THE POPULATION GENETICS OF RED SQUIRRELS IN A FRAGMENTED HABITAT

BY
REBECCA TODD, BSc

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

DECEMBER 1999

CONTENTS

CHAPTER CONTENTS

ABSTRACT

The genetics of eight small red squirrel (*Sciurus vulgaris* L.) populations in northern Belgium is investigated by analysing variation in a section of the mitochondrial control region and five microsatellite loci. The full sequence of the mitochondrial control region in red squirrels is determined and is compared to that of other mammals. The isolation of microsatellite loci is also described.

The eight fragment populations are compared with two large Belgian populations and one large population in the Bavarian Forest, Germany. Virtually no variation is found in the control region within any of the Belgian squirrels, although the German population is found to be highly variable. However, the Belgian and German samples show comparable levels of diversity at the microsatellite loci. The lack of variation in the control region of the Belgian squirrels suggests that they have lost variation, due to either selective or demographic pressures. A combination of a bottleneck and metapopulation structuring could lead to reduced diversity levels and explain the observed patterns of variation.

The recent effects of habitat fragmentation and population expansion can be seen in the microsatellite data. Three of the fragment populations show evidence of recent bottlenecks or founder events, probably due to the recent colonisation of these areas by squirrels from nearby expanding populations. Estimations of F_{ST} and R_{ST} show that there is some differentiation among the populations, but none of the populations are significantly differentiated from any of the others. There is no correlation between genetic differentiation and geographic distance indicating that migration is influenced by other factors as well as distance. The fragment populations all contain more allelic diversity than would be expected in populations of their size at mutation-drift equilibrium. Migration between the populations appears to be maintaining nuclear variation and counteracting the effects of random genetic drift.

ACKNOWLEDGEMENTS

Thanks must obviously go to Prof. David Parkin for selecting me to do a research project in his laboratory and supervising me through it. Dr. Jon Wetton provided joint supervision and invaluable help whilst working as a post-doctoral researcher in the laboratory; he was greatly missed when he left.

A huge amount of gratitude to Dr. Bruce Winney for his unending support and guidance throughout my time in Nottingham, particularly over the last two years. He has provided constant patient encouragement, inspiration by example and helpful critical advice.

A great big thank you to Nick Harvey, and everyone else in the "birdlab" over the last four years, for being supportive and putting up with me towards the end as I felt the pressure of time!

I must thank Goedele Verbeyen at the University of Antwerp for collecting samples from the fragment populations, passing on the demographic data she painstakingly collected and spending three days with me in the field, showing me the populations and introducing me to the red squirrels.

I would also like to acknowledge the earlier hard work of Luc Wauters, who initiated the study of the fragmented populations of red squirrels around Antwerp, and Sibylle Münch for providing samples from the German population of Waldhausser.

Lastly, but by no means least, thanks to my parents for being sympathetic and supportive throughout and to my sister, Monica, without whose ear at the other end of the telephone line, life would have been harder.

This thesis is dedicated to my grandparents, John and Margaret Todd,
who took great interest in my education whilst they were alive
and who I know would have been very proud.

CHAPTER ONE:

INTRODUCTION

1.1 INTRODUCTION

Habitat fragmentation is an increasingly important subject in conservation biology as more and more habitats are affected by human activities. Natural habitats are constantly being destroyed or altered and large continuous areas of a single habitat are being broken up into smaller areas or fragmented as a result of the increasing demands being made on the land by its human occupants. The survival of a species in a fragmented habitat depends on both demography and genetics. There have been many investigations into the effects that habitat fragmentation have had on the demography of a wide range of organisms, including the red squirrel (Verboom and van Apeldoorn 1990; van Apeldoorn *et al.* 1994; Wauters 1994a and 1994b; Andrén 1994; Delin 1996). In recent years, attention has turned to the genetic effects of habitat fragmentation on a wide range of organisms including plants (Young *et al.* 1996; Aldrich *et al.* 1998; Young *et al.* 1999), amphibians (Hitchings and Beebee 1998; Seppä and Laurila 1999), insects (van Dongen *et al.* 1998), and mammals (Dallas *et al.* 1995; Gaines *et al.* 1997; Becher and Griffiths 1998). This study will investigate the genetic effects of habitat fragmentation on a species for which the demographic and socioecological effects of population subdivision are well understood.

The aim of this project is to determine the effects that habitat fragmentation has had on the genetics of a group of red squirrel populations in northern Belgium. Two types of molecular marker, the control region of the mitochondrial genome and a set of microsatellite loci, were analysed using polymerase chain reaction techniques. This allowed the genotypes of the individuals occupying each population to be determined from small samples of ear tissue collected in a non-invasive manner. The amount of variation in each population was quantified and the relatedness of the populations determined. Conclusions regarding the effects of habitat fragmentation on the genetics of these populations have been drawn.

This introduction will start by discussing the possible structures of populations restricted to habitat fragments and the effects these structures will have on the demography of the populations. The genetics of populations in habitat fragments is essentially the genetics of small populations and the relevant population genetic theory is introduced in the following section. Many studies have been carried out on red squirrels in fragmented habitats around Europe; the results of these investigations are reviewed, in the light of some general red squirrel biology, and the study populations are described. In the final sections, the molecular markers used to investigate the genetics of the populations are introduced and what is known of their characteristics and evolution is reviewed.

1.2 HABITAT FRAGMENTATION

Population subdivision affects different organisms in various ways. Many species naturally form aggregations such as herds or flocks, so forming subpopulations and some habitats are naturally patchy forcing a degree of subdivision onto their residents. In modern times, however, more and more species are being affected by habitat fragmentation due to human activities, in ways that are not natural to them.

A fragment is a detached, isolated or incomplete part broken away from a whole. Habitat fragmentation is just that: the breaking up of an area of habitat into several isolated patches (Andrén, 1994). Fragmentation occurs when a barrier is created that reduces the movement of individuals between areas of their habitat. Barriers may take many forms, such as roads, a fence or a new field, and a barrier for one species may not be a barrier for another (Kirby 1995). In the same way, a "corridor" (a link between two habitat patches that facilitates movement of individuals between them) may not be effective for all species.

The main effects of the fragmentation of a species' habitat are the direct loss of habitat and loss of continuity between areas (Andrén, 1994). Individuals become restricted to smaller patches and become isolated from individuals occupying other patches. The survival of the species in the remaining patches depends on the size of the fragments, how near they are to other patches, and the impacts of the new surrounding environment (Rolstad 1991).

Species occupying a fragmented habitat commonly take on a "metapopulation" structure. A metapopulation has been described as "a population of populations" (Hanski and Gilpin 1991), it is a system of local populations connected by dispersing individuals. The concept was developed by Levins in 1970 in order to develop a mathematical model to investigate ideas about group selection. He wanted to know whether a gene that was disadvantageous in a single population could persist in a larger system of many populations, a metapopulation. He showed that it was only possible in unusual circumstances (Hanski and Gilpin 1991).

The metapopulation concept was developed from MacArthur and Wilson's theories of "island biogeography" (developed in the mid-1960s) which were used to explain the "species-area relationship". This is the long recognised pattern that the number of different species seen on an island increases with the area of the island (Diamond and May 1981) until the maximum capacity is reached. The theory of island biogeography describes the dynamics of island occupation and shows that the number of species on an island is the result of a dynamic equilibrium between immigration of new colonising species and the extinction of previously

established ones (Mackenzie *et al.* 1998). This theory can be applied to a single species (Hanski and Gilpin 1991) and describes the situation where there is a large area of habitat, the "mainland", with many "islands" of habitat nearby. The islands show a dynamic equilibrium of extinction and recolonisation from the mainland. Crucial to this model is the concept of "turnover" which describes the extinction of populations and recolonisation of empty habitat patches (Gilpin 1991).

The metapopulation concept takes the idea further and describes a situation where populations are equally subject to turnover, extinction occurs at random and recolonisation occurs by dispersal from any of the extant populations that are themselves also subject to turnover (Hanski and Gilpin 1991). These models share the same basic processes of extinction and colonisation of islands of habitat. Both assume that suitable habitat patches are isolated from one another by hostile habitat and that individuals of each species only use one patch, each patch having a local population (Andrén, 1994). The rate of extinction relative to colonisation determines the proportion of inhabited patches and the turnover rate (van Apeldoorn *et al.* 1992). They differ in that the populations in Levins' metapopulation are equally vulnerable to extinction whereas a population on the mainland of the mainland-island model is very unlikely to go extinct.

In reality, these two models describe extremes of a continuum and most groups of populations probably have dynamics encompassing elements of both models. Van Apeldoorn *et al.* (1992) described two extreme situations that could be envisaged for a fragmented habitat. At one extreme, all the patches are more or less equal in size, quality and surroundings, and have equal extinction and recolonisation rates. Therefore local populations go extinct at random and dispersal of individuals is equal in all directions. This is the metapopulation model of Levins'. At the other extreme, small fragments are scattered around larger areas which have a very low probability of extinction. The small areas have a much higher chance of extinction and dispersal is therefore not equal in all directions. This is the "mainland-island" model.

Yet another possible dynamic system experienced by groups of populations is the "source-sink" structure (van Apeldoorn *et al.* 1992). Such a system occurs when many individuals regularly occupy suboptimal habitats or "sinks" (Gaggiotti and Smouse 1996). In these areas reproduction is insufficient to balance mortality and the population is maintained by immigration from "source" areas where natality is greater than mortality, so the imbalances in the two types of area are balanced by dispersal (Dias 1996). This is similar to the "mainland-island" structure but differs in that "islands" in the mainland-island model do not necessarily

have a natality deficit, they may have more deaths than births occasionally due to stochastic factors but this is not an intrinsic state. In sink areas the habitat is less favourable and the population is constantly dependent on immigrants from a source, the source populations need not necessarily be larger than the sinks, they simply occupy better habitat areas and so are more productive (Hanski and Gilpin 1991).

An extreme example of a "source-sink" population structure is seen in the checkered white butterfly of central California. Every winter the butterflies retreat to a small "source" area where they overwinter successfully, then in the spring emigrants set up many small colonies ("sinks") that can be a considerable distance from the source. These populations only persist for a few generations until they become extinct at the onset of the next winter (Gaggiotti and Smouse 1996). A less extreme example is that of the Amargosa vole that inhabits desert marsh; it is widespread throughout the habitat most of the time but on occasional flood years it retreats to isolated hilltops (Gaggiotti and Smouse 1996). Most species living in interacting populations do not fit one type of structure, but probably display elements of all the systems described, but in all of them the importance that migration has on their risk of extinction is evident.

Nowadays, all of these population structures are usually considered to be kinds of metapopulations, although they do not fit Levins' original simpler description in which all populations are equivalent and extinction and colonisation are essentially stochastic processes (Dias 1996). More realistic single species models of metapopulations that allow for variations in the area of patches and distances between patches have been developed recently and a more appropriate general definition would be of an ensemble of interacting populations with a finite lifetime (i.e. an expected time to extinction) (Hanski and Gilpin 1991). However, the central concept remains that the persistence of a species in an area depends on the colonisation and extinction of individual populations (Andrén, 1994).

1.2.1 The demographic effects of habitat fragmentation

To understand the effects of habitat fragmentation on a species, the effects of local population extinction and recolonisation must be understood and the importance of migration considered. The effects of habitat fragmentation are a complex mix of changes in habitat quality and dispersal patterns, so understanding the habitat selection and the dispersal success of a species is key to understanding the effects of fragmentation (Matthysen *et al.* 1995).

The most important changes resulting from fragmentation are habitat loss, reduction in the size of patches available for occupation and an increasing distance between patches. The relative importance of each of these depends on the degree of fragmentation (Andrén, 1994). The most important effect is a reduction in the total number of individuals in the metapopulation which is a direct result of habitat loss. Andrén (1994) reviewed the effects of habitat fragmentation on a range of bird and mammal species. He found that the decline in metapopulation size was directly proportional to the loss of habitat when a reasonable proportion of the original habitat remains. At some threshold (he suggested the threshold may be between 10% and 30% of habitat remaining for birds and mammals) smaller and more isolated patches appear and the area and isolation of the habitat fragments becomes important, strengthening the effects of habitat loss so that the metapopulation size decreases at a faster rate. As he pointed out, metapopulation theory predicts a threshold for fragmentation below which a metapopulation can no longer exist because the extinction rate increases with decreasing patch size and at the same time, the colonisation rate decreases with increasing patch isolation.

The survival of a species in a particular patch depends on the probabilities of extinction and colonisation. The maintenance of an individual population is dictated by both demography and genetics, the demographic effects being both of environmental and demographic stochasticity and the genetic effects resulting from inbreeding and a loss of genetic variation (Gilpin 1987). Genetics may be important to long term survival but it is demography that is likely to be of more immediate importance to a small population (Lande 1988). Demographic stochasticity describes differences between individuals in the probabilities and rates of survival and reproduction, whereas environmental stochasticity refers to environmental changes that affect many individuals at once, for example, a bad winter may have a negative affect on the dynamics of the whole population (Gilpin 1987; Lande 1988). Extreme forms of environmental stochasticity are catastrophes such as volcanic eruptions and disease pandemics which may drastically affect the whole metapopulation, possibly even wiping it out. Overall, environmental stochasticity is thought to have a more dominant effect than demographic variation, as populations experience environmental fluctuations constantly (Lande 1988).

The demographic stability of the population occupying a patch is greatly influenced by patch "quality": higher quality patches are occupied by larger, more stable populations. Habitat fragments will vary in overall habitability for a species, the closer the area is to ideal for the species the higher its quality. Van Apeldoorn *et al.* (1992) used 13 variables to assess the quality of habitat fragments for the bank vole (measuring its area, percentage grass, shrub

and herb cover, number of tree trunks appropriate for nesting, height of herb level etc.) and showed that the higher quality patches were occupied by more voles with less fluctuation in numbers. They found that the population dynamics of the bank vole metapopulation resembled a source-sink structure with some woodlots being occupied during the reproductive season, but not all year round.

The presence of a large proportion of edge will reduce the overall quality of a patch and the smaller the patch area, the greater the proportion of habitat that is close to the edge. There is often a deterioration of habitat quality near boundaries with other habitats (Lande 1988) and many reasons for this can be envisaged (Kirby 1995). A wooded area experiences changes in tree density, light and humidity near edges. There may be changes in the distribution of species, especially invertebrates, near edges and the possible invasion of species from neighbouring habitat. Areas near roads may suffer from noise, chemicals and rubbish polluting the environment. All these factors may make edges of habitat patches less desirable and reduce the survival chances of a species in a patch which has a large proportion of edge.

