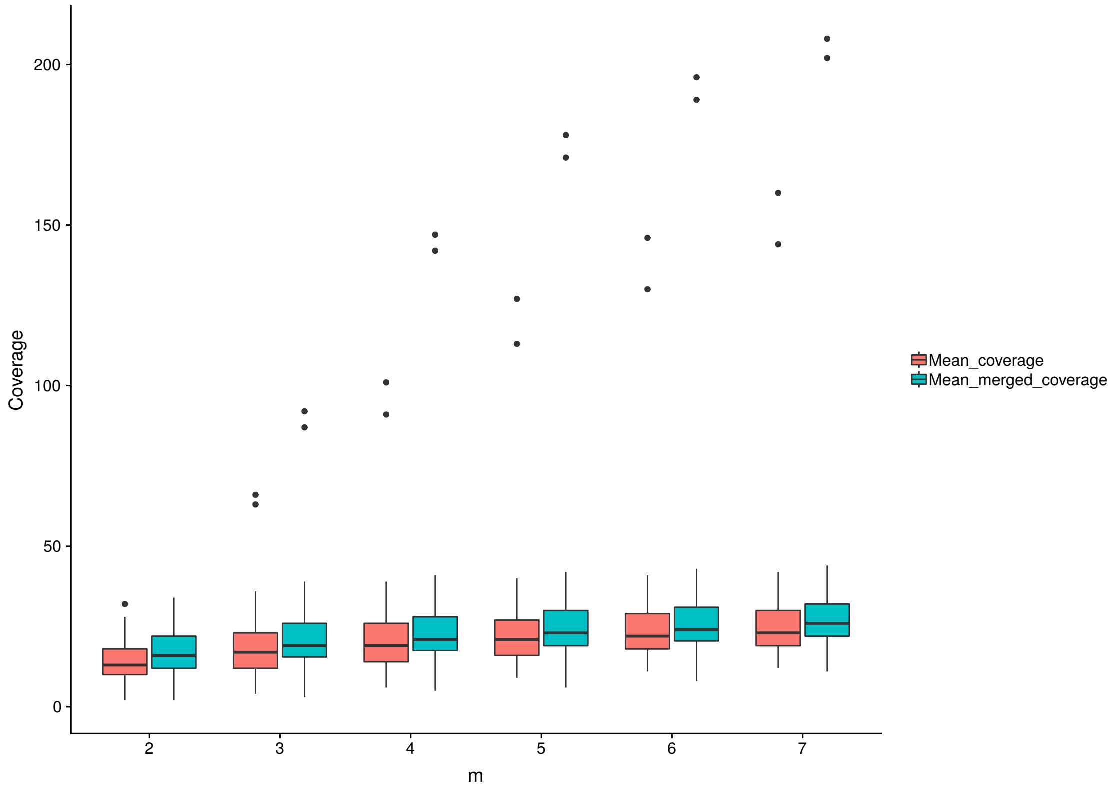


FIGURE A.2: Distribution of the mean coverage before and after merging loci for each iteration of the m parameter.Results for the combined dataset including samples from *Apodemus flavicollis* and *Apodemus sylvaticus*. Data can be found on github: *https://github.com/Marisa89/ddRADseq_poland/blob/master/Tables/Apodemus/Table_coverage_Apodemus.csv*

A.4 Estimation of the best parameters for *Apodemus flavicollis* dataset.

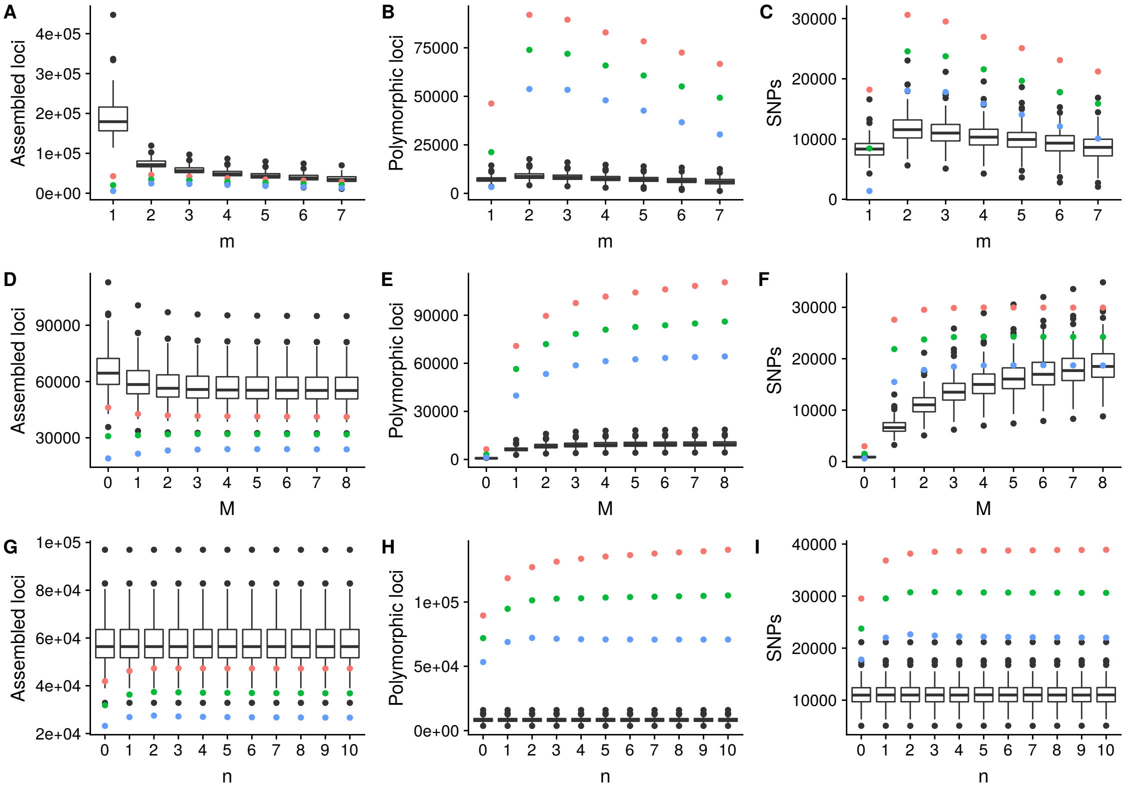


FIGURE A.3: Number of assembled loci, polymorphic loci and SNPs for iterating values of m, M and n parameters. Blue circles represent data found in at least 40% of the population, green circles in the 60% and red circles in the 80%. Results for *Apodemus flavicollis* samples. Data used to generate the graph can be found on:*https://github.com/Marisa89/ddRADseq_poland/blob/master/Tables/A.flavicollis/Table_selection_best_parameters_Aflavicollis.xlsx*

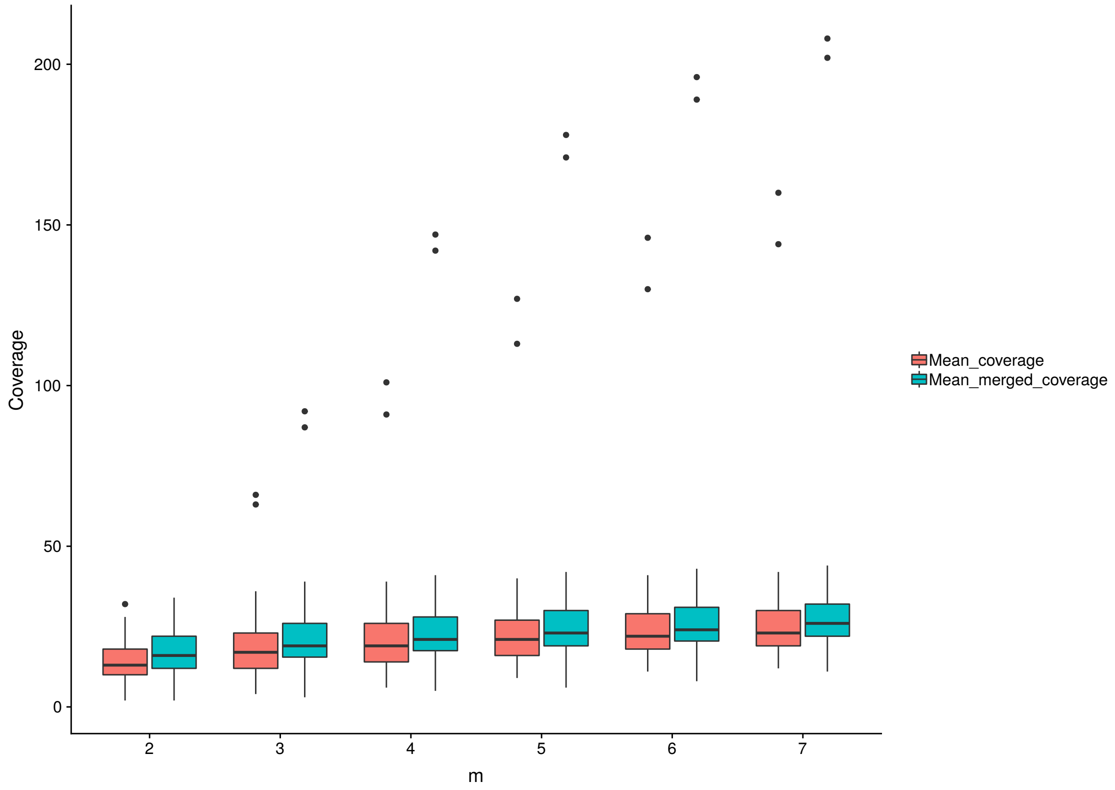


FIGURE A.4: Distribution of the mean coverage before and after merging loci for each iteration of the m parameter for *Apodemus flavicollis* samples. Data used to build the graph is available in: *https://github.com/Marisa89/ddRADseq_poland/blob/master/Tables/A.flavicollis/Table_coverage_Aflavicollis.csv*