The probability of the survival of the population in an area is increased by immigration both to boost population numbers and to recolonise an area if the population goes extinct. The rate of migration depends on the dispersal ability of the species and the geographic arrangement of the habitat fragments. Larger areas produce more migrants but the probability of immigration falls exponentially with increasing distance between patches (Gilpin 1991). The type of habitat that must be crossed between patches will affect the immigration success rate (van Apeldoorn *et al.* 1992).

A very neat experiment in habitat fragmentation was carried out by Gonzalez *et al.* (1998a). They investigated the distribution and abundance of species in a miniature, moss-based ecosystem that was artificially fragmented, with and without corridors to aid migration. This habitat was easy to manipulate on an appropriate scale relative to the size and dispersal abilities of the animal populations that live in it and it contains well-known, easily sampled communities of microarthropods. Initially they tested the effects of fragmentation by artificially manipulating the habitat into two areas, one a large continuous patch and the other an area of small patches or fragments. They found that after one year the distribution and abundance of the species in the fragmented habitat had declined significantly, exactly as metapopulation theory predicts. The set-up for the second experiment is illustrated in figure 1.1, in this experiment they added two more areas to investigate the effect of corridors. They created a third area that contained patch fragments which were connected by corridors of habitat to

facilitate movement between the patches. To control for the additional habitat area that these corridors provided, the fourth area contained fragments and pseudo corridors that provided additional area but no real connection between patches. In the truly fragmented area and in the pseudo-corridor area 41% of the species had suffered extinction after 6 months, whereas only 14.5% of species in the fragments connected by corridors were extinct. The extant species in this area also showed decreases in abundance, but this would be expected from the reduction in area. These experiments illustrate how habitat fragmentation results in a decline in the distribution and abundance of animal species, but that the creation of corridors to facilitate migration between areas significantly ameliorates these effects. The maintenance of dispersal is vital for maintaining the abundance of species in fragmented habitats.

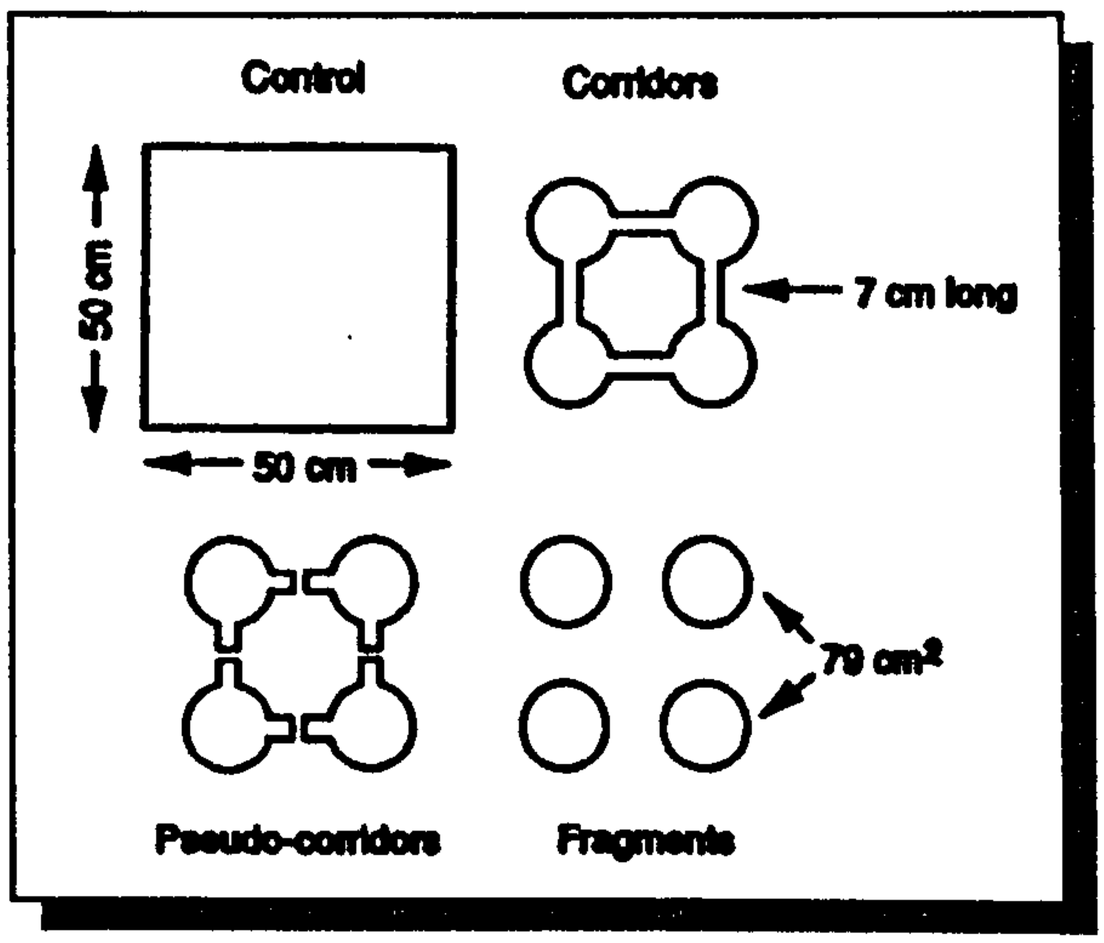


Figure 1.1: Diagramatic representation (not to scale) of the design of the second experiment carried out by Gonzalez *et al.* to investigate the effects of corridors on the species content of habitat fragments (reproduced from Gonzalez *et al.* 1998a).

However, the influence of migration on a metapopulation can be more complex. Lindenmayer and Lacy (1995) carried out a simulation study of the impacts of population subdivision on the mountain brushtail possum (*Trichosurus caninus*) in south-eastern Australia. They found that in some circumstances with very small population sizes, migration can have a negative effect, depressing rates of population and metapopulation growth and increasing the fluctuations in growth rate. They attributed this to the inherent demographic instability of metapopulations containing very small populations. For example, in such an unstable metapopulation there are likely to be many unoccupied patches and migrants may arrive in these patches and find few opportunities to breed, so dispersal events may reduce the reproductive rates of the metapopulation. Dispersing individuals are less likely to be replaced by immigrants when the populations are all very small, so the overall fecundity of the populations is reduced and the extinction rates increase due to the effects of migration. Such an extreme situation is perhaps not very likely in reality unless the metapopulation is very close to extinction anyway. Generally, they found that migration has a positive effect, increasing population and metapopulation growth, reducing fluctuations in growth rate and reducing population extinction rate.

They also confirmed that a single large population always performs better than an ensemble of populations of the same total size, even with high rates of migration. The negative effects of subdivision included lower and more fluctuating rates of population and metapopulation growth, increased probability of population extinction, a shorter expected lifetime of each population and smaller metapopulation sizes. The larger the metapopulation the less severe the effects. It is important to bear in mind when considering these results, however, that this was a simulation study that made many assumptions about the environment and populations and inevitably portrayed a simplified story.

Matthysen *et al.* (1995) reviewed the results of studies investigating the real population ecology of several different species, including red squirrels, small passerine birds and butterflies, restricted to forest fragments. They found that overall fragmentation had a more profound effect on dispersal patterns and population structure than on reproductive output and survival. There was little evidence for lowered reproduction, condition or survival in any of the species, despite the fact that these habitats have probably been fragmented for centuries. They did find that the populations in fragmented habitats generally showed reduced densities (numbers per unit area) compared with populations in continuous habitats, which they thought was due to immigration deficits resulting from changes in the immigration/emigration balance. Changes were seen in the timing of dispersal and in habitat selection. Dispersers between fragments showed a higher mortality than those in a continuous habitat and they were generally less successful. These results are useful because dispersal patterns and disperser success is likely to be of great importance to the survival of a metapopulation.

The species most vulnerable to the effects of habitat fragmentation are interior rather than edge species with large minimum area requirements, poor mobility outside the habitat and a dependence on habitats that are or were continuous (Kirby 1995). Highly mobile species will not suffer greatly provided that their total area is not greatly reduced. For example, the great spotted woodpecker needs a large area of woodland (about 8-10 hectares) to breed but this can be composed of several separate patches (Kirby 1995). Species with little or no mobility, such as plants and invertebrates, can exist in a very small area and may not be affected at all. It is the species that fall between these two categories, with varying abilities to spread by crossing gaps or using corridors, that are most vulnerable to habitat fragmentation (Kirby 1995). It is interesting to note that, in Britain, red squirrels are considered to be amongst the most vulnerable of mammals (Kirby 1995).

1.3 THE GENETICS OF FRAGMENTED POPULATIONS

Investigating the genetics of populations generally involves measuring the amount and distribution of genetic variation across populations. Essentially, genetic variation can be measured as "allelic diversity", which refers to the number of different alleles at a given locus in the population, or "heterozygosity", which is the proportion of individuals that are heterozygous at a locus. The average heterozygosity is determined over several loci. Heterozygosity also represents the probability that two alleles taken at random from the population are different (Ayala 1982).

These two measures are very closely related; the more alleles there are in a population, the greater the potential for heterozygosity. Yet, the two can behave very differently: one population may have fewer alleles at a locus (lower allelic diversity) than another population, but the same proportion of individuals may be heterozygous at that locus. An extreme example of the potential discrepancy between allelic diversity and heterozygosity could be a population that usually reproduces by self-fertilisation, as some plants do, when most individuals would be homozygous (Ayala 1982). Different individuals, however, may carry different alleles if the locus is variable, so allelic diversity could be considerable whilst heterozygosity was virtually zero. Therefore, measures of heterozygosity alone may not be an accurate representation of the variation in a population. Both measures depend upon the resolving power of the analysis technique and the type of marker analysed (section 1.5), so absolute values quantifying variation are only useful for direct comparisons with other populations or groups where the same techniques and marker have been employed (Gilpin 1991).

Nei defined a third measure of genetic variation that could be considered a compromise between allelic diversity and average heterozygosity, called "gene diversity" (Nei 1987). This is one minus the sum of the squared frequencies of all the alleles and reflects both the number of alleles and the evenness of their frequencies (Lacy, 1995). Low frequency alleles are not well accounted for in this measure so, as with heterozygosity, it may not represent the actual variation present in a population accurately if many of the alleles are at a very low frequency. Gene diversity is also known as "expected heterozygosity" because it is equal to the actual observed heterozygosity in large, randomly mating populations with no outside influences on allele frequencies, such as selection, mutation and migration (i.e. at Hardy-Weinberg equilibrium, see section 5.1.2.1).

The genetics of fragmented populations is effectively the genetics of small populations. When the area became fragmented each remaining isolated population would have experienced a dramatic reduction in size or a "bottleneck". If a new population was founded in an uninhabited patch by a few individuals migrating to the area then it has experienced the "founder effect". This has the same initial effects as a bottleneck as both result in very small effective sizes. A population may then remain small or it may rapidly increase in size. If habitat fragmentation has restricted the species to small areas, it will probably be limited in size, but after a founding event it may expand as it fills a new area.

Whilst the populations remain small their genetics will be dominated by the effects of "random genetic drift" (usually referred to simply as "drift"). This is a process of random sampling of alleles from generation to generation. Each new generation of a breeding population contains only a selection of the possible combination of gametes from the parental generation and which alleles survive into the next generation is a process of chance. The result is chance fluctuations in allele frequencies and the eventual loss of many alleles. The smaller the population, the greater the effect of drift and the faster the loss of genetic variation.

Population bottlenecks, the founder effect and random genetic drift all act on populations to remove variation. Mutation and migration are homogenising factors, acting to restore some of that variation. How these different processes theoretically interact is the subject of the following sections.

1.3.1 Bottlenecks and founder events

When a large population experiences a sudden dramatic reduction in size (a "bottleneck") much of the allelic diversity present in the original population is lost. A small random sample of the alleles from the original gene pool survives to form the bottlenecked population. The same thing occurs when a random sample of individuals forms a new population in a founder event, this is called the "founder effect". In the case of nuclear genes, each individual in the new sample can only carry a maximum of two alleles at each locus; for mitochondrial genes, each individual only carries one. Hence, if a population is reduced to only a few individuals, then only a few alleles will remain. Nei *et al.* (1975) were the first to study the theoretical effects of a bottleneck and to examine exactly how alleles are lost from reduced populations. The processes have since been modelled and examined in great detail by Maruyama and Fuerst (1984 and 1985).

Which alleles and how many are maintained not only depends on the size of the bottleneck but also on the genetics of the original (source) population. At most variable loci in a large outbred population, there exist some alleles at high frequency and many alleles at very low frequencies, the rare alleles, but few alleles at intermediate frequencies; this results in a U – or J – shaped allele frequency distributions (when the incidence of each allele frequency class is plotted). Therefore, some alleles are more likely than others to be preserved during a bottleneck (Chakraborty *et al.* 1980; Fuerst and Maruyama 1986). Rare alleles, present at low frequencies, are most likely to be lost during a random sampling process such as a bottleneck or founder event. The more variable the original population, the greater the reduction in allelic diversity during a bottleneck. The number of alleles found in a population after a bottleneck or founder event depends on the original population size, the new population size, the mutation rate and variability of the locus concerned and the number of generations since the reduction in size (Maruyama and Fuerst 1985).

A dramatic bottleneck will lead to a rapid loss of alleles from the population, but conversely, much of the heterozygosity will be retained (Maruyama and Fuerst 1985; Fuerst and Maruyama 1986; Allendorf 1986). When frequency of loci is plotted against heterozygosity the distributions tend to be L – shaped showing that most loci in a population have little or no variability (Fuerst and Maruyama 1986). This pattern becomes more pronounced as the average heterozygosity in the population decreases. Most loci have one high frequency allele and several rare alleles, as suggested by the J-shaped allele frequency distributions. It is the rare alleles found at these loci that are quickly lost during random sampling processes. Yet, the loss of these rare alleles would have little effect on average heterozygosity as most individuals were already homozygous at these loci. A few loci are highly polymorphic with several different alleles at high frequency and they would be expected to remain polymorphic even during a severe bottleneck (Fuerst and Maruyama 1986). It is these high frequency alleles that contribute most to the measure of heterozygosity and so maintain its levels during bottlenecks and founder events.

As a result, a small population that exists immediately after a bottleneck will be expected to show a deficiency of alleles with respect to relatively unchanged levels of heterozygosity (Maruyama and Fuerst 1985). If that population then remains small the ongoing effects of random genetic drift will continue to erode both the remaining allelic diversity and heterozygosity. The reduction in heterozygosity is less dramatic than that of allelic diversity in the early stages of a bottleneck, but, once a population is reduced to a small size, heterozygosity is also gradually removed over the following generations as there are fewer alleles to maintain it.

Allelic diversity and heterozygosity will continue to be lost after a bottleneck whilst the population remains small. The loss of heterozygosity is most rapid in the early generations after the bottleneck but it will eventually reach a minimum level from which it can slowly recover (Nei *et al.* 1975). If the population size recovers quickly after a bottleneck then the heterozygosity will not be greatly affected, but once it has been reduced it takes a very long time to recover. The number of generations required for heterozygosity to return to its original level can be roughly calculated as the reciprocal of the mutation rate, e.g. if the mutation rate is 10^{-8} (as in *Drosophila*) then the number of generations required is approximately 10^{8}, or 10^{7} years if one year represents 10 generations (Nei *et al.* 1975).

Allelic diversity is most dramatically lost in the first generation affected by the bottleneck but loss continues during the following generations whilst the population is small, at a lessening rate. As the population recovers, the allelic diversity is expected to increase much faster than the heterozygosity. An expanding population, recovering from a bottleneck in its recent past, can be expected to show an excess in number of alleles relative to the heterozygosity in the same way as a bottlenecked population shows an allelic deficiency (Nei *et al.* 1975; Maruyama and Fuerst 1984). This is due to alleles being reintroduced into the population by mutation and at the early stages these new alleles are at low frequency, contributing little to the measure of heterozygosity. As with the allelic deficiency, this is only a temporary state and heterozygosity eventually catches up with allelic diversity. This situation may be seen in a founder population, if the population rapidly expands into the new area after the founder event.

Bottlenecks are frequently inferred when a population is shown to have low genetic variability, but it is rarely possible to compare the historical population with the post bottleneck population to directly measure the change in genetic variation. However, such a study was carried out on the greater prairie chicken (*Tympanuchus cupido*) in Illinois, USA (Bouzat *et al.* 1998). This species was known to have been reduced to less than 50 individuals in 1993 after a decline that started in the middle of the 19th century. A comparison of the allelic diversity found in museum samples and current samples of the species at six polymorphic microsatellite loci revealed that the alleles present in the current population represented a subset of the alleles found in the historical population with certain specific alleles having been lost. This kind of comparison is the most powerful test for a bottleneck and Luikart *et al.* (1998) have developed a formal statistical test for examining data from the same population before and after a bottleneck.

The classic and oft quoted example of a population bottleneck is the South African subspecies of cheetah (*Acinonyx jubatus jubatus*) (O'Brien *et al.* 1983; O'Brien *et al.* 1985). A low level of variability at allozyme loci and in the major histocompatibility complex led to the conclusion that these cheetahs are genetically depauperate and that this was probably due to an historical bottleneck. However, comparisons with other carnivores showed that, although the cheetah has relatively low levels of variation, such levels are not unique in cat family (Merola 1994). It may be that this species is naturally low in genetic variation, as was proved to be the case for the eastern barred bandicoot (*Perameles gunnii*). No polymorphism was detected in an isolated and endangered population of this species in Australia so it was concluded that the population had suffered a bottleneck. However, a later study of a large population of the same species in Tasmania showed that this species had naturally low levels of variation and there was no real evidence for a bottleneck (Merola 1994). A lack of variation alone does not prove a bottleneck, comparisons with non-bottlenecked populations must be made.

In another famous example of a bottleneck, the comparison with non-bottlenecked populations has been made but, more importantly, the bottleneck itself is documented. The northern elephant seal (*Mirounga angustirostrus*) was heavily expoited during the 19th century, it was hunted heavily until the 1860s when it became rare. In 1884, 153 were killed and no more were seen until 1892 since when the population has recovered slowly to about 15,000 individuals in 1960 and 120,000 in 1980 (Hoelzel *et al.* 1993). In 1974, Bonnell and Selander discovered that this species showed genetic homogeneity in a survey of 24 allozyme loci in 159 samples. This result was confirmed by Hoelzel *et al.* (1993) who extended it to 43 allozyme loci in 67 individuals, in addition they found only two control region haplotypes in their sample. In contrast, the southern elephant seal, which has also been persecuted but not to such a great extent showed similar variability at allozyme loci to other large mammal populations (Hoelzel *et al.* 1993).

Hoelzel *et al.* (1993) used a simulation model to determine the size and duration of the bottleneck required to explain the lack of mitochondrial variation in this species, given the demographic data available. They determined that the bottleneck must have been of less than 30 seals for a 20 year duration or, at the other extreme, less than 20 seals for a single year. Hedrick (1995) has reassessed their results and confirmed that a bottleneck of this size would explain the loss of mitochondrial variation, but a more severe bottleneck would be required to explain the loss of nuclear variation at the allozyme loci.

Sometimes founder events can be studied directly. When a new population is established following a founder event, it may be possible to compare the original and the new populations to assess directly their genetical differences. Ardern *et al.* (1997) examined the effects of two founder events involving five or fewer New Zealand robins (*Petroica australis australis*). As they assessed the levels of variation in each population using multilocus fingerprinting it was hard to accurately quantify allelic diversity, but the number of fragments can be taken to indicate the number of different alleles and that was shown to be reduced in the new populations. The heterozygosity however was slightly reduced but remained at high levels despite the severity of the bottleneck. As the founder populations had quickly expanded into their new areas, the heterozygosity was never significantly reduced. They pointed out that "a sudden population contraction followed by rapid recovery to large population size does not necessarily have profound consequences on the reduction of genetic variation."

A maintenance of heterozygosity was also seen in translocated populations of the Laysan finch (*Telespiza cantans*) in Hawaii, studied by Fleischer *et al.* (1991). Approximately 100 individuals were introduced to Southeast island in 1967 and the population has since grown to about 500 individuals. With that many individuals in the founder population, the reduction in variation would only be expected to be slight but they actually found higher levels of allozyme heterozygosity in the founder population than in the source. This is probably because the number of individuals translocated to the new island was so high, the sample may have had a higher average heterozygosity by chance than the source population. The population has since expanded so heterozygosity would not have been expected to decrease after the bottleneck. A similar pattern was found by Leberg (1992) in experimentally bottlenecked populations of the eastern mosquitofish (*Gambusia holbrooki*). He found that in many populations single and multiple-locus heterozygosity actually increased after a founder event, despite there only being six founders in each population. Again, this is due to chance sampling events; a large number of loci must be examined for heterozygosity to reliably indicate a bottleneck has occurred. The inclusion of allelic diversity comparisons would be more reliable.

Allelic diversity and heterozygosity are both valid measures of genetic variation in a population but they each respond very differently to the dramatic population size reductions that result from bottlenecks and founder events. The loss of alleles is largely dictated by the bottleneck size whereas the loss of heterozygosity is determined more by the rate of population growth after the bottleneck (Nei *et al.* 1975). A substantial bottleneck may initially leave the heterozygosity only slightly reduced, if reduced at all, as much of this variation is

maintained in the loci possessing several alleles at intermediate frequencies in the population. Chance effects due to the random sampling of individuals from the source population can mean that resulting populations actually have higher heterozygosities than the original populations (Leberg 1992). The allelic diversity however, will be affected to a degree that is directly proportional to the size of the bottleneck. Allelic diversity, therefore, is much more sensitive to the effects of bottlenecks and founder events than average heterozygosity (Allendorf 1986; Fuerst and Maruyama 1986; Ardern *et al.* 1997; Brookes *et al.* 1997).

1.3.2 Random genetic drift

Populations which are small suffer the ongoing consequences of random genetic drift. The effects of drift on a population are best described in terms of a sampling process. Each generation a random sample of the gametes in the population forms the following generation. This leads to fluctuations in gene frequency that are random in direction. The size of the change in gene frequency depends on the size of the breeding population in the following generation. The larger the number of individuals making up the next generation the closer the agreement between the allele frequencies of the parental and progeny generations (Ayala 1982).

The effects of drift can be quantified in terms of the variance in allele frequencies and for this, the "effective population size" must be determined. The effective population size (*Ne*) (Wright 1931) is the size of an "idealised population" that would give rise to the same variance in allele frequencies (the same rate of drift) as that observed in the natural population (Lacy 1995; Wang and Caballero 1999). The idealised population is a concept credited to both Fisher and Wright, pioneers in population genetic theory. It is essentially an unrealistically simplified population which allows theoretical study by removing the complicating factors, such as non-random mating and overlapping generations, that are normally found in nature. It allows the effects of drift, mutation, selection and migration to be studied and quantified. An idealised population is "a monoecious population with constant size over discrete generations, random mating including selfing in random amounts, an equal probability of contributing gametes to the next generation from different parents with respect to autosomal loci without mutation and selection" (Wang and Caballero 1999).

Clearly, no natural population will satisfy these criteria. Many characteristics of natural populations violate the requirements of an idealised population, including fluctuations in the sex ratio, fluctuations in population size, overlapping generations, variance in reproductive success and non-random mating. These violations mean that the effective population size is

usually much smaller than the actual population size. This is a very important consideration in conservation as it means that an apparently large population could behave genetically like a very small one (Gilpin 1991). In a metapopulation, one of the most important violations of the idealised population concept is non-random mating, individuals are more likely to reproduce with others within their area than from different patches.

The main consequences of genetic drift for small populations are a loss of variation within populations, differentiation between populations and increased homozygosity (Falconer 1981). Over each generation, drift leads to a gradual loss of variation as alleles are randomly lost from the population. Low frequency alleles have a higher probability of being lost, but drift is a random process and any allele, by chance, may become fixed in the population (reach a frequency of one, where no other allele at that locus is present in the population). The probability of a particular allele becoming fixed is equal to its initial frequency, so the lower its frequency the more likely it is to be lost. The rate of allele loss from a population of a constant size and losing variation at a constant rate is (Fuerst and Maruyama 1986):

$n(n-1)/2Ne$ where n is the number of alleles remaining in the population and Ne is the effective population size,

whereas the loss of heterozygosity occurs at a rate of:

$1/2Ne$

This means that until the number of alleles at a locus is reduced to 1.6, allelic diversity is reduced faster than heterozygosity, but when the number of alleles falls below 1.6 then it is heterozygosity that is lost at the faster rate.

Drift is random and acts independently on separate populations. Thus, over time, the effects of drift will be different in each population resulting in differences in their genetic composition or "differentiation". This is an important effect when a population becomes subdivided into many populations as it means that variation may be maintained within the metapopulation as differences in the allelic compositions between the populations. Increased homozygosity also occurs over time as alleles are lost. When the number of alleles in a population is reduced, there is a greater chance of an individual mating with another individual carrying the same allele at a particular variable locus rendering the offspring homozygous at that locus. This is also the main genetic consequence of inbreeding.

1.3.3 Inbreeding

Inbreeding is the mating of individuals who share common ancestors (Lacy, 1995). The degree of relatedness of individuals whose ancestry is derived from the same population depends on the size of the population: the smaller the population the less remote in time are the common ancestors. Pairs mating at random in a small population are more closely related than individuals in a large one (Falconer 1981). If two individuals share a common ancestor they may both carry replicates of the same allele derived from that ancestor; the more recent the common ancestor, the more likely it is that they will share alleles that are "identical by descent" (identical as a result of being copies of the same ancestral gene in one ancestral individual). If the two individuals then mate, they may both pass on the identical allele to the offspring, so that the offspring may be homozygous for the allele. The smaller the population the more likely it is that relatives will mate. The genetics of small populations are often considered in terms of the rate of inbreeding (Falconer 1981).

"Inbreeding depression" refers to the loss of fitness that has often been observed to be associated with inbreeding in populations. Studies of domestic, laboratory and zoo populations have shown evidence of inbreeding depression (Keller *et al.* 1994) and lines propagated by continued brother-sister mating or selfing become sterile or inviable after several generations (Lande 1988), but evidence for its occurrence in the wild is lacking. It is often difficult to detect accurately the occurrence of inbreeding in natural populations (Pusey and Wolf 1996) and it is hard to rule out other causes of fitness variation. Inbreeding depression is thought to be the result of an increase in the number of homozygotes for deleterious recessive alleles that are rare in a large population (Lande 1988; Pusey and Wolf 1996) as homozygotes for deleterious alleles usually show reduced viability or fecundity. For example, the grizzly bear population in Yellowstone National Park has had a very small effective population size so is likely to be subject to inbreeding and reduced litter size in this species has been suggested to be the result of inbreeding depression (Gilpin 1987). Another proposed cause of loss of fitness due to increased homozygosity is a loss of "heterosis" (Pusey and Wolf 1996; Young *et al.* 1996). This is increased fitness due to a heterozygotic genotype and, if a heterozygote at a particular locus is fitter than the homozygote, the fitness of the population is reduced with increasing homozygosity.

Jiménez *et al.* (1994) made an experimental study of inbreeding depression in the white-footed mouse (*Peromyscus leucopus noveboracensis*). Inbred and noninbred descendants of wild caught mice were released back into the field from which their ancestors originated and their survival compared. Inbred mice showed continued weight loss and reduced survivorship in the wild when compared to their outbred counterparts, illustrating the possible negative

effects of inbreeding. In another study, island populations of song sparrows (*Melospiza melodia*) were followed during two population crashes caused by severe winters and inbred individual showed a much lower rate of survival than outbred individuals (Keller *et al.* 1994). Both studies suggest that inbreeding can have a detrimental effect on individual fitness, but mechanisms for the reduction in fitness remain undefined.

It has often been suggested that the "unmasking" of recessive deleterious alleles by inbreeding may lead to them being purged from the population by natural selection, eventually leaving the population in greater genetic health than before (Pusey and Wolf 1996; Young *et al.* 1996) but this is a controversial theoretical idea and the relationship between inbreeding and selection is not well understood (Hedrick 1994).

1.3.4 Migration

Migration is the movement of individuals from one population to another with the intention of settling in the new location. The rate of migration may be the most important variable affecting fragmented populations. Migration (also referred to as "gene flow") is an homogenising factor, countering the effects of random genetic drift. It was Sewall Wright's view that small amounts of gene flow between isolated small populations are necessary for long term persistence (Mills and Allendorf 1996).

Very small amounts of gene flow can have powerful genetic consequences on the degree of divergence between populations (Mills and Allendorf 1996). In theory, differentiation between populations can be lost completely after one generation of total migration or random mating (Wang and Caballero 1999); a lack of differentiation and random mating is referred to as "panmixia". At the other end of the scale there would be no migration between populations and they would differentiate over time due to the effects of drift. Usually there is some, but not total, migration between populations in a fragmented habitat and an equilibrium between the counteracting effects of drift and migration may be possible.