A.5 Cross-validation errors

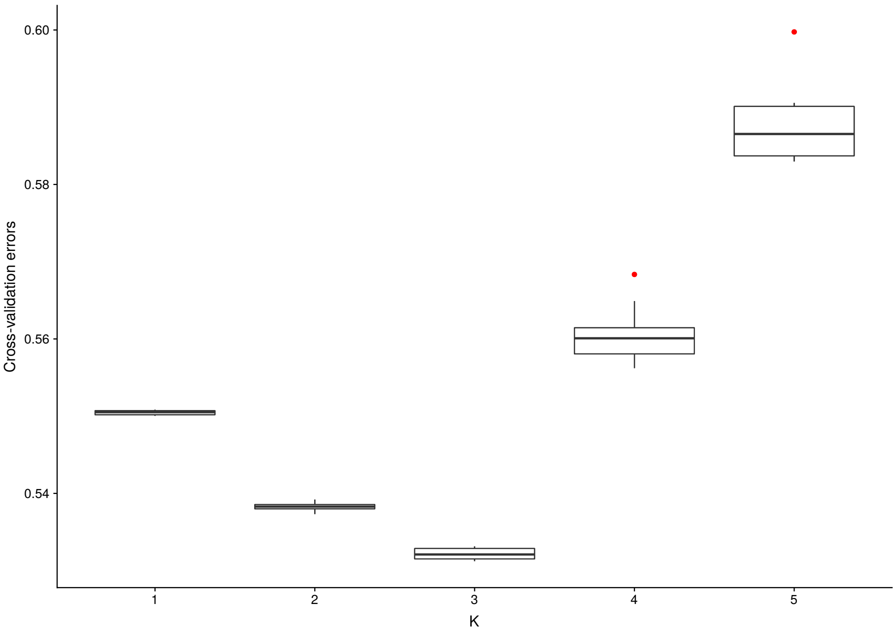


FIGURE A.5: Cross-validation errors obtained for values of K between 1 and 5 for 10 runs with different seeds for all samples

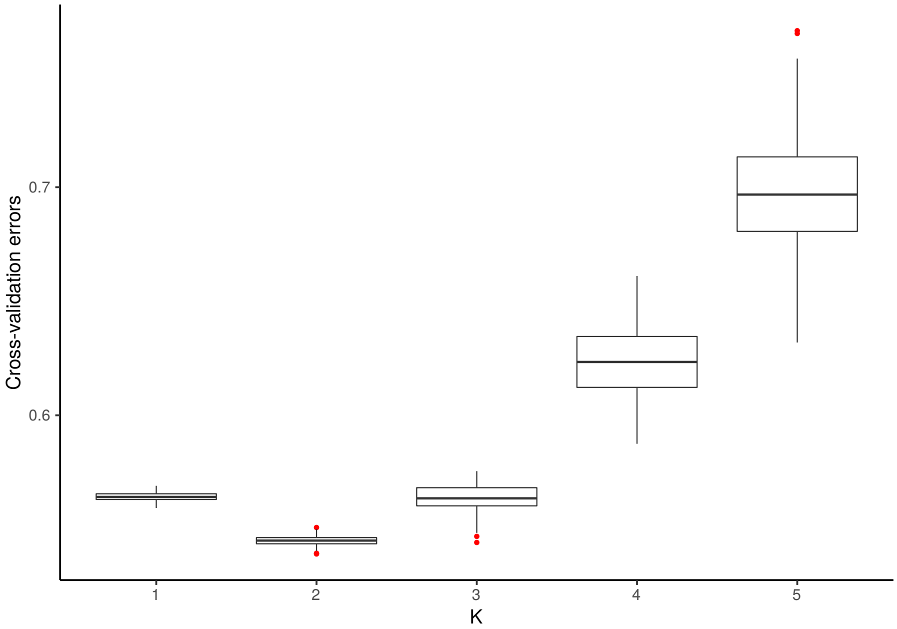


FIGURE A.6: Cross-validation errors obtained for values of K between 1 and 5 for the 100 permutations performed with randomly-drawn equal number of samples per population (n = 15)

A.6 Catalogue of loci used for species differentiation

Due to the size of the catalogue, the files has been uploaded into Dropbox. They are available on the following link: *https://www.dropbox.com/sh/3757wzer94eef85/AADRXN6GT5J6QJ-JHFySi34Aa?dl=0*

A.7 117 loci with the highest divergence

TABLE A.3: List of loci with the highest divergence between both species

3211	11103	20032	35338	45028	62435
4189	11112	20410	35417	45908	62495
4759	12321	20426	35799	47463	62544
4835	12823	21475	36256	51367	62719
4967	13690	22268	36342	51435	62846
5241	13708	23146	36597	51533	64055
5937	13820	23682	36821	53072	64057
6024	14596	24277	37171	53520	64228
6497	14916	25086	37193	53551	64457
6678	15177	25874	38518	53831	64631
7484	15553	26440	39788	54014	65038
7873	16614	26520	39844	57051	65147
8108	16806	27415	39936	57466	65161
8225	17192	30030	40266	59850	65163
9035	17594	31857	40440	60100	66267
9762	18137	32033	41161	60367	66602
10097	18207	32483	42388	60452	67679
10594	19036	32926	42581	61260	
10967	19729	33371	42639	61310	
11041	19799	33510	42900	62087	

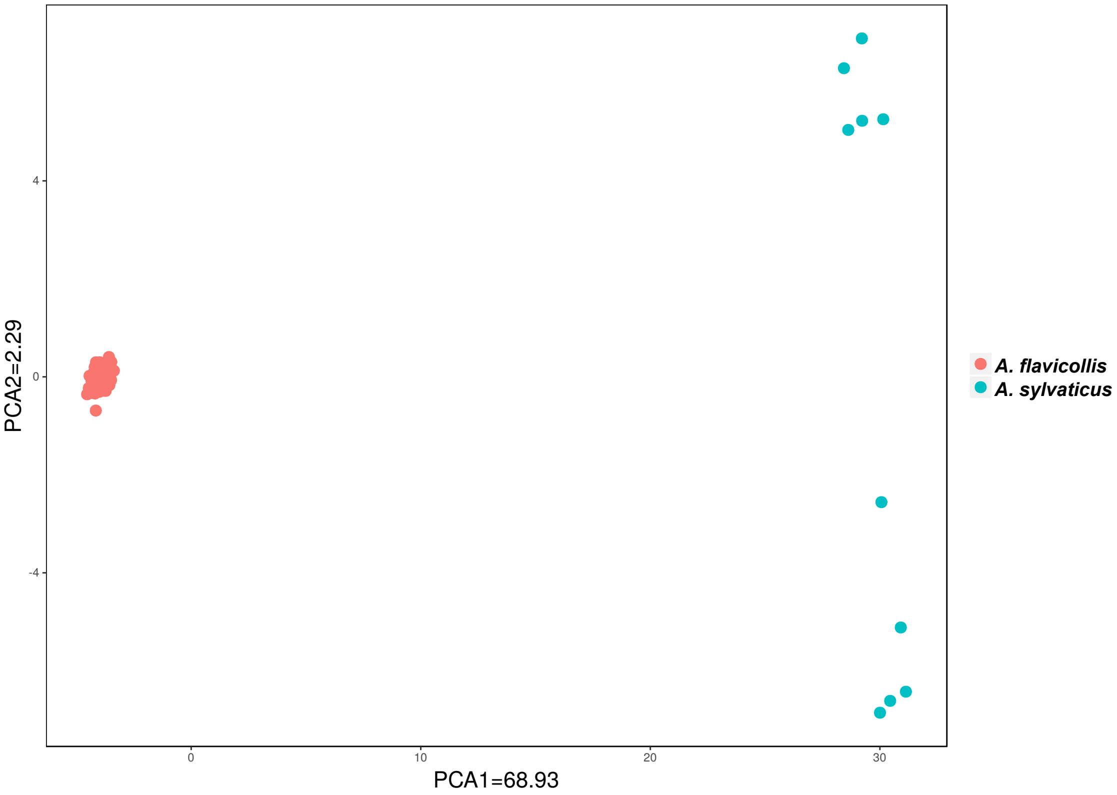


FIGURE A.7: Principal Component Analysis using the 117 loci with the highest divergence only for the Polish samples