If there is no gene flow between populations in a metapopulation, genetic variation can be maintained in the system as differences between the populations. If there is panmixia and no differentiation between populations, then variation in the form of interpopulation differences will be lost. It requires very little gene flow to eliminate between population differences. A famous rule of thumb in conservation biology is that one immigrant per generation is all that is required to maintain a healthy population, regardless of its size (Varvio *et al.* 1986; Mills and Allendorf 1996). This is counterintuitive, as it might be predicted

that a larger population would required more migration, but it results from the opposing effects of drift. The larger the population the less drift the population experiences so the fewer migrants are required to counterbalance it. A smaller population suffers more drift so requires more migration to offset it. Hence, divergence is dependent on the amount of migration and a similar amount of migration is required by all populations regardless of size (Mills and Allendorf 1996). The recommendation of one migrant per generation represents a trade-off between the effects of migration reducing between population variation but also increasing within population variation, one migrant per generation does not produce panmixia (Mills and Allendorf 1996).

In some circumstances, habitat fragmentation can lead to increased gene flow over distances. Comparison of fragmented and non-fragmented populations of the tree *Acer saccharum* in Ohio, USA showed that the fragmented populations showed less genetic divergence between cohorts and in fact, interpopulation gene flow had increased after fragmentation (Young *et al.* 1996). This may be because fragmentation meant that greater gene flow was required for reproduction to be successful and so local population structure was broken down. This is only likely to be the case in very particular circumstances and is likely to be an unusual result.

1.3.5 Mutation and selection

Random genetic drift is a "dispersive" process, tending to scatter the gene frequencies in a stochastic manner. Migration, mutation and selection are "systematic" (Falconer 1981) or "deterministic" (Ayala 1982) processes tending to bring populations towards stable equilibria in which the amount of variation at each locus remains unchanged whilst circumstances remain the same. Without these opposing forces, alleles would all eventually be fixed or lost from the population (unless the population was of infinite size) and there would be no genetic variation. Mutation is the ultimate source of all new variation, migration and selection change the allele frequencies of the variation present.

In populations of finite size, gene frequency changes are governed primarily by drift if:

$$4Nex << 1$$

where *Ne* is the effective population size and *x* is the mutation or migration rate, or the selection coefficient (which represents the strength of selection) (Ayala 1982).

Therefore, if *4Nex* is around one or more then the gene frequencies are determined by the deterministic process. For example, for a mutation rate of 10^{-4} (which is typical for microsatellites (Ellegren *et. al.* 1995; McDonald and Potts 1997; Schlötterer, 1998b)) a population size of about 2500 would be required for mutation to have a greater effect than drift on gene frequencies. In small populations, mutation has a negligible effect on the genetics as drift is so strong. For the recommended migration rate of one immigrant per generation (where the migration rate is the number of immigrants/ population size), *4Nex*=4, so migration at this rate will effectively balance the effects of drift. Overall, this means that relatively small amounts of gene flow can have a significant effect on small populations but mutation is unlikely to have any effect at all because of its slow rate (Wang and Caballero 1999). In theory, populations can reach an equilibrium between drift and mutation either when the population is large enough for mutation to have a significant effect, or when heterozygosity is so low that further losses can be balanced by mutation (Lacy 1987).

Selection will only have an effect if it is very strong (Lacy 1987 and 1997). Selection will tend to bring an allele to a high frequency if it is advantageous, to a low frequency if it is deleterious, or to an intermediate frequency if the heterozygote is the genotype of the highest fitness. The effects of selection depend on the trait in question, usually it will reduce variation by favouring particular alleles but balancing selection (when the heterozygote is favoured) tends to preserve variation (Lacy 1987).

The marker loci being employed in this study are likely to be selectively neutral; however, this does not necessarily mean that they are totally immune to the effects of selection. If a neutral locus is linked to a locus that is affected by selective pressure, then its gene frequency in the population will also depend on the same selection process. This is called "hitch-hiking" as the neutral locus moves to a new frequency dependent on a different, but nearby, locus (Ballard and Kreitman 1995). This is particularly important when analysing the mitochondrial genome as it is in effect one single large linked locus because it does not undergo recombination (section 1.5.1.1). If a particular locus has been affected by selection it should be recognisable when comparing the data from that locus with several other loci, therefore it is important to consider many loci in a population genetic investigation.

1.3.6 Overview

"At all levels of organisation life depends on the maintenance of a certain balance among its factors." Sewall Wright (1932)

When an area of habitat is fragmented the populations inhabiting the area suffer a reduction in area and numbers. This will necessarily lead to a loss of variation, both in terms of alleles and heterozygosity. The formation of a metapopulation will involve bottlenecks where populations are reduced in size and founder events when areas of habitat are recolonised after extinction events. Both will involve large losses of allelic diversity. A severe bottleneck can exert a disproportionate effect on the long term effective population size as the effective size is given by the harmonic mean of a fluctuating population size (Gilpin 1991; Amos and Harwood 1998).

Small populations are constantly affected by random genetic drift which erodes both allelic diversity and heterozygosity. The smaller the population the faster the loss of variation. Heterozygosity can be lost directly as a function of a reduction in the number of alleles at a locus or as a result of the increased likelihood of in inbreeding in small populations. It is this latter process that is likely to have the most effect on heterozygosity levels (Young *et al.* 1996).

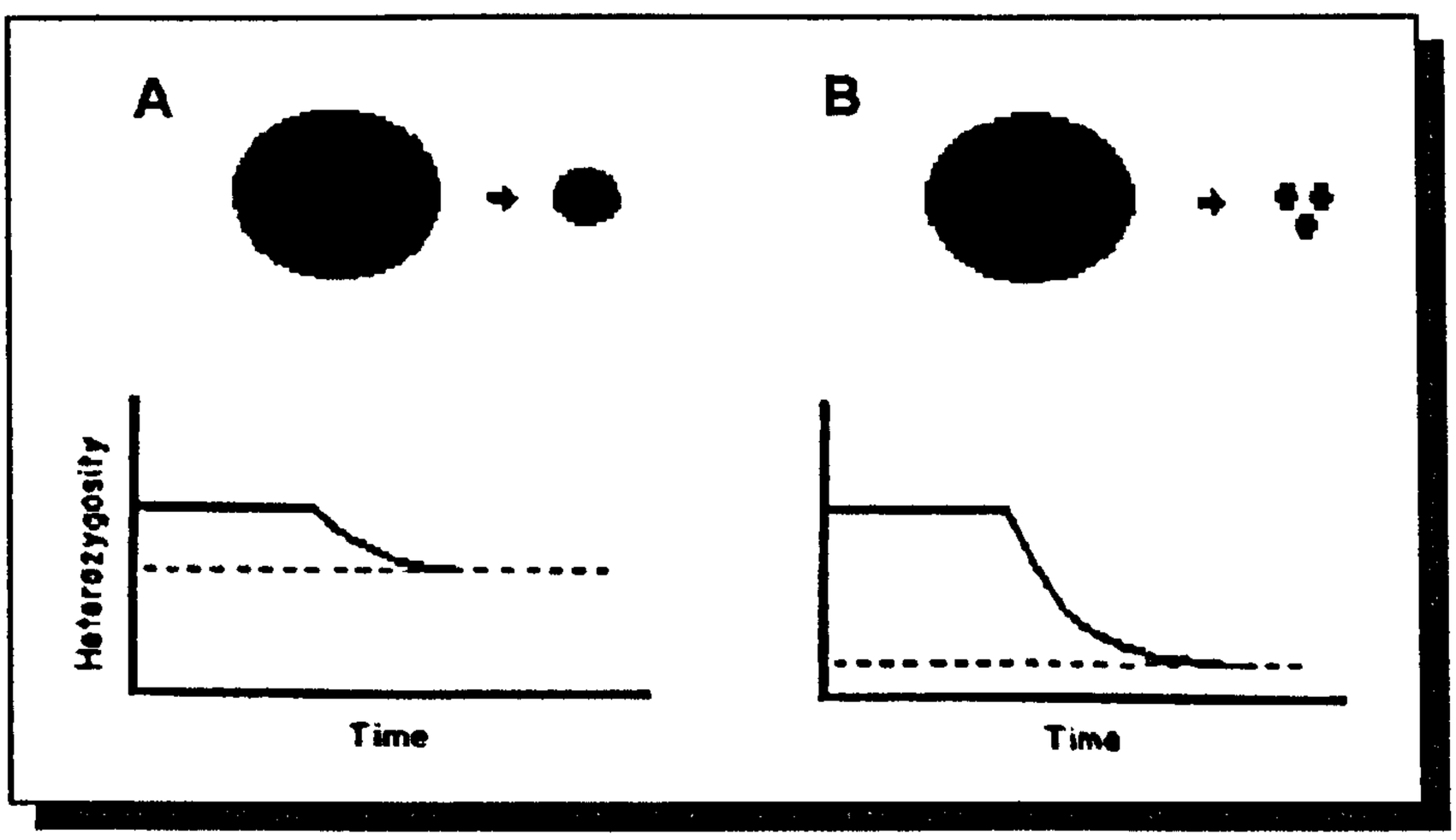


Figure 1.2: The two cases of heterozygosity loss modelled by Gilpin (1991). A: Habitat reduction to a single patch. B: Habitat reduction to three smaller patches of the same total area. The greater the fragmentation the greater the loss of heterozygosity. (Reproduced from Gilpin (1991).)

Gilpin (1991) modelled the effects of metapopulation structure on population genetics and showed that variation is lost from fragmented populations, the more fragmented the faster and greater the loss. This is illustrated in figure 1.2. This model considered a very severe metapopulation structure where the only migration was in the form of the recolonisation of areas after extinction events. Recolonisation/extinction events greatly reduce the potential for differentiation between populations and the metapopulation overall becomes very low in variation (Lacy and Lindenmayer 1995). It has been suggested that extreme metapopulation structuring of this sort may explain the low levels of genetic variation found in some natural populations, such as the cheetah (Pimm *et al.* 1989; Gilpin 1991; Hedrick 1996; but see O'Brien 1989), usually assumed to be the result of severe bottlenecks. Population fragmentation without turnover, on the other hand, can lead to the preservation of diversity in the form of between population differences.

The genetics of small populations in a fragmented landscape is dominated by the conflicting actions of drift and migration. The divergence between populations and the loss of variation from populations depends on the migration rate. If gene flow is sufficiently high then fragmented populations behave as a single unit with a total size equal to the total number of individuals in the system. If gene flow is restricted then the populations will diverge (Amos and Harwood 1998). The opposing forces of migration and drift may settle down to equilibrium levels where they balance each other. This equilibrium will be influenced by the population structure which, in natural metapopulations, is likely to be complicated and hierarchical (Wang and Caballero 1999). Some populations will exchange more migrants than others and so be more closely related.

Mutation is unlikely to have a role in the genetics of small populations because of its slow rate, but it is the ultimate source of all variation in the system and is the only source of new alleles (unless immigrants from separate metapopulations arrive carrying new alleles). Selection is also unlikely to play much part due to the dominating effects of drift, unless it is very strong (Lacy and Lindenmayer 1995). This leads to the possibility of slightly deleterious alleles drifting to fixation in small populations as a result of the ineffectiveness of selection (Lande 1988; Lacy and Lindenmayer 1995).

Rarely have studies investigating the effects of habitat fragmentation looked at both demographic and genetic effects. One such study is that of Lacy and Lindenmayer (1995), the second part of a larger project which also looked in more detail at demographic effects alone. They simulated the impacts of habitat fragmentation on the mountain brushtail possum (*Trichosurus caninus* Ogilby) in south eastern Australia, including in the simulation

the effects of demographic stochasticity as well as all the factors affecting the genetics, such as migration. They found, as was to be expected, that the larger the metapopulation the slower the loss of heterozygosity and migration offset this effect, although losses from fragmented populations connected by high rates of migration were always greater than from one large population. Including demographic stochasticity in their simulation led to some important observations. Demographic stochasticity results in fluctuations in population size across generations, greatly reducing the effective population size and increasing the effects of drift. This increase in the rate of drift meant that in their simulations, alleles were lost too fast from the fragmented populations for significant levels of variation to be maintained as differences between populations. Migration does reduce the effectiveness of a fragmented structure to maintain variation in the form of between population differences but in the long term the advantages of migration are not just in maintaining within population variation. Migration means that populations suffer fewer extinctions and metapopulations remain larger and more stable than in scenarios without migration.

Theory and simulations such as these can be used to predict the "expected" behaviour of metapopulations but they do not predict the actual fate of real genes and populations (Lacy 1987). Many of the processes involved are stochastic so there will be a lot of chance and unpredictability as to the actual fates of metapopulations. The distribution of genetic variation in populations occupying a fragmented habitat depends on the distribution of variation before fragmentation, the size and number of fragmented populations, the time since isolation and the species socioecology including, crucially, dispersal patterns and abilities (Pope 1996). The survival of populations and metapopulations in the short term will be determined by demographic stochasticity, migration is very important to population survival. Genetics may be important for the long term survival of metapopulations.

1.4 THE EURASIAN RED SQUIRREL (*Sciurus vulgaris* L.)

Squirrels belong to the order Rodentia (class: Mammalia); the name rodent is derived from the Latin *rodere* meaning "to gnaw" and all rodents have a successful generalised body plan with grinding cheek teeth and efficient gnawing front teeth. The success of the rodents can also be attributed to fast breeding and their ability to adapt quickly to changing surroundings and colonise new habitats. There are three types of rodent, the myomorphs (mouse-like rodents), hystricomorphs (porcupine-like rodents) and the sciuromorphs (squirrel-like rodents). The Sciuridae (the squirrels) is one family of sciuromorphs and contains flying squirrels, ground dwelling squirrels and tree squirrels. There are two genera of tree squirrels found in the cold and temperate forests of the northern hemisphere: *Sciurus* and *Tamiasciurus*. The Eurasian red squirrel, *Sciurus vulgaris*, belongs to the *Sciurus* genus of tree squirrels and its range is shown in figure 1.3; it probably has the largest range of all the squirrels. Many sub-species have been suggested for the red squirrel even though its distribution is mostly continuous and the taxonomic status of any sub-species is not clear (Gurnell 1987).