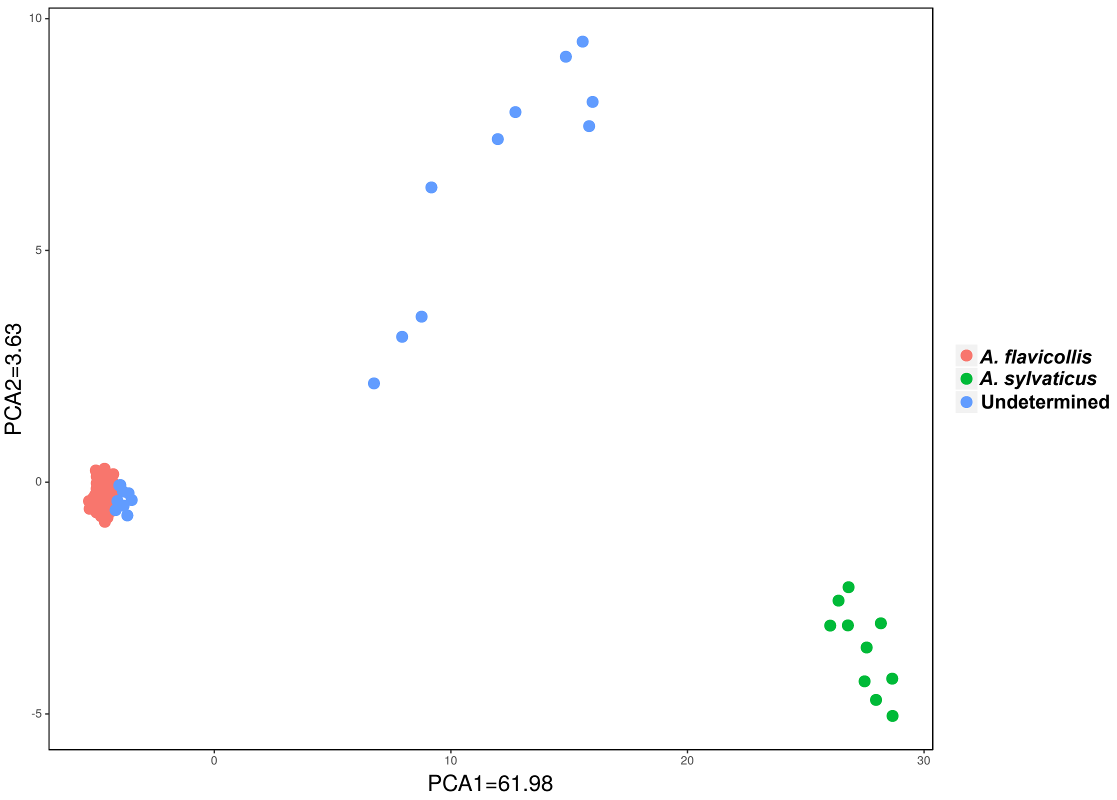


FIGURE A.8: Principal Component Analysis using the 117 loci with the highest divergence includying other European samples

A.8 European and Tunisian samples information

TABLE A.4: Details of the 20 European and Tunisian samples used to check the catalague

ID	**Source**	**Code**
AT1	Johan Michaux	JRM-203
AT2	Johan Michaux	JRM-204
LT1	Karol Zub	JB-466
LT2	Karol Zub	JB-468
LT3	Karol Zub	JB-470
LT4	Karol Zub	JB-485
LT5	Karol Zub	JB-475
RO1	Johan Michaux	JRM-2729
RO2	Johan Michaux	JRM-2720
RO3	Johan Michaux	JRM-2721
WL1	National Museums Scotland	NMS.Z.2009.101.1295M
WL2	National Museums Scotland	NMS.Z.2009.101.1296M
WL3	National Museums Scotland	NMS.Z.2009.101.1203M
WL4	National Museums Scotland	NMS.Z.2009.101.1294M
TN1	Johan Michaux	JRM-138
TN2	Johan Michaux	JRM-139
TN3	Johan Michaux	JRM-140
SC1	National Museums Scotland	NMS.Z.2009.101.1M
SC2	National Museums Scotland	NMS.Z.2009.101.2M
SC3	National Museums Scotland	NMS.Z.2009.101.3M

A.9 Code

All the code is available at:

https://github.com/Marisa89/ddRADseq_poland/tree/master/Code 1- Demultiplex_concatenation.sh
2- Iteration_parameter_selection.sh
3- Graphs_Iteration_parameters.R
4- PCA_plots_species.R
5- PCA_plots_flavicollis.R
6- Generate_files_for_divergence.sh
7- SNP_error_rate.sh
8- Loci_Allele_error_rate.sh
9- Allele_error_rate.R
10- Permutations.sh
11- Permutations_genetic_diversity_and_Fst_tables.R
12- Admixture_different_seed.sh

The code used to calculate divergence is available at: *https://github.com/jarekbryk/divergenceR*

Chapter B

Appendix Chapter 3

B.1 Sequences of the designed adapters

All the sequences below have been modified from Franchini *et al.* (2017)

TABLE B.1: i5 adapter′s sequences

Adapter name	Sequence
i5top_#01_AAGACTGG	CGCTCTTCCGATCTVBBNAAGACTGGTGCA/3Phos/
i5top_#02_ATGTTGGC	CGCTCTTCCGATCTVBBNATGTTGGCTGCA/3Phos/
i5top_#03_ATTGGCTG	CGCTCTTCCGATCTVBBNATTGGCTGTGCA/3Phos/
i5top_#04_CCTCATCT	CGCTCTTCCGATCTVBBNCCTCATCTTGCA/3Phos/
i5top_#05_CGGAATTG	CGCTCTTCCGATCTVBBNCGGAATTGTGCA/3Phos/
i5top_#06_CAAGGTGA	CGCTCTTCCGATCTVBBNCAAGGTGATGCA/3Phos/
i5top_#07_GACTTGAG	CGCTCTTCCGATCTVBBNGACTTGAGTGCA/3Phos/
i5top_#08_GAATCACG	CGCTCTTCCGATCTVBBNGAATCACGTGCA/3Phos/
i5top_#09_GGATTGTC	CGCTCTTCCGATCTVBBNGGATTGTCTGCA/3Phos/
i5top_#10_TCCTTCAC	CGCTCTTCCGATCTVBBNTCCTTCACTGCA/3Phos/
i5top_#11_TGTCAGTG	CGCTCTTCCGATCTVBBNTGTCAGTGTGCA/3Phos/
i5top_#12_TTCTGAGG	CGCTCTTCCGATCTVBBNTTCTGAGGTGCA/3Phos/
i5bottom_#01_AAGACTGG	/5Phos/CCAGTCTTNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#02_ATGTTGGC	/5Phos/GCCAACATNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#03_ATTGGCTG	/5Phos/CAGCCAATNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#04_CCTCATCT	/5Phos/AGATGAGGNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#05_CGGAATTG	/5Phos/CAATTCCGNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#06_CAAGGTGA	/5Phos/TCACCTTGNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#07_GACTTGAG	/5Phos/CTCAAGTCNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#08_GAATCACG	/5Phos/CGTGATTCNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#09_GGATTGTC	/5Phos/GACAATCCNVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#10_TCCTTCAC	/5Phos/GTGAAGGANVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#11_TGTCAGTG	/5Phos/CACTGACANVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
i5bottom_#12_TTCTGAGG	/5Phos/CCTCAGAANVVBAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

TABLE B.2: i7 adapter's sequences

Adapter name	Sequence
i7-top_#01_AGAGTTCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNAGAGTTCG
i7-top_#02_ACCTGTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNACCTGTTG
i7-top_#03_AATCGCCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNAATCGCCT
i7-top_#04_CTGGTTCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNCTGGTTCA
i7-top_#05_CGACAAGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNCGACAAGA
i7-top_#06_CAGTCGAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNCAGTCGAA
i7-top_#07_GTCAGAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNGTCAGAAC
i7-top_#08_GGCAATCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNGGCAATCT
i7-top_#09_GTGGTCTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNGTGGTCTT
i7-top_#10_TTGTTCCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNTTGTTCCG
i7-top_#11_TCGCATTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNTCGCATTC
i7-top_#12_TCGAACCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNTCGAACCA
i7-bottom_#01_AGAGTTCG	TACGAACTCTNVVBAGATCGGAAGAGCA
i7-bottom_#02_ACCTGTTG	TACAACAGGTNVVBAGATCGGAAGAGCA
i7-bottom_#03_AATCGCCT	TAAGGCGATTNVVBAGATCGGAAGAGCA
i7-bottom_#04_CTGGTTCA	TATGAACCAGNVVBAGATCGGAAGAGCA
i7-bottom_#05_CGACAAGA	TATCTTGTCGNVVBAGATCGGAAGAGCA
i7-bottom_#06_CAGTCGAA	TATTCGACTGNVVBAGATCGGAAGAGCA
i7-bottom_#07_GTCAGAAC	TAGTTCTGACNVVBAGATCGGAAGAGCA
i7-bottom_#08_GGCAATCT	TAAGATTGCCNVVBAGATCGGAAGAGCA
i7-bottom_#09_GTGGTCTT	TAAAGACCACNVVBAGATCGGAAGAGCA
i7-bottom_#10_TTGTTCCG	TACGGAACAANVVBAGATCGGAAGAGCA
i7-bottom_#11_TCGCATTC	TAGAATGCGANVVBAGATCGGAAGAGCA
i7-bottom_#12_TCGAACCA	TATGGTTCGANVVBAGATCGGAAGAGCA