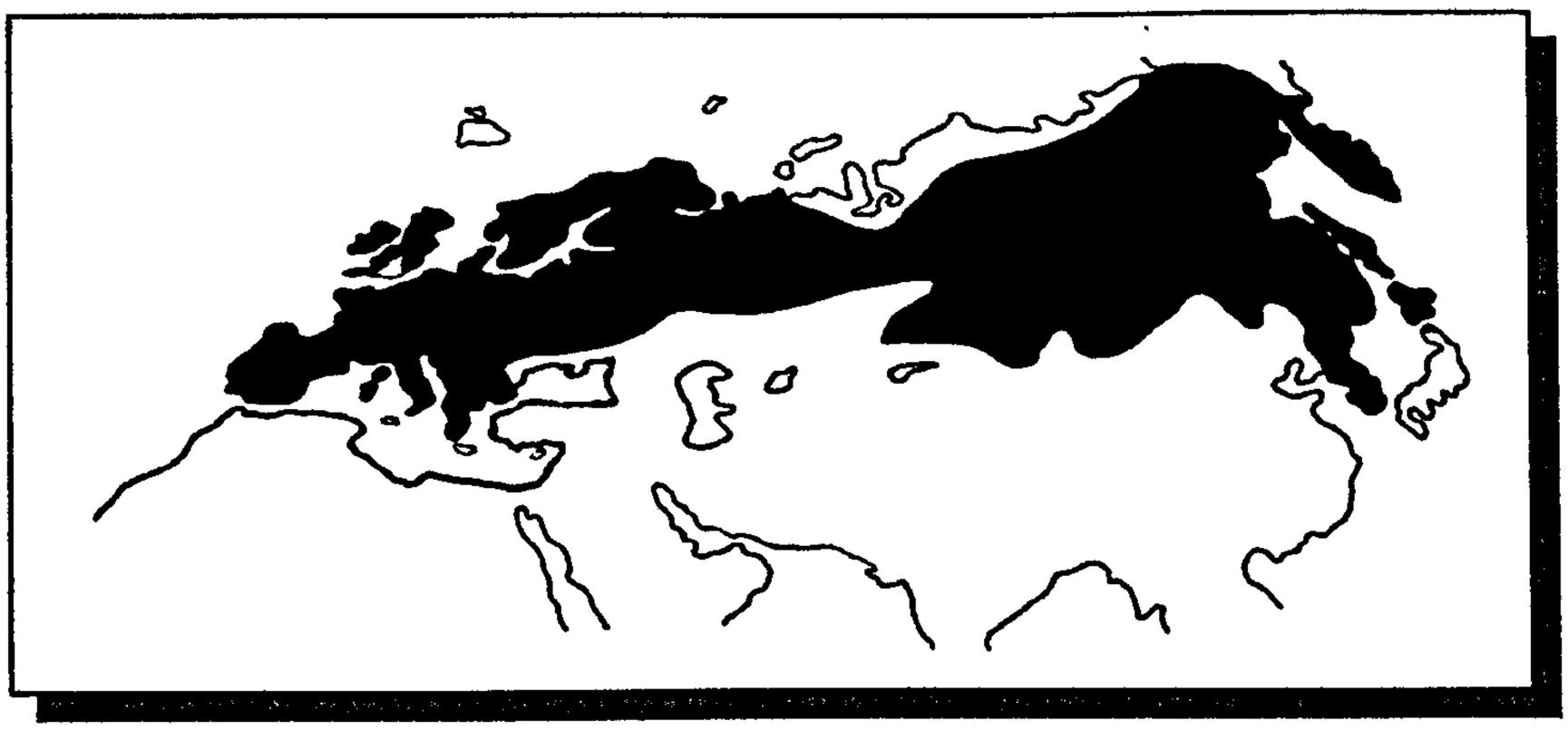


Figure 1.3: The distribution of *S.vulgaris* (based on Gurnell, 1987)

The large and bushy tail is the most distinguishing feature of squirrels (figure 1.4). The name *Sciurus* comes from the ancient Greek *skia* meaning "shadow" and *oura* meaning "tail" so literally a squirrel lives in the shadow of its tail. Red squirrels can have three different dorsal coat colours: red, melanic black and intermediate brown, all have white undersides and totally white squirrels have been known. The frequencies of these forms vary from region to region and are thought to be related to climate and genetic background (Gurnell 1987). There are moults in spring and autumn and the winter coat is bushier and more intense in colour. It has been suggested that coat colour may be a form of crypsis, the dorsal colours blending in with the surrounding forest and the white underside, seen from below, blending in with the sky above.

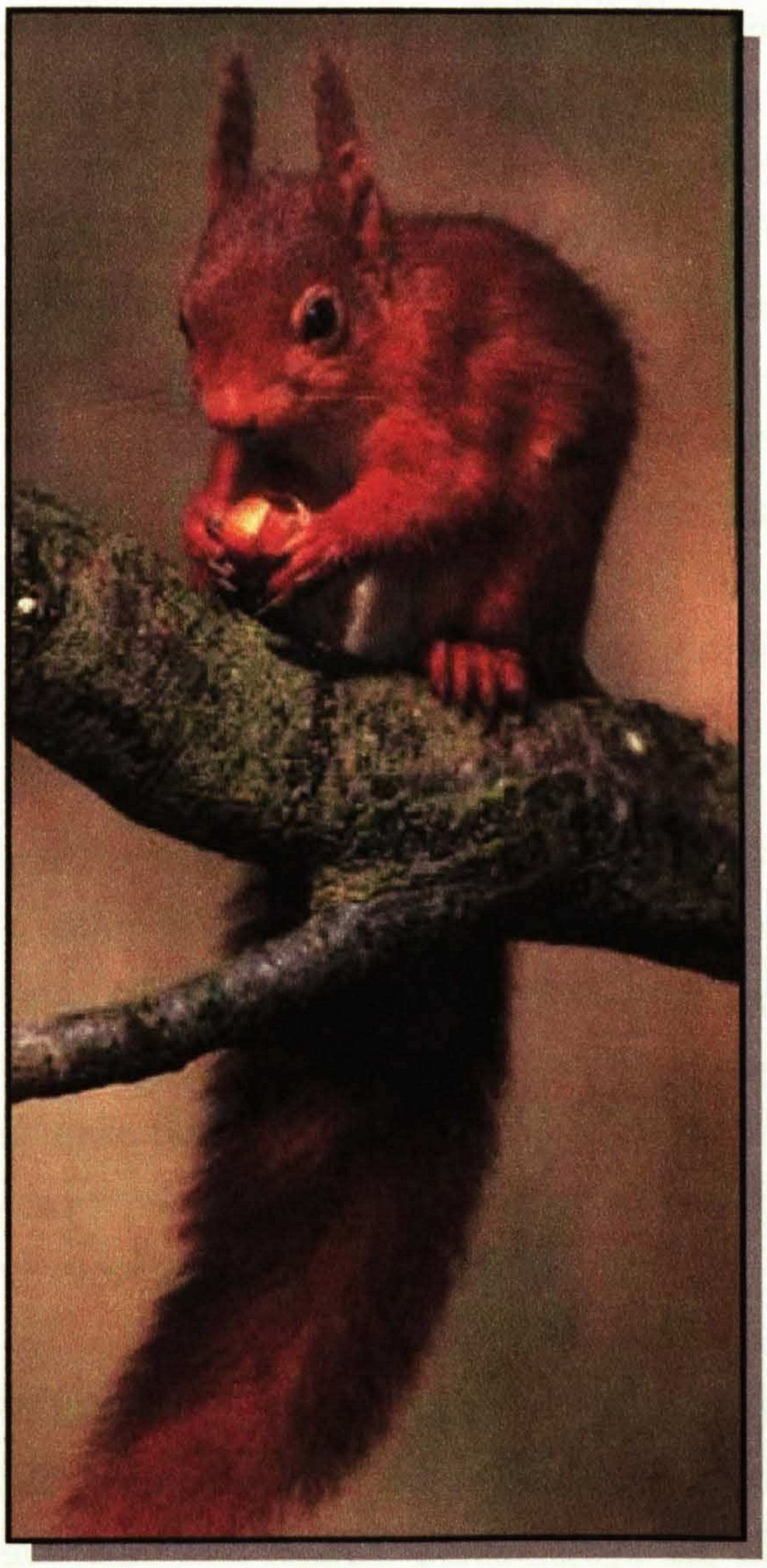


Figure 1.4: The Eurasian red squirrel, *Sciurus vulgaris* L. (reproduced from the BBC Wildlife magazine, May 1997)

Squirrels are granivore-herbivores, generally relying on a few important foods associated with the trees in which they live. Red squirrels mainly feed on a combination of spruce and pine seeds, coniferous buds and fungi. The seeds have the highest nutrient content and the squirrels tend to feed mostly on these. They are active during the day, spending most of their time in the canopy, searching for seeds. *S. vulgaris* can be found in pure coniferous, mixed or deciduous woodland but they show a strong preference for coniferous and mixed woods (Verboom and van Apeldoorn 1990). Percentage conifer content can be used as a measure of habitat quality (Verboom and van Apeldoorn 1990).

Each squirrel has a home range, the area over which the animal moves during its normal daily activity, and home ranges of individuals may overlap considerably. The home range behaviour seen in European red squirrels is consistent with a low density but highly predictable food supply as used by the squirrels; the squirrels cover an area that contains enough food for their needs. More extreme territorial behaviour, where territories are defended against other individuals, is expected if the benefits from holding a territory outweigh the costs of defence (Wauters and Dhont 1992). This would be the case either if the food supply was of a higher density and a smaller area could supply all the animals needs, or if the seed supply was more erratic and more intense hoarding in a particular limited area was required, as is the case for some North American squirrel species (Wauters and Dhont 1992). In Europe, tree seeds are available most of the year, either as cones on the trees or as hoarded supplies (Wauters and Dhont 1992), so European squirrels show home range behaviour but do not defend exclusive areas.

Range size is affected by population density, food supply and habitat quality. In lower quality deciduous woods, red squirrels require larger home ranges to compensate for the lower habitat quality. Red squirrels use part of their home range, the "core area", very intensively, whilst the edges are only visited occasionally. The core area usually comprises between 40 – 50% of the total home range (Wauters and Dhont 1992).

Space use is also affected by gender, age and season of the year. As would be expected in a species with a polygynous mating system, males and females have different behavioural strategies to improve mating success and this is reflected in their space use (Wauters and Dhont 1992). Male squirrels have larger home ranges that overlap considerably, whereas females have smaller ranges that overlap less. Both sexes have a dominance hierarchy that determines the distribution of the individuals. At around 18 months old, the largest and heaviest females, which are the dominant females, settle in areas with the most food and small, defendable core areas can ensure sufficient food supply. Subordinates overlap with the range edges of one or more dominant females. Dominant males have larger ranges than subordinates, overlapping the ranges of as many receptive females as possible. Therefore males invest in larger areas to gain access to females and reproductive opportunities, whereas females invest in defensible core-areas with sufficient food resources to raise their young successfully.

Within their home ranges, squirrels have nests which can be dreys, dens or holes in the ground. Dreys are twig and leaf structures built in the trees usually about half way up so they are sheltered and hidden. Winter dreys are elaborate waterproof structures made with an outer layer of interwoven twigs and a softer inner lining of moss, bark, leaves, lichen and other materials. Summer dreys are simpler structures and may only be a twig and leaf platform on which the squirrels can rest. Dens are holes in the main trunks of trees and may be lined with soft material and used for nesting or simply used for escape. Red squirrels tend to use more dreys than dens and, as they mostly forage in the canopy, they rarely use holes in the ground. Drey counts can be used to estimate the number of red squirrels in an area; it has been estimated that 2.7 to 3 dreys were found for every squirrel in coniferous forests in Belgium (Gurnell 1987).

1.4.1 Population dynamics

The breeding season of squirrels depends on temperature and food availability. If temperatures are high and there is plenty of food, breeding may start early in the year. Time from conception to independence of young is only 4 months so in a good year there is time for two litters to be produced, but only some females will do so. Only the dominant females in a population will raise offspring (Wauters and Dhont 1992). Females of 9 months of age are sexually mature, but they start breeding later in the year, whereas females of two years and older breed earlier and may produce two litters.

Only 15-25% of young survive to the second year of life, after which year to year survival averages at 50-70% but depends greatly on circumstances. Squirrel populations suffer loses from emigration, predation and other causes of death such as disease and, commonly these days, road kills. Occasionally epidemics will wipe out large numbers of individuals, for example the intestinal disease coccidiosis killed nearly one million red squirrels in Finland in 1943-4 (Gurnell 1987). It is hard to tell what kills red squirrels although they appear not to die of cold directly but of poor body condition and food supply associated with bad weather. The maximum life span of tree squirrels is between 5 and 10 years, but as few as 1% may reach old age.

Young squirrels, both males and females, commonly "disperse" away from their natal ranges. Juveniles have very small home ranges prior to dispersal. Dispersal starts at around 4 months of age when the young explore the surroundings of their natal range and expand their area of occupation (Wauters *et al.* 1994a). The distance of dispersal varies greatly and depends on food supply, intraspecific behaviour and the patchiness of the forest. Dispersal may only be over a short range (associated with slight changes in the area of the home range) or over long distances, involving movement between populations. Dispersal peaks during two periods, one in spring/summer and another in the autumn (Gurnell 1987). The spring/summer dispersal is the movement of young born late in the previous year and males increasing their home ranges in relation to breeding activities. In the autumn, the spring/summer born disperse and some adults may move if the food supply is poor. Once an adult is established, it tends to remain in the same general area for life unless food shortage forces a move. In areas with small patches of habitat, squirrels may move between several small ranges in different patches.

Overall, there are marked seasonal changes in population numbers and a clear annual cycle (Gurnell 1987). In late spring and early summer there is an increase in numbers as young from autumn litters are recruited to a population and immigrants become established. Wauters and Dhont (1985, reported in Gurnell 1987) found immigration exceeds emigration during this period in Belgian populations, suggesting some squirrels may overwinter elsewhere and be recruited in spring. The populations peak in summer and decrease in the autumn as the juveniles disperse. Numbers drop further during the winter due to poor food supplies and harsh weather. As shown in figure 1.5, populations can change markedly in density between years; this may lead to a greatly reduced effective population size. The patterns of density variation can generally be explained by food availability and weather conditions. Notably, there is not much difference in the overall density of red squirrels in coniferous and deciduous habitats despite the difference in habitat quality. This may be due to their stable solitary form of social organisation, based on large home ranges which will result in low densities.

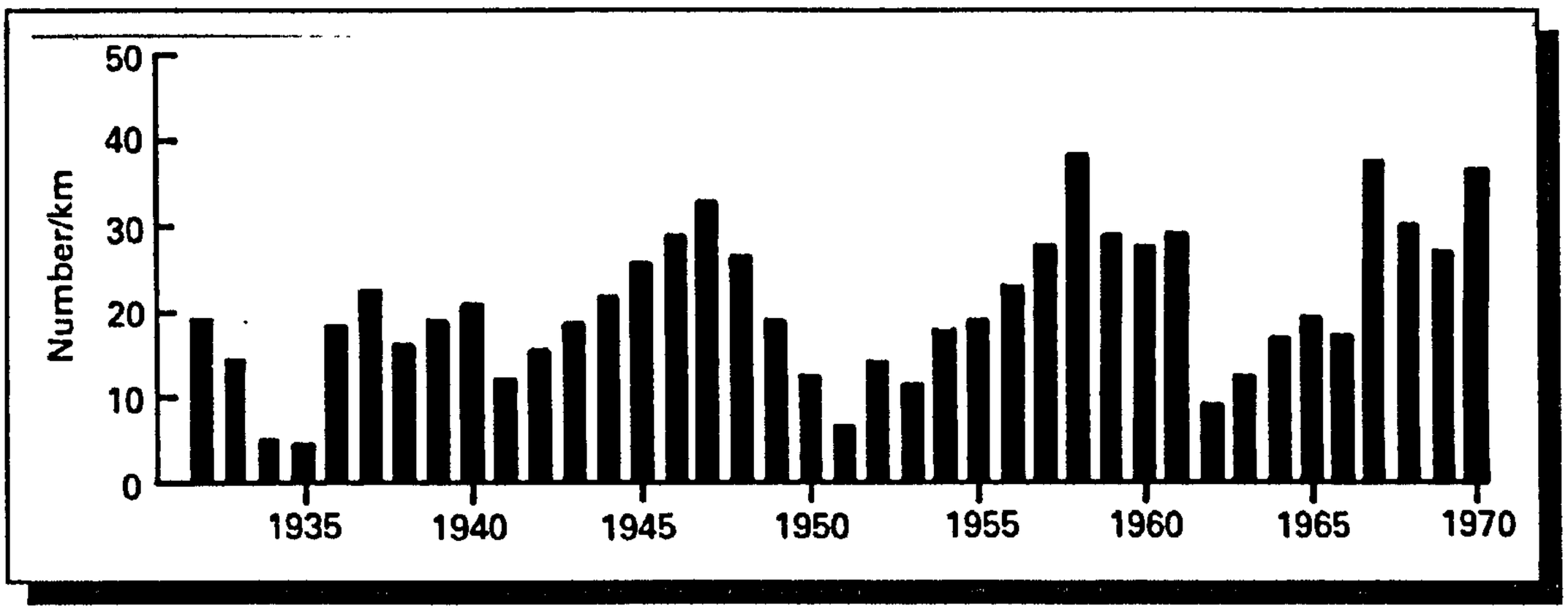


Figure 1.5: Year-to-year changes in red squirrel population sizes in eastern Siberia (reproduced from Gurnell, 1987)

1.4.2 The effects of habitat fragmentation on red squirrels

In much of western Europe, forests and woods have been destroyed and fragmented. Red squirrels occur in such habitats and, as with many small mammals, they are habitat specialists (showing a strong preference for coniferous and mixed woodland (van Apeldoorn *et al.* 1994)). This leaves them restricted to fragments of suitable habitat ("woodlots") separated by unsuitable areas. The success of an animal species in such an environment depends on dispersal and colonisation abilities, degree of specialisation in habitat use, the distances between habitat patches and other landscape elements such as roads and hedgerows that may influence movement (Verboom and van Apeldoorn 1990; van Apeldoorn *et al.* 1994).

A group of red squirrel populations in a fragmented habitat constitutes a metapopulation (van Apeldoorn *et al.* 1994). The dynamics of a metapopulation are determined by the characteristics of the landscape and the species. The rate at which a population loses members depends on the size and quality of the habitat patch, and the probability of colonisation of empty patches depends on their isolation (van Apeldoorn *et al.* 1994). The environment between the fragments is important in determining the ability of the animals to move between patches. The presence of barriers such as roads or rivers, and of corridors such as tree-rows and hedgerows will affect the ability of red squirrels to disperse between fragments. Andrén and Delin (1994) found that mean daily movements were 430m for males and 180m for female red squirrels in Sweden. They observed maximum movements of 2800m for males and 680m for females, but dispersal over 4500m has been observed between Belgian populations (G. Verbeyen, pers. com.). This suggests that when necessary much larger gaps between patches can be covered than those witnessed in the Swedish squirrels.

The populations studied by Verboom and van Apeldoorn (1990) in the Netherlands show the dynamics of a source-sink metapopulation with several large source areas of high quality habitat that remain occupied during periods of low squirrel numbers, and many small sink areas of suboptimal habitat that are only occupied during high density periods (van Apeldoorn *et al.* 1994). At times of low population density, the sink areas that are occupied tend to be those most accessible to source areas but not necessarily those of highest quality. In times of high population density some "sink" patches remain unoccupied as they are unreachable from the source patches.

The probability of fragment occupation by red squirrels is determined by certain characteristics including the size and the quality of the fragment and the distance to another fragment. The larger the woodlot the more likely it is to be occupied, but the most important characteristic is quality. For red squirrels this basically refers to the conifer content of the patch: the more conifers there are in a patch the more likely it is to be occupied (Verboom and van Apeldoorn 1990; van Apeldoorn *et al.* 1994). Also important is the degree of isolation of the fragment: the greater the amount of woods in the surrounding area and the shorter the distance to a permanently inhabited woodlot, the more likely the fragment is to be occupied (Verboom and van Apeldoorn 1990).

Red squirrels in Belgium have been observed to have the same basic social structure in fragmented as in continuous woodland (Wauters *et al.* 1994a) although home range size and use was influenced strongly by the size and quality of the patch. No effects on reproductive output, condition or survival have been seen (Matthysen *et al.* 1995), although reproduction is less synchronous in fragment populations. This may be due to increased variation in the time when females reach optimal condition for reproduction and to females re-entering oestrus more quickly after a nest loss. Fragment populations show lower abundance than populations in continuous forest and lower immigration rates (Wauters *et al.* 1994b). The fragments have apparently higher juvenile survival rates which is attributable to higher philopatry rather than true survival (Matthysen *et al.* 1995). However, increased juvenile survival is not enough to compensate for lower immigration rates and this is the most likely explanation for the reduced abundance. Dispersal rate from fragment populations is also reduced and dispersers suffer higher mortality in a fragmented landscape, thus reducing the viability of the entire metapopulation (Matthysen *et al.* 1995).

Wauters *et al.* (1994b) studied a group of fragmented populations in northern Belgium located near the populations included in this study. They found a much higher population density in the large "mainland" forest areas (these areas, Merodese Bossen and Peerdsbos, are included in this study; see section 1.4.3) than in the fragments (mean of 1.07 ha^{-1} as opposed to 0.57 ha^{-1}) due mainly to higher immigration rates in the mainland areas. There is also a tendency for squirrels of both sexes to be heavier in the mainland areas (Wauters *et al.* 1996). Multilocus DNA fingerprinting was carried out, with the application of one probe, and it was shown that the mainland populations were genetically more diverse than the fragment populations. A negative correlation between immigration rate and band-sharing (a means of quantifying genetic similarity within populations) suggested that the reduced diversity in the fragment populations was due to reduced immigration rather than small population size. These results confirm the importance of migration to fragmented populations as indicated by theory and simulation studies.

No detectable negative effects associated with inbreeding were found in this study (Wauters *et al.* 1994b). In a further study, Wauters *et al.* (1996) used fluctuating asymmetry as an indicator of stress in the small fragment populations of squirrels. Fluctuating asymmetry occurs when an individual fails to complete identical development of an otherwise bilaterally symmetric trait on both sides of the body and is thought to indicate stress experienced during development (Wauters *et al.* 1996). It has been argued that this may reflect environmental and/or genetic stress. They measured the length of the right and left hind feet and found that they did demonstrate asymmetry (having tested that the measurements were repeatable).

Squirrels in the fragment populations had a higher degree of fluctuating asymmetry than those in larger woodland areas. They also found that heavier squirrels were more symmetrical and, because heavier squirrels have a greater competitive ability and are more likely to mate and reproduce successfully (Wauters and Dhont 1992), it can be argued that asymmetry is higher in lower quality individuals. However, three of the small populations had similar levels of asymmetry to the large populations and there was no correlation between genetic diversity and fluctuating asymmetry. It is, of course, possible that the small populations are suffering genetic stress that has not been indicated by the molecular markers used, but the differences in level of asymmetry between populations were small leading the authors to conclude that this and other studies should "serve as a warning against placing too much reliance on fluctuating asymmetry as an indicator of stress in natural, small populations."

Overall, migration plays an extremely important role in the biology of red squirrels in a fragmented habitat. Red squirrels can disperse over several kilometres when necessary, but shorter distances seem to be the norm. The quality of each woodlot affects its potential for occupation by squirrels, the higher the quality the more likely it is to be occupied, but the degree of patch isolation is also important, more so in more highly fragmented areas. The isolation of a patch is not just a factor of the distance to another occupied area, but also the type of habitat between the areas. The presence of corridors aiding dispersion, or barriers to it, can be crucial for individual populations.

Red squirrels show the same social structure in both continuous and fragmented habitats, but they have lower population densities in fragmented areas due to lower immigration rates and reduced disperser success. The small Belgian fragment populations studied showed lower genetic diversity at minisatellite loci which is also likely to be due to reduced immigration, further illustrating the importance of migration in determining the effects of habitat fragmentation on red squirrels.

1.4.3 The study populations

The aim of this study is to investigate the effects of habitat fragmentation on the genetics on a group of red squirrel populations in northern Belgium, furthering the work of Wauters *et al.* (1994b). The woodland habitat around Antwerp is heavily fragmented and has probably been so for centuries (Matthysen *et al.* 1995). Woodlots are surrounded by agricultural areas, parkland and housing. A map showing the locations of each of the woodlots included in this study is given in figure 1.6. The numbers of squirrels in this region is currently increasing since the introduction of a ban on hunting in 1992 (G. Verbeyen, pers. com.).

Two large areas of continuous woodland are included in the study to allow for comparisons to be made between the genetics of the large and small populations. These are the same two areas of Merodese Bossen and Peerdsbos that were also included in the study by Wauters *et al.* (1994b). They are located some distance from the small fragment populations. Merodese Bossen (300ha) is a coniferous woodland, dominated by scots pine, and is therefore of higher quality than Peerdsbos (600ha) which is a mixed mature wood dominated by deciduous trees, but does contain a few stands of pine and spruce (Wauters and Dhont 1992). Despite the difference in quality, squirrel density in both these areas is around 1 adult/ha (Wauters *et al.* 1994b). Tissue samples, in the form of small ear plugs, were collected from some of the red squirrels occupying these areas by Luc Wauters, working at the University of Antwerp, in the early 1990s.

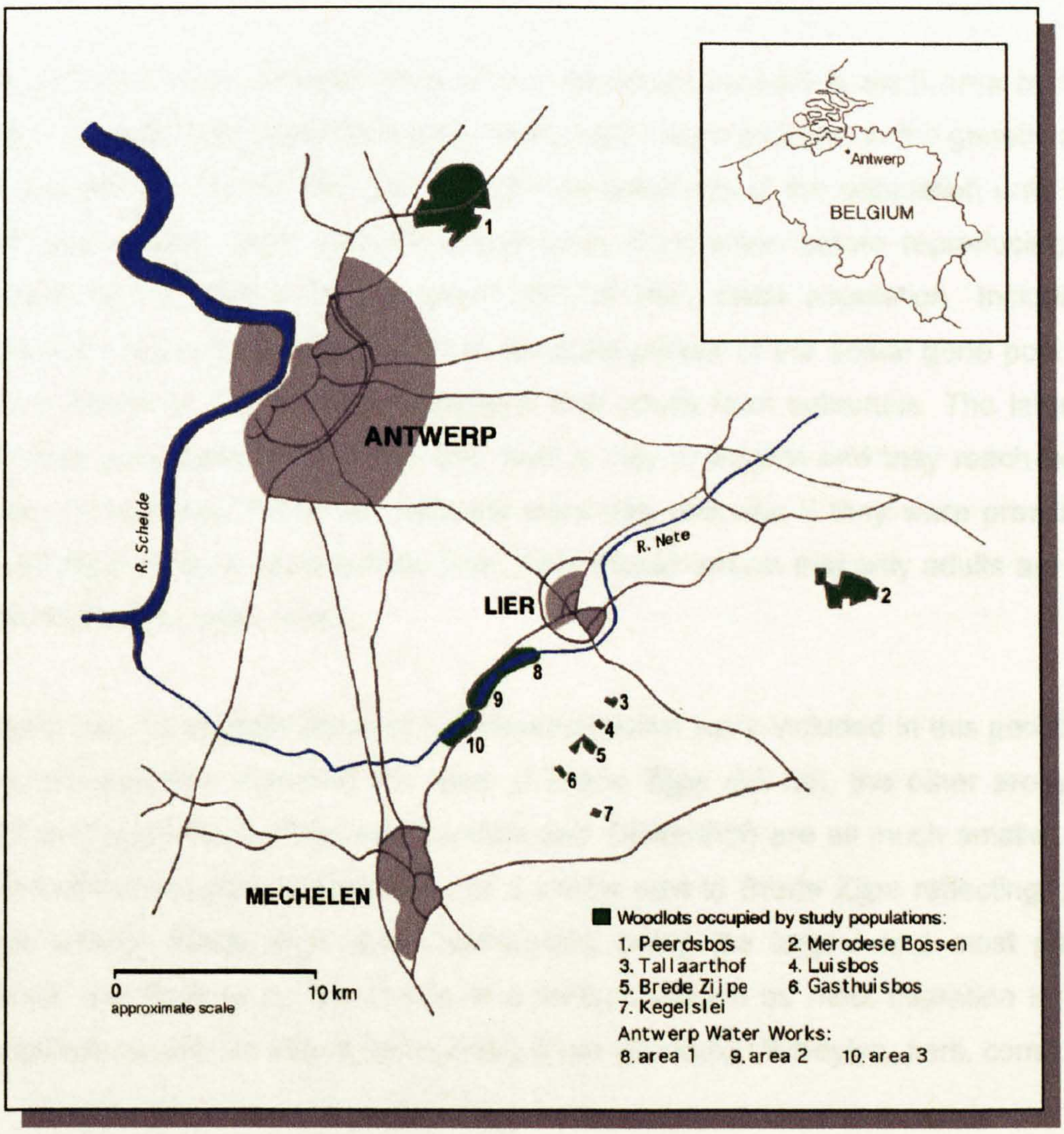


Figure 1.6: A map showing the location of the populations of red squirrels in northern Belgium included in this study.

The other populations are all small fragments, ranging in size from 3 to 17 adults. The social structure and demography of these populations has been followed for several years by Goedele Verbeyen from the University of Antwerp. All the squirrels are regularly trapped, weighed and measured. They are all fitted with a small electronic device, placed just under their skin, which transmits a unique identifying code enabling them to be correctly identified wherever they are trapped. As the populations are so small, it is possible to identify, tag and trace every individual. Each squirrel was periodically fitted with a radio collar in order to track their movements and to determine their home range within a patch. A few migration events were followed whilst squirrels were wearing radio collars and other migrations could be identified when tagged squirrels were trapped in different locations to where they were tagged. The fates of many squirrels will never be known; they disappear from an area and, unless a carcass is found or they show up in a different area, it is not possible to determine whether they have died or emigrated.