TABLE B.3: Combinatorial outer adapter sequences

Adapter name	**Sequence**
i501_AGCATGGA	AATGATACGGCGACCACCGAGATCTACAC{AGCATGGA}ACACTCTTTCCCTACACGAC*G
i502_CCTGGAAT	AATGATACGGCGACCACCGAGATCTACAC{CCTGGAAT}ACACTCTTTCCCTACACGAC*G
i503_GCAAGCAA	AATGATACGGCGACCACCGAGATCTACAC{GCAAGCAA}ACACTCTTTCCCTACACGAC*G
i504_TGAGGATG	AATGATACGGCGACCACCGAGATCTACAC{TGAGGATG}ACACTCTTTCCCTACACGAC*G
i701_ACACTCAG	CAAGCAGAAGACGGCATACGAGAT{CTGAGTGT}GTGACTGGAGTTCAGACGTGTGC*T
i702_CAGTCGAA	CAAGCAGAAGACGGCATACGAGAT{TTCGACTG}GTGACTGGAGTTCAGACGTGTGC*T
i703_GGCTCAAT	CAAGCAGAAGACGGCATACGAGAT{ATTGAGCC}GTGACTGGAGTTCAGACGTGTGC*T
i704_TTCCGCTT	CAAGCAGAAGACGGCATACGAGAT{AAGCGGAA}GTGACTGGAGTTCAGACGTGTGC*T

TABLE B.4: Adapters designed to substitute problematic adapters, These adapters have not been tested yet

Adapter name	Sequence
quaddRAD-i7-bottom_03_ACACTCAG	TACTGAGTGTNVVBAGATCGGAAGAGCA
quaddRAD-i7-bottom_08_GATGACTC	TAGAGTCATCNVVBAGATCGGAAGAGCA
quaddRAD-i7-bottom_09_GTTCCAAG	TACTTGGAACNVVBAGATCGGAAGAGCA
quaddRAD-i7-top_03_ACACTCAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNACACTCAG
quaddRAD-i7-top_08_GATGACTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNGATGACTC
quaddRAD-i7-top_09_GTTCCAAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVBBNGTTCCAAG

B.2 Library preparation and sequencing performance

B.2.1 Library preparation

quaddRADseq library was prepared, in collaboration with Dr Marek Kucka and Dr Frank Yingguang Chan, from the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, following Franchini *et al.* (2017) protocol, using a different combination of enzymes (SbfI and MseI) and our own set of adapters. The protocol was first tested using 16 samples with different qualities, and testing the complete set of adapters.

Inner adapters were prepared by annealing each single-stranded oligo with its complementary oligo parter, following Franchini *et al.* (2017). 5 μl of each oligo (bottom and top) were mix with 40 μl of AB buffer (500 mM NaCl, 100 mM Tris-Cl, ph= 8.0). In a thermocycler the mix of adapters was warmed at 98°C for 2.5 min and then, they were cool down -1°C per minute until reaching 15°C. Once the adaptors were ready, they were kept at -20°C and they were used within 2 weeks.

Afterwards, the genomic DNA was digested and ligated to the adapters in a 40 μl single step reaction with CutSmart® buffer, 15 Units of HF-*SbfI* and 15 units of HF-*MseI* (R3642L and R0525L, respectively, both from New England Biolabs, Frankfurt am Main, Germany), 4 μl 10mM of rATP, 400 U of T4 DNA ligase, 0.75 μl of 10 μM inner adapters, 60 ng of DNA and water. The mix was incubated at 30 °C for 3 hours and the reaction was stopped with 10 μl of 50mM EDTA. Samples were cleaned and size selected using 0.4x and 0.8x Sera-Mag SpeedBeads solution (GElifesciences, Marlborough, MA, USA) and eluted in 30 μM .

Three inner adapters, (Number 3, 8 and 9) were identified as potentially problematic, as they could partially reproduce the cutting site used by the enzymes. These samples were digested and adapters were ligated in two different steps, to avoid the digestion of the ligated adapters. The digestion was performed in a 30 μl reaction with with CutSmart® buffer, 15 Units of HF-*SbfI* and 15 units of HF-*MseI* (R3642L and R0525L, 60 ng of DNA and water. The mix was incubated at 37 °C for 90 minutes and 65°C for 20 minutes. Adapters were ligated in a 40μl reaction with CutSmart® buffer, 4 μl 10mM of rATP, 400 U of T4 DNA ligase, 0.75 μl of 10 μM inner adapters, 30 ul of digested DNA and water. The mix was incubated at room temperature for 90 minutes and the reaction was stopped with 10 *μl* of 50mM EDTA.

Afterwards the outer adapters were ligated through a PCR step conducted in 50 *μl* with 4 *μl* of each 5mM primers, 1 *μl* of 10 mM dNTPs, 10.5 *μl* of purified water,10 *μl* of 5x Q5-HF Buffer, 0.5 *μl* of Q5-HF DNA Polymerase (New England Biolabs, Frankfurt am Main, Germany) and 20 *μl* of template DNA.

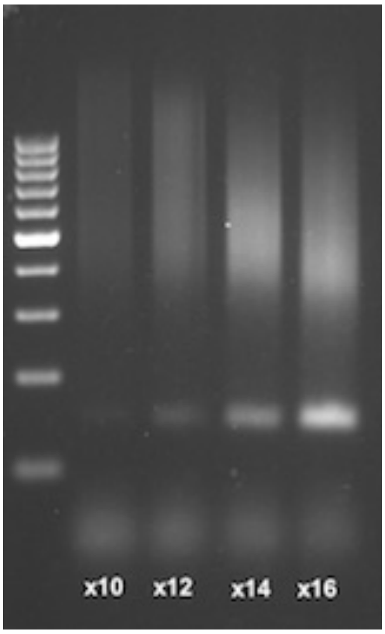


FIGURE B.1: Determination of the minimum number of cycles required for library amplification through agarose gel electrophoresis

Different number of cycles were tested, in order to find the lower number of cycles that produced a clear amplification of the library (Figure B.1. After an initial denaturation step of 30 seconds at 98°C, PCR reaction was carried out in 14 cycles (15 seconds at 98°C , 30 seconds at 67°C and 60 seconds at 72°C). Final elongation step was performed at 72 °C for 2 minutes. Clean up was performed using 0.8x Sera-Mag SpeedBeads solution (GElifesciences, Marlborough, MA, USA) and DNA was eluted in 22 *μl*. Libraries were checked in a 1% agarose gel (Figure B.2.

Samples were multiplexed by adding 40ng of DNA from each sample. Fragments between 300-600 bp were manually cut from an agarose gel (Sage Science)(Figure B.3. Gel extraction was performed using the GeneJet gel extraction kit (K0691, Thermofisher, Waltham, Massachusetts, United States) from approximately 1 g.

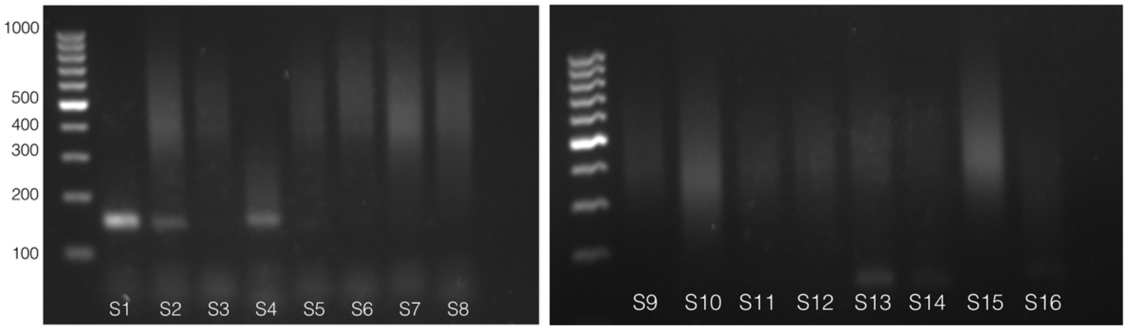


FIGURE B.2: Distribution of fragment lengths after digestion and library amplification for all the samples tested

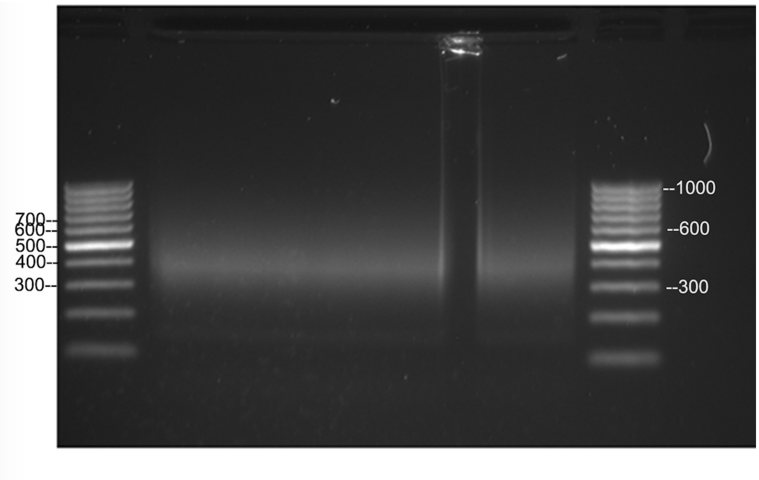


FIGURE B.3: Size selection on agarose gel. Fragments between 300 and 600 bp were manually cut from the gel

Due to the small number of samples included in the library, the quaddrad library was multiplexed for sequencing along with a whole genome from *Apodemus sylvaticus* (20%:80%).

The quaddrad library and the final sequencing library were analyzed with 2100 Bioanalyzer (Agilent Technologies)(Figure 4.2 and paired-end sequenced with Hiseq3000 with cBot (Illumina) at the Genome Center of the Max Planck institute for Developmental Biology in Tübingen B.4.

B.2.2 Processing of RAD-tags and selection of the best parameters

Reads were demultiplexed based on outer adapters by the Genome Center of the Max Planck institute for Developmental Biology in Tübingen. The

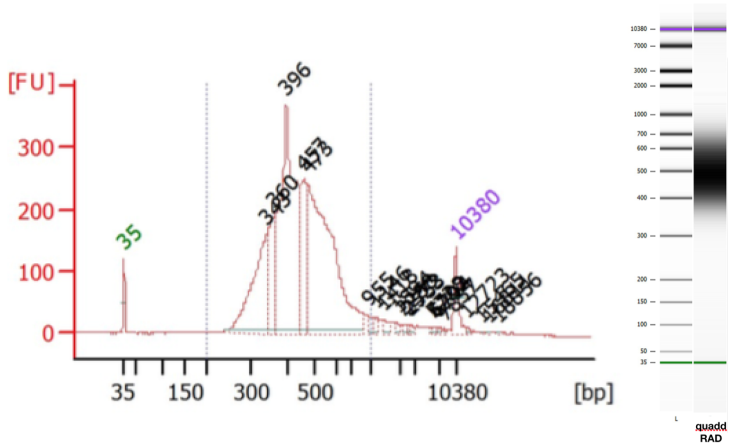


FIGURE B.4: Distribution of sizes for the quaddRAD library. The higher the fluorescence (FU), the higher the amount of DNA of a specific size.

inner demultiplex, even when it was unneccesary, as all the samples could have been classified by the outer adapters, was performed using Stacks version 1.48 (Catchen *et al.*, 2011). PCR duplicates were removed using clone_filter program. Reads were demultiplexed and quality filtered using process_radtags program. Reads containing adapter sequences, uncalled bases or low quality scores or that were marked by Illumina's chastity/purity filter as failing were discarded. Barcode rescue was enabled, allowing 2 mismatches on the tag sequence and sequences were truncated to a final length of 136 bp. Chimeric sequences produced during sequencing were extracted using process_radtags, considering pairs of tags used in different samples.

Chapter C

Appendix Chapter 4

C.1 Sample information

TABLE C.1: Locality, coordinates, species identification, sample ID and combination of adapter sequences used for all the samples included on the library preparation for the Phylogeography project. Samples sequenced by duplicate or triplicate are also indicated.