Samples of tissue were collected from all the individuals present in each area by Goedele Verbeylen. However, only reproductively mature adults were included in the genetic analysis. Juveniles were excluded as they are not effective members of the population until they are reproductively mature. Many juveniles disperse to other areas before reproducing and so never make a contribution to the gene pool of their natal population. Including only reproductively mature adults gives a more accurate picture of the actual gene pool of each population. However, it is hard to distinguish true adults from subadults. The latest that a squirrel could have been born in any one year is July or August and they reach adulthood after about 10 months. Therefore, squirrels were only included if they were present in the area as adults in June of each sample year. This should ensure that only adults are counted (Goedele Verbeylen, pers. com.).

Populations occupying eight areas of fragmented habitat were included in this genetic study. The largest population occupied the area of Brede Zijpe (55 ha), the other areas around Brede Zijpe (Gasthuisbos, Kegelslei, Luisbos and Tallaarthof) are all much smaller. Despite this, Gasthuisbos supports a population of a similar size to Brede Zijpe reflecting, perhaps, its higher quality. Brede Zijpe and Gasthuisbos, being the largest and most permanent populations, are likely to be the centre of a metapopulation as most migration is between these populations and the others surrounding them (Goedele Verbeylen, pers. com.).

There is also some migration between these five and the populations of Antwerp Water Works. The three populations around Antwerp Water Works are defined by the reservoirs that they surround. Areas 1 and 2 surround different but contiguous reservoirs, they are only

separated by a small road so they may not constitute separate populations; squirrels are likely to move relatively freely between these two areas. Area 3, however, is separated from the other areas by several roads and an urban area so squirrels are very unlikely to move freely between areas 2 and 3.

Two very small areas of habitat that are periodically occupied were not included in the genetic analysis as the populations were too small and temporary to give meaningful results. One of the populations, Klooster, was only occupied by one adult pair during the study year of 1996. The other population, Duivenbos, was first monitored in the autumn of 1996 when one adult female, who had been present in the area for an unknown length of time, was joined by a migrant from Brede Zijpe. The colonisation of these low quality areas perhaps illustrates the current increase in numbers being experienced by red squirrels in northern Belgium. These two sites are probably sink areas that are only occupied during high density periods and are maintained by immigration from the higher quality areas such as Brede Zijpe. Monitoring of Klooster was carried out during the springs of 1995 and 1996. Three individuals were present in this area in 1995 but they had all disappeared by July. One had migrated to Brede Zijpe, another had died and the fate of the third squirrel is unknown. By the following spring however, two different squirrels were present. The transient nature of this population and the extinction/recolonisation event it experienced indicates its status as a sink area.

Also included in the study is a random sample of red squirrels collected by Sibylle Münch from the Waldhäuser population in Germany. This population is located within the Bavarian Forest, the largest area of continuous forest in northern Europe. it will be useful for comparison with the Belgian squirrels as it is unlikely to have experienced any of the fragmentation effects experienced by them.

1.5 MOLECULAR MARKERS

Evolution proceeds through inheritance. Offspring inherit genes from their parents, who in turn inherited them from their parents. This process of the transmission of genetic material between the generations is inheritance and the lines of inheritance through history are phylogenies. Examining the inherited material itself, the genetic material, can shed light on the phylogenies that have defined all life on earth. The parts of the genome examined to study phylogenies are called molecular markers and the information they reveal can shed light on the answers to many biological questions about a wide range of subjects, from behavioural patterns to evolutionary relationships.

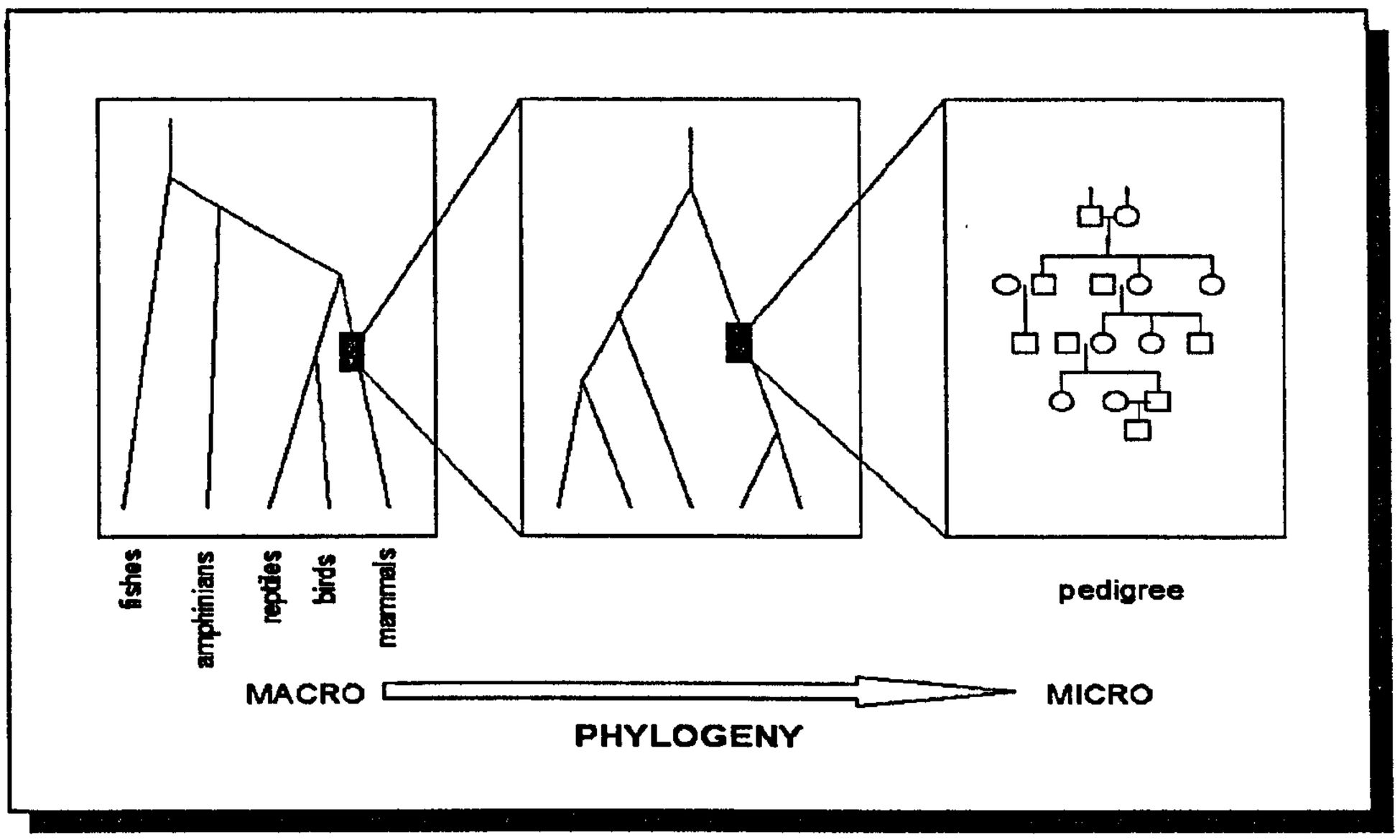


Figure 1.7: The hierarchical nature of phylogenetic assessment (based on Avise 1994).

Phylogenies exist on many levels (figure 1.7), tracing lines of descent between groups of species, species (termed systematics), populations and individuals (broadly, population genetics). Different molecular markers are appropriate for the study of phylogenies at different levels. The original molecular studies made use of protein variation but most studies now examine the DNA itself as this is believed to be more accurate, uncomplicated by the processes of transcription and translation. Some parts of an organism's genome evolve much faster than others, for example the introns (non-coding regions) of genes are generally fast evolving whereas some gene sequences themselves have hardly changed in millions of years. Faster evolving sequences are useful for comparing individuals within and between populations of the same species, as in this study, whereas slow evolving sequences are useful for studying deeper phylogenetic events. Comparisons are made between the DNA

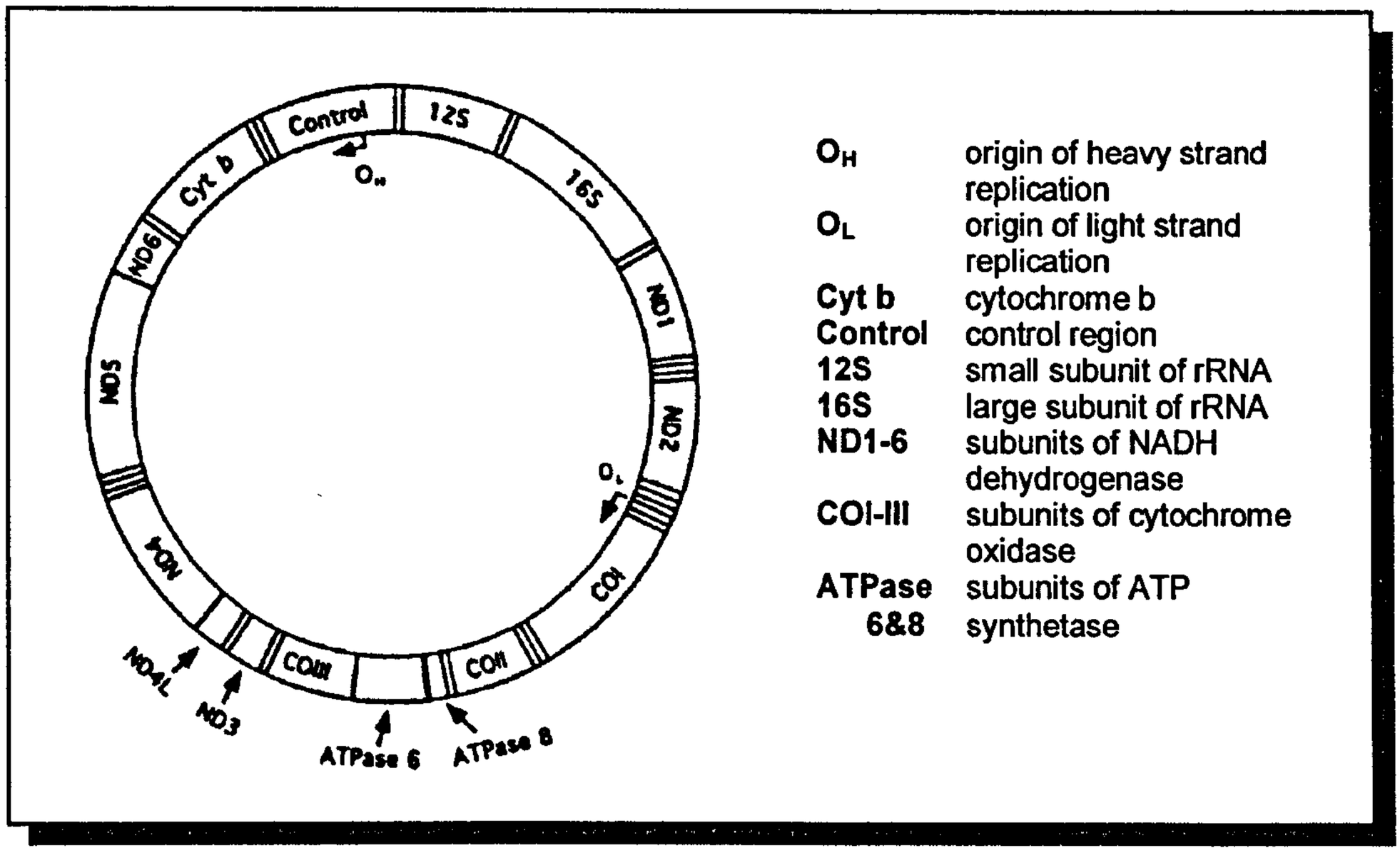
sequences of different organisms or individuals, the differences are then interpreted and conclusions drawn. An understanding of molecular evolution, of how the revealed genetic variation has developed, is essential for the correct interpretation of results of such studies. The two molecular markers employed in this study and their modes of evolution, as far as they are understood, are now discussed.

1.5.1 The mitochondrial genome

During the late 1970's and 1980's, the mitochondrial genome (figure 1.8) became the molecular marker of choice. Previously studies had relied on analysing proteins to quantify genetic differentiation and so study evolution, but it quickly became apparent that, because of its many useful features, mitochondrial DNA held much more potential as an informative marker. Avise *et al.* (1987) described the mitochondrial genome as the "bridge between population genetics and systematics" and Wilson *et al.* (1985) were just as optimistic when they stated that "animal mitochondrial DNA continues to provide perhaps the best hope of linking genetic change to organismal, which is a prerequisite for understanding the genetic basis of evolutionary change at the organismal and population levels."

There are several features of the mitochondrial genome that make it a useful tool for both population genetics and systematics. Firstly, mitochondrial genomes are found in all eukaryotic cells in very high numbers so direct comparisons can be made between a wide range of organisms. In animal cells it displays a rapid rate of evolution, although the different genes in the genome vary in their rate of change and can be used for different depths of phylogenetic analysis. Finally, its transmission genetics are simple and uncomplicated as it appears to be maternally inherited and shows little evidence of recombination. A detailed discussion of these features is included further on. However, it was the ease with which it could be isolated and purified, due to its unusual buoyant density and high copy number, that initially made mitochondrial DNA so popular (Wilson *et al.* 1985). For the first time homologous sections of DNA could be isolated and compared with relative ease.

During the late 1980s and 1990s, whilst the mitochondrial genome proved to be an extremely useful marker, limitations on its use became apparent. Some of the assumptions about the genome, such as strict maternal inheritance, have not in fact been proven; considerable doubt exists as to whether mitochondrial biology is really as simple as was originally thought. There can be differences in the characteristics of mitochondrial genomes in different organisms and this must be taken into consideration when comparisons are made (Harrison 1989).



O_H	origin of heavy strand replication
O_L	origin of light strand replication
Cyt b	cytochrome b
Control	control region
12S	small subunit of rRNA
16S	large subunit of rRNA
ND1-6	subunits of NADH dehydrogenase
COI-III	subunits of cytochrome oxidase
ATPase 6&8	subunits of ATP synthetase

Figure 1.8: A diagram of the gene content and order of the mammalian mitochondrial genome. All genes are marked except the 22 tRNAs. (Based on Quinn and Wilson 1993.)