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Plate1						
Austria	47.22	9.79	*A. flavicollis*	AT1	501-1-1-701	
Germany	51.05	13.74	*A. flavicollis*	DE1	501-2-2-701	
Italy	42.46	13.93	*A. flavicollis*	IT1	501-4-4-701	
Russia	52.69	30.62	*A. flavicollis*	RU1	501-5-5-701	
Slovakia	48.72	21.86	*A. flavicollis*	SK1	501-6-6-701	
Sweden	60.13	18.64	*A. flavicollis*	SE1	501-7-7-701	
France	44.97	5.53	*A. flavicollis*	FR1	501-10-10-701	
Poland	52.25	17.09	*A. flavicollis*	PL	501-11-11-701	
Slovakia	49.37	22.45	*A. flavicollis*	SK2	501-12-12-701	
Austria	47.22	9.79	*A. flavicollis*	AT2	501-1-1-702	
Germany	51.05	13.74	*A. flavicollis*	DE2	501-2-2-702	
Macedonia	41.00	21.18	*A. flavicollis*	MK1	501-4-4-702	
Russia	52.69	30.62	*A. flavicollis*	RU2	501-5-5-702	
Slovenia	45.59	14.03	*A. flavicollis*	SI1	501-6-6-702	
Sweden	60.13	18.64	*A. flavicollis*	SE2	501-7-7-702	
Italy	38.11	15.65	*A. flavicollis*	IT2	501-10-10-702	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Poland	52.25	17.09	*A. flavicollis*	PL1	501-11-11-702	
Slovakia	49.37	22.45	*A. flavicollis*	SK3	501-12-12-702	
Austria	47.22	9.79	*A. flavicollis*	AT3	501-1-1-703	
Germany	51.05	13.74	*A. flavicollis*	DE3	501-2-2-703	
Macedonia	41.00	21.18	*A. flavicollis*	MK2	501-4-4-703	
Russia	52.69	30.62	*A. flavicollis*	RU3	501-5-5-703	
Slovenia	45.59	14.03	*A. flavicollis*	SI2	501-6-6-703	
Sweden	60.13	18.64	*A. flavicollis*	SE3	501-7-7-703	
Italy	38.11	15.65	*A. flavicollis*	IT3	501-10-10-703	
Poland	52.25	17.09	*A. flavicollis*	PL2	501-11-11-703	
Slovakia	49.37	22.45	*A. flavicollis*	SK4	501-12-12-703	
Austria	47.22	9.79	*A. flavicollis*	AT4	501-1-1-704	
Greece	40.95	22.45	*A. flavicollis*	GR1	501-2-2-704	
Romania		22.90	*A. flavicollis*	RO1	501-4-4-704	
Russia	59.93	30.34	*A. flavicollis*	RU4	501-5-5-704	
Slovenia	45.59	14.03	*A. flavicollis*	SI3	501-6-6-704	
Denmark	56.26	9.50	*A. flavicollis*	DK1	501-7-7-704	DK7
Italy	38.11	15.65	*A. flavicollis*	IT4	501-10-10-704	
Poland	49.36	22.51	*A. flavicollis*	PL3	501-11-11-704	
Spain	41.79	2.39	*A. flavicollis*	ES1	501-12-12-704	
Germany	52.02	8.52	*A. flavicollis*	DE4	502-1-1-701	
Greece	40.95	22.45	*A. flavicollis*	GR2	502-2-2-701	
Romania	45.37	22.90	*A. flavicollis*	RO2	502-4-4-701	
Russia	59.93	30.34	*A. flavicollis*	RU5	502-5-5-701	
Slovenia	45.59	14.03	*A. flavicollis*	SI4	502-6-6-701	
Denmark	56.26	9.50	*A. flavicollis*	DK2	502-7-7-701	
Italy	38.11	15.65	*A. flavicollis*	IT5	502-10-10-701	
Poland	49.36	22.51	*A. flavicollis*	PL4	502-11-11-701	
Spain	41.79	2.39	*A. flavicollis*	ES2	502-12-12-701	ES4
Germany	51.16	12.93	*A. flavicollis*	DE5	502-1-1-702	
Greece	39.57	20.76	*A. flavicollis*	GR3	502-2-2-702	
Romania	45.37	22.90	*A. flavicollis*	RO3	502-4-4-702	
Estonia	59.44	24.75	*A. flavicollis*	RU6	502-5-5-702	
Slovenia	45.67	14.69	*A. flavicollis*	SI5	502-6-6-702	
Denmark	56.26	9.50	*A. flavicollis*	DK3	502-7-7-702	
Poland	52.25	17.09	*A. flavicollis*	PL5	502-10-10-702	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Poland	49.36	22.51	*A. flavicollis*	PL6	502-11-11-702	
Spain	41.79	2.39	*A. flavicollis*	ES3	502-12-12-702	
Germany	51.16	12.93	*A. flavicollis*	DE6	502-1-1-703	
Italy	42.46	13.93	*A. flavicollis*	IT6	502-2-2-703	
Russia	48.68	44.45	*A. flavicollis*	RU7	502-4-4-703	
Slovakia	48.72	21.86	*A. flavicollis*	SK5	502-5-5-703	
Slovenia	45.67	14.69	*A. flavicollis*	SI6	502-6-6-703	
France	45.36	2.34	*A. flavicollis*	FR2	502-7-7-703	
Poland	52.25	17.09	*A. flavicollis*	PL7	502-10-10-703	
Poland	49.36	22.51	*A. flavicollis*	PL8	502-11-11-703	
Belgium	50.25	5.67	*A. sylvaticus*	BE1	502-12-12-703	
Germany	51.16	12.93	*A. flavicollis*	DE7	502-1-1-704	
Italy	42.46	13.93	*A. flavicollis*	IT7	502-2-2-704	
Russia	48.68	44.45	*A. flavicollis*	RU8	502-4-4-704	
Slovakia	48.72	21.86	*A. flavicollis*	SK6	502-5-5-704	
Slovenia	45.67	14.69	*A. flavicollis*	SI7	502-6-6-704	
France	44.97	5.53	*A. flavicollis*	FR3	502-7-7-704	
Poland	52.25	17.09	*A. flavicollis*	PL9	502-10-10-704	
Russia	51.68	39.21	*A. flavicollis*	RU9	502-11-11-704	
Belgium	50.25	5.67	*A. sylvaticus*	BE2	502-12-12-704	
Belgium	50.25	5.67	*A. sylvaticus*	BE3	503-1-1-701	
Belgium	50.25	5.67	*A. sylvaticus*	BE4	503-2-2-701	
Belgium	50.65	4.87	*A. sylvaticus*	BE5	503-4-4-701	
Denmark	56.26	9.50	*A. sylvaticus*	DK4	503-5-5-701	
Denmark	56.26	9.50	*A. sylvaticus*	DK5	503-6-6-701	
Denmark	56.26	9.50	*A. sylvaticus*	DK6	503-7-7-701	
France	43.16	6.62	*A. sylvaticus*	FR4	503-10-10-701	
France	43.00	6.39	*A. sylvaticus*	FR5	503-11-11-701	
France	43.00	6.39	*A. sylvaticus*	FR6	503-1-1-702	
France	43.00	6.39	*A. sylvaticus*	FR7	503-2-2-702	
France	42.48	3.13	*A. sylvaticus*	FR8	503-4-4-702	
France	42.48	3.13	*A. sylvaticus*	FR9	503-5-5-702	
Spain	41.79	2.39	*A. flavicollis*	ES4	503-6-6-702	ES2
Denmark	56.26	9.50	*A. flavicollis*	DK7	503-7-7-702	DK1
Plate2						
Tunisia	36.77	8.68	*A. sylvaticus*	TN1	503-1-1-703	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Germany	54.19	13.19	*A. flavicollis*	DE8	503-2-2-703	
Belgium	50.72	5.75	*A. sylvaticus*	BE6	503-4-4-703	
Germany	51.28	13.55	*A. sylvaticus*	DE9	503-5-5-703	
Italy	38.11	15.65	*A. sylvaticus*	IT8	503-6-6-703	
Italy	42.46	13.93	*A. sylvaticus*	IT9	503-7-7-703	
Spain	41.79	2.39	*A. sylvaticus*	ES5	503-10-10-703	
Germany	52.83	13.83	*A. flavicollis*	DE10	503-11-11-703	
Germany	53.81	9.62	*A. flavicollis*	DE11	503-12-12-703	
Tunisia	36.77	8.68	*A. sylvaticus*	TN2	503-1-1-704	
Sweden	60.61	15.63	*A. flavicollis*	SE4	503-2-2-704	
France	43.60	3.90	*A. sylvaticus*	FR10	503-4-4-704	FR15
Germany	51.28	13.55	*A. sylvaticus*	DE12	503-5-5-704	
Italy	38.11	15.65	*A. sylvaticus*	IT10	503-6-6-704	
Italy	42.46	13.93	*A. sylvaticus*	IT11	503-7-7-704	
Spain	41.79	2.39	*A. sylvaticus*	ES6	503-10-10-704	
Germany	52.83	13.83	*A. flavicollis*	DE13	503-11-11-704	
Germany	53.81	9.62	*A. flavicollis*	DE14	503-12-12-704	
Tunisia	36.77	8.68	*A. sylvaticus*	TN3	504-1-1-701	
Sweden	60.61	15.63	*A. flavicollis*	SE5	504-2-2-701	
France	43.60	3.90	*A. sylvaticus*	FR11	504-4-4-701	
Ireland	54.33	-5.72	*A. sylvaticus*	IE1	504-5-5-701	
Italy	38.11	15.65	*A. sylvaticus*	IT12	504-6-6-701	
Slovenia	46.05	14.47	*A. sylvaticus*	SI8	504-7-7-701	
Sweden	56.67	12.86	*A. sylvaticus*	SE6	504-10-10-701	
Germany	52.83	13.83	*A. flavicollis*	DE15	504-11-11-701	
Germany	53.81	9.62	*A. flavicollis*	DE16	504-12-12-701	
Germany	51.10	12.34	*A. flavicollis*	DE17	504-1-1-702	
Sweden	56.88	14.81	*A. flavicollis*	SE7	504-2-2-702	
France	43.60	3.90	*A. sylvaticus*	FR12	504-4-4-702	
Ireland	54.33	-5.72	*A. sylvaticus*	IE2	504-5-5-702	
Italy	38.11	15.65	*A. sylvaticus*	IT13	504-6-6-702	
Slovenia	46.05	14.47	*A. sylvaticus*	SI9	504-7-7-702	
Sweden	57.65	14.70	*A. sylvaticus*	SE8	504-10-10-702	
Germany	51.53	9.94	*A. flavicollis*	DE18	504-11-11-702	
Germany	50.95	10.72	*A. flavicollis*	DE19	504-12-12-702	DE35
Germany	51.10	12.34	*A. flavicollis*	DE20	504-1-1-703	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Sweden	56.16	15.59	*A. flavicollis*	SE9	504-2-2-703	
France	43.60	3.90	*A. sylvaticus*	FR13	504-4-4-703	
Ireland	54.33	-5.72	*A. sylvaticus*	IE3	504-5-5-703	
Italy	38.17	16.00	*A. sylvaticus*	IT14	504-6-6-703	
Slovenia	45.57	13.79	*A. sylvaticus*	SI10	504-7-7-703	
Sweden	56.17	15.64	*A. sylvaticus*	SE10	504-10-10-703	
Germany	51.53	9.94	*A. flavicollis*	DE21	504-11-11-703	
Germany	50.95	10.72	*A. flavicollis*	DE22	504-12-12-703	
Germany	51.10	12.34	*A. flavicollis*	DE23	504-1-1-704	
Sweden	62.40	17.30	*A. flavicollis*	SE11	504-2-2-704	
France	43.60	3.90	*A. sylvaticus*	FR14	504-4-4-704	
Ireland	53.41	-8.24	*A. sylvaticus*	IE4	504-5-5-704	
Italy	38.17	16.00	*A. sylvaticus*	IT15	504-6-6-704	
Slovenia	45.57	13.79	*A. sylvaticus*	SI11	504-7-7-704	
Germany	52.91	12.17	*A. flavicollis*	DE24	504-10-10-704	
Germany	51.53	9.94	*A. flavicollis*	DE25	504-11-11-704	
Germany	50.95	10.72	*A. flavicollis*	DE26	504-12-12-704	
Germany	54.19	13.19	*A. flavicollis*	DE27	501-1-1-701	
Belgium	50.72	5.75	*A. sylvaticus*	BE7	501-2-2-701	
Germany	51.28	13.55	*A. sylvaticus*	DE28	501-4-4-701	
Ireland	53.41	-8.24	*A. sylvaticus*	IE5	501-5-5-701	
Italy	38.17	16.00	*A. sylvaticus*	IT16	501-6-6-701	
Spain	41.79	2.39	*A. sylvaticus*	ES7	501-7-7-701	
Germany	52.91	12.17	*A. flavicollis*	DE29	501-10-10-701	
Germany	51.53	9.94	*A. flavicollis*	DE30	501-11-11-701	
Lithuania	54.59	24.00	*A. flavicollis*	LT1	501-12-12-701	
Germany	54.19	13.19	*A. flavicollis*	DE31	501-1-1-702	
Belgium	50.72	5.75	*A. sylvaticus*	BE8	501-2-2-702	
Germany	51.28	13.55	*A. sylvaticus*	DE32	501-4-4-702	
Ireland	53.41	-8.24	*A. sylvaticus*	IE6	501-5-5-702	
Italy	42.46	13.93	*A. sylvaticus*	IT17	501-6-6-702	
Spain	41.79	2.39	*A. sylvaticus*	ES8	501-7-7-702	
Germany	52.91	12.17	*A. flavicollis*	DE33	501-10-10-702	
Germany	53.81	9.62	*A. flavicollis*	DE34	501-11-11-702	
Lithuania	54.59	24.00	*A. flavicollis*	LT2	501-12-12-702	
Lithuania	54.59	24.00	*A. flavicollis*	LT3	501-1-1-703	

Country	**Latitude**	**Longitude**	**Species**	**Sample**	**Adapters**	**DUP**
Lithuania	54.59	24.00	*A. flavicollis*	LT4	501-2-2-703	
Lithuania	54.59	24.00	*A. flavicollis*	LT5	501-4-4-703	
Lithuania	55.85	26.20	*A. flavicollis*	LT6	501-5-5-703	
Lithuania	55.85	26.20	*A. flavicollis*	LT7	501-6-6-703	
Lithuania	55.98	24.61	*A. flavicollis*	LT8	501-7-7-703	
Lithuania	55.85	26.20	*A. flavicollis*	LT9	501-10-10-703	
Lithuania	55.98	24.61	*A. flavicollis*	LT10	501-11-11-703	
Lithuania	55.85	26.20	*A. flavicollis*	LT11	501-1-1-704	
Lithuania	55.52	21.11	*A. flavicollis*	LT12	501-2-2-704	
Lithuania	55.85	26.20	*A. flavicollis*	LT13	501-4-4-704	
Lithuania	55.52	21.11	*A. flavicollis*	LT14	501-5-5-704	
Germany	50.95	10.72	*A. flavicollis*	DE35	501-6-6-704	DE19
France	43.60	3.90	*A. sylvaticus*	FR15	501-7-7-704	FR10
Plate 3						
Lithuania	55.52	21.11	*A. flavicollis*	LT15	502-1-1-701	
Serbia	44.73	20.41	*A. flavicollis*	RS2	502-2-2-701	
Serbia	44.73	20.41	*A. flavicollis*	RS10	502-4-4-701	
Serbia	44.73	20.41	*A. sylvaticus*	RS18	502-5-5-701	
Denmark	55.29	8.69	*A. sylvaticus*	DK11	502-6-6-701	
England	52.34	0.52	*A. sylvaticus*	EN8	502-7-7-701	
France	46.08	3.03	*A. sylvaticus*	FR23	502-10-10-701	
France	50.08	1.57	*A. sylvaticus*	FR31	502-11-11-701	
Iceland	64.39	-15.29	*A. sylvaticus*	IS4	502-12-12-701	
Lithuania	54.66	24.83	*A. flavicollis*	LT21	502-1-1-702	
Serbia	44.73	20.41	*A. sylvaticus*	RS3	502-2-2-702	
Serbia	44.73	20.41	*A. flavicollis*	RS11	502-4-4-702	
Serbia	44.73	20.41	*A. flavicollis*	RS19	502-5-5-702	
England	50.46	-4.73	*A. sylvaticus*	EN1	502-6-6-702	
France	48.55	-3.40	*A. sylvaticus*	FR16	502-7-7-702	FR32
France	45.80	1.13	*A. sylvaticus*	FR24	502-10-10-702	
Germany	54.28	8.84	*A. sylvaticus*	DE37	502-11-11-702	
Iceland	64.10	-21.80	*A. sylvaticus*	IS5	502-12-12-702	
Lithuania	55.98	24.61	*A. flavicollis*	LT16	502-1-1-703	
Serbia	44.73	20.41	*A. flavicollis*	RS4	502-2-2-703	
Serbia	44.73	20.41	*A. sylvaticus*	RS12	502-4-4-703	
Germany	51.28	13.55	*A. sylvaticus*	DE36	502-5-5-703	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
England	50.46	-4.73	*A. sylvaticus*	EN2	502-6-6-703	
France	48.55	-3.40	*A. sylvaticus*	FR17	502-7-7-703	
France	45.80	1.13	*A. sylvaticus*	FR25	502-10-10-703	
Germany	54.28	8.84	*A. sylvaticus*	DE38	502-11-11-703	
Iceland	64.10	-21.80	*A. sylvaticus*	IS6	502-12-12-703	
Lithuania	54.66	24.83	*A. flavicollis*	LT22	502-1-1-704	
Serbia	44.73	20.41	*A. flavicollis*	RS5	502-2-2-704	
Serbia	44.73	20.41	*A. sylvaticus*	RS13	502-4-4-704	
Sweden	59.40	13.51	*A. sylvaticus*	SE12	502-5-5-704	
England	50.46	-4.73	*A. sylvaticus*	EN3	502-6-6-704	
France	48.55	-3.40	*A. sylvaticus*	FR18	502-7-7-704	
France	45.80	1.13	*A. sylvaticus*	FR26	502-10-10-704	
Germany	54.28	8.84	*A. sylvaticus*	DE39	502-11-11-704	
Iceland	64.10	-21.80	*A. sylvaticus*	IS7	502-12-12-704	
Lithuania	55.98	24.61	*A. flavicollis*	LT17	503-1-1-701	
Serbia	44.73	20.41	*A. sylvaticus*	RS6	503-2-2-701	
Serbia	44.73	20.41	*A. sylvaticus*	RS14	503-4-4-701	
Sweden	59.40	13.51	*A. sylvaticus*	SE13	503-5-5-701	
England	50.46	-4.73	*A. sylvaticus*	EN4	503-6-6-701	
France	48.55	-3.40	*A. sylvaticus*	FR19	503-7-7-701	
France	45.80	1.13	*A. sylvaticus*	FR27	503-10-10-701	
Germany	54.28	8.84	*A. sylvaticus*	DE40	503-11-11-701	
Iceland	64.10	-21.80	*A. sylvaticus*	IS8	503-12-12-701	
Lithuania	55.98	24.61	*A. flavicollis*	LT23	503-1-1-702	
Serbia	44.73	20.41	*A. sylvaticus*	RS7	503-2-2-702	
Serbia	44.73	20.41	*A. flavicollis*	RS15	503-4-4-702	
Denmark	55.25	8.82	*A. sylvaticus*	DK8	503-5-5-702	
England	52.28	0.54	*A. sylvaticus*	EN5	503-6-6-702	
France	46.08	3.03	*A. sylvaticus*	FR20	503-7-7-702	
France	50.08	1.57	*A. sylvaticus*	FR28	503-10-10-702	
Iceland	64.39	-15.29	*A. sylvaticus*	IS1	503-11-11-702	
Norway	58.98	5.57	*A. sylvaticus*	NO1	503-12-12-702	
Lithuania	54.66	24.83	*A. flavicollis*	LT18	503-1-1-703	
Serbia	44.73	20.41	*A. flavicollis*	RS8	503-2-2-703	
Serbia	44.73	20.41	*A. sylvaticus*	RS16	503-4-4-703	
Denmark	55.25	8.82	*A. sylvaticus*	DK9	503-5-5-703	

Country	**Latitude**	**Longitude**	**Species**	**Sample**	**Adapters**	**DUP**
England	52.28	0.54	*A. sylvaticus*	EN6	503-6-6-703	
France	46.08	3.03	*A. sylvaticus*	FR21	503-7-7-703	
France	50.08	1.57	*A. sylvaticus*	FR29	503-10-10-703	
Iceland	64.39	-15.29	*A. sylvaticus*	IS2	503-11-11-703	
Norway	58.98	5.57	*A. sylvaticus*	NO2	503-12-12-703	
Serbia	44.73	20.41	*A. flavicollis*	RS1	503-1-1-704	
Serbia	44.73	20.41	*A. sylvaticus*	RS9	503-2-2-704	
Serbia	44.73	20.41	*A. sylvaticus*	RS17	503-4-4-704	
Denmark	55.29	8.69	*A. sylvaticus*	DK10	503-5-5-704	
England	52.34	0.52	*A. sylvaticus*	EN7	503-6-6-704	
France	46.08	3.03	*A. sylvaticus*	FR22	503-7-7-704	
France	50.08	1.57	*A. sylvaticus*	FR30	503-10-10-704	
Iceland	64.39	-15.29	*A. sylvaticus*	IS3	503-11-11-704	
Norway	58.98	5.57	*A. sylvaticus*	NO3	503-12-12-704	
Norway	58.98	5.57	*A. sylvaticus*	NO4	504-1-1-701	
Norway	59.47	10.09	*A. sylvaticus*	NO5	504-2-2-701	
Norway	59.47	10.09	*A. sylvaticus*	NO6	504-4-4-701	
Norway	59.47	10.09	*A. sylvaticus*	NO7	504-5-5-701	
Norway	59.47	10.09	*A. sylvaticus*	NO8	504-6-6-701	
Scotland	55.96	-3.26	*A. sylvaticus*	SC1	504-7-7-701	SC8
Scotland	55.96	-3.26	*A. sylvaticus*	SC2	504-10-10-701	
Scotland	55.96	-3.26	*A. sylvaticus*	SC3	504-11-11-701	
Scotland	55.96	-3.26	*A. sylvaticus*	SC4	504-1-1-702	
Scotland	57.46	-4.24	*A. sylvaticus*	SC5	504-2-2-702	
Scotland	57.46	-4.24	*A. sylvaticus*	SC6	504-4-4-702	
Scotland	57.46	-4.24	*A. sylvaticus*	SC7	504-5-5-702	
Scotland	55.96	-3.26	*A. sylvaticus*	SC8	504-6-6-702	SC1
France	48.55	-3.40	*A. sylvaticus*	FR32	504-7-7-702	FR16
Spain	43.25	-4.05	*A. sylvaticus*	ES9	504-1-1-703	
Plate 4						
England	52.57	-2.86	*A. flavicollis*	EN9	504-2-2-703	
France	45.78	3.09	*A. flavicollis*	FR41	504-4-4-703	
Poland	50.90	20.93	*A. flavicollis*	PL10	504-5-5-703	
Poland	54.10	17.47	*A. flavicollis*	PL11	504-6-6-703	
France	45.78	3.09	*A. flavicollis*	FR42	504-7-7-703	
England	54.20	-2.95	*A. sylvaticus*	EN10	504-10-10-703	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
England	51.13	1.26	*A. sylvaticus*	EN11	504-11-11-703	EN22
Portugal	41.32	-8.73	*A. sylvaticus*	PT1	504-12-12-703	
Spain	43.25	-4.05	*A. sylvaticus*	ES10	504-1-1-704	
England	52.57	-2.86	*A. flavicollis*	EN12	504-2-2-704	
France	45.78	3.09	*A. flavicollis*	FR43	504-4-4-704	
Poland	50.90	20.93	*A. flavicollis*	PL12	504-5-5-704	
Poland	53.74	19.96	*A. flavicollis*	PL13	504-6-6-704	
England	51.13	1.26	*A. sylvaticus*	EN13	504-7-7-704	
England	52.37	-1.96	*A. sylvaticus*	EN14	504-10-10-704	
Poland	52.61	15.86	*A. sylvaticus*	PL14	504-11-11-704	
Portugal	41.32	-8.73	*A. sylvaticus*	PT2	504-12-12-704	
Spain	43.25	-4.05	*A. sylvaticus*	ES11	501-1-1-701	
England	52.24	0.99	*A. flavicollis*	EN15	501-2-2-701	
Poland	54.13	19.42	*A. flavicollis*	PL15	501-4-4-701	
Poland	53.85	18.95	*A. flavicollis*	PL16	501-5-5-701	
Poland	53.74	19.96	*A. flavicollis*	PL17	501-6-6-701	
England	51.13	1.26	*A. sylvaticus*	EN16	501-7-7-701	
England	52.37	-1.96	*A. sylvaticus*	EN17	501-10-10-701	
Poland	52.61	15.86	*A. sylvaticus*	PL18	501-11-11-701	
Portugal	41.32	-8.73	*A. sylvaticus*	PT3	501-12-12-701	
Sweden	56.00	14.10	*A. sylvaticus*	SE14	501-1-1-702	
England	52.24	0.99	*A. flavicollis*	EN18	501-2-2-702	
Poland	54.13	19.42	*A. flavicollis*	PL19	501-4-4-702	
Poland	53.85	18.95	*A. flavicollis*	PL20	501-5-5-702	
Lithuania	55.52	21.11	*A. flavicollis*	LT19	501-6-6-702	
England	51.13	1.26	*A. sylvaticus*	EN19	501-7-7-702	
England	52.37	-1.96	*A. sylvaticus*	EN20	501-10-10-702	
Poland	53.37	16.61	*A. sylvaticus*	PL21	501-11-11-702	
Portugal	41.32	-8.74	*A. sylvaticus*	PT4	501-12-12-702	
Sweden	56.00	14.10	*A. sylvaticus*	SE15	501-1-1-703	
England	51.23	0.90	*A. flavicollis*	EN21	501-2-2-703	
Poland	54.13	19.42	*A. flavicollis*	PL22	501-4-4-703	
France	43.16	6.62	*A. sylvaticus*	FR44	501-5-5-703	
Poland	49.19	22.43	*A. flavicollis*	PL23	501-6-6-703	
England	51.13	1.26	*A. sylvaticus*	EN22	501-7-7-703	EN11
England	52.37	-1.96	*A. sylvaticus*	EN23	501-10-10-703	

Country	**Latitude**	**Longitude**	**Species**	**Sample**	**Adapters**	**DUP**
Poland	54.29	19.50	*A. sylvaticus*	PL24	501-11-11-703	
Portugal	39.70	-7.30	*A. sylvaticus*	PT5	501-12-12-703	
Sweden	56.00	14.10	*A. sylvaticus*	SE16	501-1-1-704	
England	51.23	0.90	*A. flavicollis*	EN24	501-2-2-704	
Poland	50.90	20.93	*A. flavicollis*	PL25	501-4-4-704	
Poland	53.92	16.57	*A. flavicollis*	PL26	501-5-5-704	
Lithuania	54.66	24.83	*A. flavicollis*	LT120	501-6-6-704	
England	54.20	-2.95	*A. sylvaticus*	EN25	501-7-7-704	
Scotland	57.46	-4.24	*A. sylvaticus*	SC9	501-10-10-704	
Spain	43.17	-6.50	*A. flavicollis*	ES12	501-11-11-704	ES27-ES24
Portugal	39.70	-7.30	*A. sylvaticus*	PT6	501-12-12-704	PT7
Sweden	56.00	14.10	*A. sylvaticus*	SE17	502-1-1-701	
France	45.78	3.09	*A. flavicollis*	FR45	502-2-2-701	
Poland	50.90	20.93	*A. flavicollis*	PL27	502-4-4-701	
Poland	50.22	19.58	*A. flavicollis*	PL28	502-5-5-701	
Poland	52.73	15.00	*A. flavicollis*	PL29	502-6-6-701	
England	54.20	-2.95	*A. sylvaticus*	EN26	502-7-7-701	
Spain	43.25	-4.05	*A. sylvaticus*	ES13	502-10-10-701	
Wales	53.21	-4.37	*A. sylvaticus*	WL1	502-11-11-701	
Spain	43.17	-6.50	*A. sylvaticus*	ES14	502-12-12-701	
England	52.24	0.99	*A. flavicollis*	EN27	502-1-1-702	
France	45.78	3.09	*A. flavicollis*	FR46	502-2-2-702	
Poland	50.90	20.93	*A. flavicollis*	PL30	502-4-4-702	
France	42.48	3.13	*A. sylvaticus*	FR47	502-5-5-702	
Spain	43.17	-6.50	*A. flavicollis*	ES15	502-6-6-702	ES27-ES12
England	54.20	-2.95	*A. sylvaticus*	EN28	502-7-7-702	
Poland	52.09	15.15	*A. sylvaticus*	PL31	502-10-10-702	
Wales	53.21	-4.37	*A. sylvaticus*	WL2	502-11-11-702	
Spain	43.17	-6.50	*A. sylvaticus*	ES16	502-12-12-702	
Spain	43.17	-6.50	*A. sylvaticus*	ES17	502-1-1-703	
Spain	43.17	-6.50	*A. sylvaticus*	ES18	502-2-2-703	
Spain	40.14	-5.25	*A. sylvaticus*	ES19	502-4-4-703	
Spain	40.14	-5.25	*A. sylvaticus*	ES20	502-5-5-703	
Spain	40.14	-5.25	*A. sylvaticus*	ES21	502-6-6-703	
Spain	40.14	-5.25	*A. sylvaticus*	ES22	502-7-7-703	
Spain	40.96	-5.97	*A. sylvaticus*	ES23	502-10-10-703	

Country	Latitude	Longitude	Species	Sample	Adapters	DUP
Spain	40.98	-5.97	*A. sylvaticus*	ES24	502-11-11-703	
Spain	40.02	-1.82	*A. sylvaticus*	ES25	502-1-1-704	
Spain	39.79	-2.15	*A. sylvaticus*	ES26	502-2-2-704	
Wales	53.21	-4.37	*A. sylvaticus*	WL3	502-4-4-704	
Wales	53.21	-4.37	*A. sylvaticus*	WL4	502-5-5-704	
Spain	43.17	-6.50	*A. flavicollis*	ES27	502-6-6-704	ES15-ES12
Portugal	39.70	-7.30	*A. sylvaticus*	PT7	502-7-7-704	PT6

Chapter D

Appendix Chapter 5: *Apodemus sylvaticus*

D.1 Demultiplexing results

Detailed results about the percentage of clones, chimeric sequences, presence of adapters, ambiguous barcodes and ambiguos rad-cutsites, shown in total number and in percentege: *https://github.com/Marisa89/Apodemus_Europe/blob/master/Demultiplexing_results.csv*

Results including the total number of samples, ambiguous rad-cutsites, low quality sequences and total number of retained reads per sample: *https://github.com/Marisa89/Apodemus_Europe/blob/master/Results_per_sample.csv*

D.2 *Apodemus sylvaticus*

D.2.1 Coverage

The data used to build the graph is available at: *https://github.com/Marisa89/Apodemus_Europe/blob/master/sylvaticus/Coverage_sylvaticus_europe.csv*

D.2.2 Selection best parameters

Tables including the number of assembled loci, polymorphic loci and SNPs per each iteration of the m, M or n parameters, per sample, as well as the population results obtained calling SNPs shared between the 40%, 60% and 80% of the samples, are available at: *https://github.com/Marisa89/Apodemus_Europe/blob/master/sylvaticus/Table_selection_best_parameters_Asylvaticus_Europe.xlsx*

D.2.3 List of samples kept after filtering

Samples kept for *Apodemus sylvaticus* analyses:
https://github.com/Marisa89/Apodemus_Europe/blob/master/sylvaticus/Samples_retained_analysis_Asylvaticus.csv

D.2.4 Genetic diversity

Polymorphic Sites, polymorphic loci, number of individuals, observed heterozigosity (Ho), Expected heterozygosity (He), Nucleotide diversity(Π) and inbreeding coefficient (FIS) average values for all *Apodemus sylvaticus* populations.
https://github.com/Marisa89/Apodemus_Europe/blob/master/sylvaticus/Genetic_diversity_pop_sylvaticus_europe.csv

D.2.5 F_{ST}

F_{ST} pairwise comparison between all *Apodemus sylvaticus* populations and average distance between them: *https://github.com/Marisa89/Apodemus_Europe/blob/master/sylvaticus/Fst_pop_sylvaticus_europe.csv*

D.3 *Apodemus flavicollis*

D.3.1 Coverage

The data used to build the graph is available at: *https://github.com/Marisa89/Apodemus_Europe/blob/master/flavicollis/Coverage_flavicollis_europe.csv*

D.3.2 Selection best parameters

Tables including the number of assembled loci, polymorphic loci and SNPs per each iteration of the m, M or n paramters, per sample, as well as the population results obtained calling SNPs shared between the 40%, 60% and 80% of the samples, are available at: *https://github.com/Marisa89/Apodemus_Europe/blob/master/flavicollis/Table_selection_best_parameters_Aflavicollis_Europe.xlsx*

D.3.3 List of samples kept after filtering

Samples kept for *Apodemus flavicollis* analyses:
https://github.com/Marisa89/Apodemus_Europe/blob/master/flavicollis/Samples_retained_analysis_Aflavicollis.csv

D.3.4 Genetic diversity

Polymorphic Sites, polymorphic loci, number of individuals, observed heterozigosity (Ho), Expected heterozygosity (He), Nucleotide diversity(Π) and inbreeding coefficient (FIS) average values for all *Apodemus flavicollis* populations.
https://github.com/Marisa89/Apodemus_Europe/blob/master/flavicollis/Genetic_diversity_pop_flavicollis_europe.csv

D.3.5 F_{ST}

F_{ST} pairwise comparison between all *Apodemus flavicollis* populations and average distance between them:
https://github.com/Marisa89/Apodemus_Europe/blob/master/flavicollis/Fst_pop_flavicollis_europe.csv

D.4 Code

Code used for plotting the results:
https://github.com/Marisa89/Apodemus_Europe/tree/master/Code