1.5.1.1 Features of the vertebrate mitochondrial genome

It is now generally accepted that mitochondria arose when a protoeukaryotic cell engulfed or was penetrated by an aerobic bacterium. This is the endosymbiont theory and was first proposed by Wallin in 1922 (Quinn 1997). The bacterium has since evolved into the mitochondrion, an organelle existing symbiotically in large numbers in all eukaryotic cells. Mitochondria still possess their own small circular, independently replicating genome and mitochondrial DNA makes up approximately 1% of the DNA in eukaryotic cells. There are around 10 copies of the genome per mitochondrion and 1000s of mitochondria per cell. The genome varies in size between organisms but is usually about 15 - 20 kb of double-stranded DNA. The two strands are referred to as the heavy (H-) strand and the light (L-) strand because of differences in their molecular weights due to differing guanine and thymine content. The H-strand codes for most of the 37 genes (figure 1.8) found in the genome, which include 13 genes for vital proteins involved in the cell's energy systems, 22 tRNAs and 2 rRNAs (Harrison 1989). Presumably, during its evolution, redundant genes (due to the presence of similar genes in the nuclear genome) were eliminated and some genes were transferred to the nuclear genome, leaving a very compact and economical mitochondrial genome (Quinn and Wilson 1993). There are no introns, few intergenic sequences, no pseudogenes, little in the way of repetitive DNA and only one stretch of non-coding sequence, referred to as the control region (except the lamprey which has two non-coding

regions (Quinn and Wilson 1993)). The compactness of the genome even extends to the codes for some genes overlapping in different reading frames, this has lead to the suggestion that there may be some selective pressure for a small genome (Moritz *et al.* 1987; Quinn and Wilson 1993), perhaps for faster replication (Gray 1989). Gene content is generally consistent but does vary between taxonomic class.

Usually the mtDNA of an individual is homoplasmic, that is all the genomes in the individual are identical, but heteroplasmy (variation within an individual) has been known to occur (Avise *et al.* 1987; Harrison 1989). In theory, heteroplasmy could arise by a mutation occurring or by paternal leakage of mtDNA into his offspring (see below). If a novel mutation arises it may spread to fixation within that lineage. During this process, cells containing mtDNA of both states will exist within one lineage and individuals in that lineage will be heteroplasmic. This heteroplasmic state is expected to be extremely transient as mitochondrial genomes go through a very narrow bottleneck between each generation resulting in very rapid segregation of molecules. Heteroplasmy is rarely detected and most individuals are homoplasmic but it has been reported for some species including the shrew (Stewart and Baker 1994; Fumagalli *et al.* 1996), hedgehog (Krettek *et al.* 1995) and horse (Ishida *et al.* 1994). In these cases the heteroplasmic state is due to a variable number of tandem repeats located in the rapidly evolving control region.

Heteroplasmy has been detected in *Drosophila simulans* which cannot be attributed to a single mutational change or variation in repeat number (Kondo *et al.* 1990). The two types of mtDNAs differed at about 50 nucleotide positions indicating that they are from very distinct lineages. This suggested that a paternal contribution of mitochondrial DNA had occurred and both lineages were existing in the same individuals. The strictly maternal inheritance of mitochondrial DNA came to be accepted as fact in the 1980's as investigations repeatedly failed to detect any paternal contribution of mitochondrial DNA to the zygote, but the techniques employed then, involving restriction enzyme analysis, were limited in resolution.

Hutchison III *et al.* (1974) were the first to conclude that, like fungi and amphibians, mammalian mtDNA was inherited maternally when they examined hybrids of equine species and found no paternal contribution to the mitochondrial genome. They noted that the whole sperm entered the egg so they reasoned that there were two possible explanations for pure maternal inheritance: either something was preventing the paternal mtDNA from replicating or it was simply diluted down by the presence of far more maternal mtDNA, to such an extent that it could not be detected. They did, however, also acknowledge that if a paternal contribution was extremely small, they would not have been able to detect it. More recently

small amounts of paternal leakage have been detected in hybrid strains of *Drosophila* (Kondo *et al.* 1990) and mice (Gyllensten *et al.* 1991; Kaneda *et al.* 1995). This has been facilitated by the advent of PCR technology which can be used to detect very small quantities of particular sequences of DNA. These studies detected low levels of paternal DNA in the offspring of *inter*specific crosses, but as yet no paternal leakage has been detected in *intra*specific crosses. This may be of significance, as will be seen later.

In most species, the entire sperm enters the egg carrying a set of mitochondrial genomes; this fact has not always been recognised. In their review of misconceptions about mitochondria and mammalian fertilisation, Ankel-Simons and Cummins (1996) explained how the belief that the mitochondria in sperm do not enter the egg arose and spread. They likened it to a Dawkinsian "meme" (a unit of cultural transmission) where an idea spreads due to its appeal rather than its accuracy. In fact, Dawkins himself propagated this misconception about mitochondrial inheritance in his books "The Blind Watchmaker" and "A River out of Eden" (Dawkins 1986 and 1995) (Ankel-Simons and Cummins 1996).

Given that the sperm, including the mitochondria, does enter the egg, then there must be a reason why no evidence can be found of a paternal contribution to future generations. A sperm contains around 50 – 75 mitochondria, whereas an egg contains around 10^5 – 10^8, so the maternal mitochondria exceed the paternal in number by a factor of at least 10^3 (Ankel-Simons and Cummins 1996). Therefore the simplest explanation for the lack of paternal inheritance is that the paternal mtDNA is diluted down beyond the limits of detection (Kondo *et al.* 1990; Ankel-Simons and Cummins 1996). If there is no mechanism to prevent the paternal mtDNA replicating then it would be possible, occasionally, for the paternal DNA to make it through the bottleneck at oogenesis and reach significant proportions, possibly taking over the maternal line (Kondo *et al.* 1990; Avise 1991). If it was as simple as this, you would expect to find, at least occasionally, individuals within a population who are heterozygous for different mitochondrial polymorphisms present in the population. Yet this has never been shown to be the case despite many wide ranging studies involving the mitochondrial genome. Another more adequate explanation is required.

Perhaps there is some mechanism in the egg to prevent the replication of paternal mtDNA, probably by eliminating it altogether (Gyllensten *et al.* 1991; Kaneda *et al.* 1995; Shitara *et al.* 1998). Such a mechanism is known to exist in *Chlamydomonas*, a green alga, where the chloroplast and mitochondrial genomes of one parent are degraded (Kaneda *et al.* 1995); a nuclear gene thought to be involved has been cloned. Kaneda *et al.* (1995) suggested some interaction between cellular components of the egg cytoplasm and sperm mitochondria may

lead to the active degradation of paternal mtDNA. They stained mouse sperm with rhodamine and followed the mitochondria during fertilisation using phase contrast microscopy. The rhodamine stained mitochondria fluoresce as long as they are intact and they found the mitochondria functioned normally until the late pronucleus stage when the signal was lost, suggesting the mitochondria had been inactivated in some way. Both microtubules and multivesicular bodies are known to interact with sperm mitochondria and they referred to an electron microscopy study on hamster eggs which found that, during the two cell stage in the embryo, multivesicular bodies gathered around the sperm midpiece and fused with the sperm mitochondria, degrading and digesting them. However, as Ankel-Simons and Cummins (1996) pointed out, care should be taken when extrapolating from rodents to mammals in general, as differences exist in patterns of cytoplasmic inheritance between rodents and other mammals. It could be possible that elimination of paternal mtDNA in this way is a recent evolutionary development in rodents.

No paternal leakage has been detected in intraspecific crosses. In interspecific crosses there may have been sufficient evolutionary divergence for the cells to fail to distinguish paternal mitochondria, so they escape degradation. As interspecific crosses are rare in the wild, paternal leakage is likely to be an extremely rare event. If paternal leakage did occur then this would open up the possibility for recombination between molecules of different descendancies and so the complication of what have been assumed to be simple genealogies. Mitochondria contain the enzymes necessary for recombination, so there does not seem to be any reason why it should not occur if it is given the opportunity (Eyre-Walker *et al.* 1999).

A recent statistical study by Eyre-Walker *et al.* (1999) looked at the high levels of homoplasy present in the human mitochondrial genome. Homoplasies are similar character states in the DNA for reasons other than inheritance from a common ancestor, for example, due either to a repeated mutation or to recombination. They found that, with current knowledge of mitochondrial molecular evolution, all the homoplasies seen could not be explained by mutation. They therefore invoked recombination during the history of the lineages to explain the homoplasies. The problem with this idea is that for recombination to account for the high levels of homoplasy found, it would have to be occurring at a rate higher than that of the nuclear genome (J. Brookfield, pers. com.) which is clearly not the case as it has very rarely been detected. In addition, if recombination was responsible then the homoplasic sites would be clustered, yet there is also no evidence of this. It seems much more likely that the knowledge of mitochondrial molecular evolution is not yet complete enough to explain the high levels of homoplasy seen.

Far more compelling evidence for mitochondrial recombination comes from recent experimental studies. Last year Saville *et al.* (1998) published evidence of mitochondrial DNA recombination in a natural population of the fungi *Armillaria gallica*. They found phylogenies where separate lineages share a number of mutations, either they have occurred in all the lineages independently or recombination has occurred allowing horizontal transfer of the mutations. This seems by far the most likely explanation, especially considering that mitochondrial DNA is inherited from both parents in this species, so heteroplasmy is widespread, and it is detected easily in the lab. In this situation, where biparental inheritance occurs, recombination does not seem surprising, but it is the first case of mitochondrial recombination detected in a natural population.

More recently a similar case has been found in a human population on the small pacific island of Nguna (Hagelberg *et al.* 1999). They found the same rare mutation present in eight different mitochondrial types in three groups of people with different ancestries. This could have occurred by mutation, but this would require several independent mutation events on this one island when the same mutation has not been seen anywhere else in the world. Alternatively, the mutation could have arisen before the three lineages split, but then several back mutations would have had to have occurred to account for all the similarities and differences in the lineages. These two explanations are very improbable. The most likely explanation involves paternal leakage making recombination possible. Paternal leakage is most likely when closely related species or long separated populations come together. If a mechanism exists for the exclusion of paternal DNA, it may be possible that these three lineages are sufficiently diverged for it not to be very effective.

Saville *et al.* (1998) felt that recombination had not previously been detected because it occurs at such low frequencies, especially when mtDNA transmission is strongly biased towards one parent. They suggested that it may be more common in hybrid zones where diverged groups come into secondary contact. Hagelberg *et al.* (1999) suggested that other recombination events may have been missed or ignored due to the assumption that it was not possible. Some instances of hypervariability in the mitochondrial genome may not be repeated mutation events at the same site, but recombination at some point in the history of the lineages. Such hypervariability is most common in African populations which are also the most genetically diverse, so perhaps these populations are more likely to experience a failure of the paternal exclusion mechanism when they interbreed making recombination between genomes of distinct lineages a more frequent event.

The debate about whether recombination really occurs hinges on the maternal inheritance question as recombination requires paternal leakage. There is no reason to believe that recombination would not occur if genomes of different descendancy were put together in an individual. If paternal leakage occurs then evidence so far seems to indicate that it only happens when individuals of highly diverged groups have the opportunity to interbreed. These are likely to be rare events. If no paternal leakage occurs in normal intraspecific matings then recombination too is very rare. If these events are rare, then they need only be considered as possibilities when examining deep phylogenies, they are unlikely to affect a population study.

If mitochondrial DNA inheritance is strictly maternal and no recombination occurs, then the mitochondrial genome comprises a single, totally linked marker. Mitochondrial genes code for proteins which have important functions in the cell, therefore the genes can be subject to selection pressures. As the mitochondrial genome is linked, any evolutionary force acting on one site in the molecule will affect the whole genome, if it causes an increase in frequency of a particular allele then the whole molecule carrying that allele will rise in frequency in the population with it. The fixation of other polymorphisms due to linkage to a gene under selection is known as "hitchhiking" and means that none of the mitochondrial genome can be assumed to be a neutral marker (Ballard and Kreitman 1995).

Another consideration when using the mitochondrial genome as a marker is the possibility of "nuclear copies". The endosymbiont theory of mitochondrial evolution holds that genes from the original bacterial genome have been transferred to the nuclear genome. The mechanisms that enabled this may still exist and, as a result, sections of the mitochondrial genome can be found copied into the nuclear genome (Pema and Kocher 1996; Zhang and Hewitt 1996). These nuclear copies were first detected in a variety of organisms (yeast, locusts, fungi and humans) in 1983 and they have since been found in many more species including the rat and akodontine rodents (Quinn 1997). Lopez *et al.* (1994), on finding such a sequence in the nuclear genome of the domestic cat, referred to the sequence as "*numt*" for *nu*clear *mit*ochondrial DNA segment.

It is clear that mitochondrial sequences have been copied into the nuclear genome frequently and independently in many different lineages (Zhang and Hewitt 1996). They vary in mitochondrial section involved and size of sequence. They can be very large and have a high copy number. An extreme example is the *numt* found in the domestic cat where 7.9kb of the mitochondrial genome is found tandemly repeated 38-76 times in the nuclear genome (Lopez *et al.* 1994). Nuclear copies vary in homology with their corresponding mitochondrial

sequence, depending on the region, the taxa involved and on the time since transfer. Mitochondrial sequences in the nuclear genome are exposed to different mutational constraints than in the mitochondrial genome and so show different evolutionary patterns to those in the authentic mitochondrial DNA. In animals, the sequences in the nuclear genome evolve more slowly than the mitochondrial forms and so become molecular "fossils" (Perna and Kocher 1996).

Nuclear copies can impair the usefulness of mitochondrial DNA as a marker in studies utilising PCR as the nuclear copy can be amplified with, or even instead of, the authentic mitochondrial sequence required. This can be especially problematic when only one sample is used as a representative of a population or species in a phylogenetic study and it can lead to very misleading results. Universal primers, designed to bind to conserved regions of DNA for use in a wide range of species, have been very useful in population and phylogenetic studies but must be used with caution. Their conserved nature means they are more likely to amplify a nuclear copy, where one exists, than is a primer that is designed specifically for the mitochondrial sequence required. They may even preferentially amplify a nuclear copy as its slower rate of evolution may mean that a conserved primer has a better match to a nuclear copy than the authentic sequence.

When carrying out a study employing mitochondrial markers, the possibility of nuclear copies must be considered. Zhang and Hewitt (1996) outlined five experimental events which suggest the possibility of a nuclear copy being amplified along with the intended mitochondrial sequence:

1) More than one band or different bands produced by PCR amplification,
2) Sequence ambiguities or background bands,
3) Unexpected deletions/insertions, frameshifts or stop codons within the sequence,
4) Nucleotide sequences are radically different from those expected,
5) Phylogenetic analysis produces unusual or contradictory results.

For example, the *numt* found in the domestic cat was identified when double bands were consistently found at several positions along a sequencing gel, the result of the simultaneous amplification of the nuclear and mitochondrial homologues (Quinn 1997). Such ambiguities in the results of direct sequencing reactions are a common means of spotting a nuclear copy and they must not be dismissed as PCR artefacts without investigation.

Nuclear copies may present problems for population and phylogenetic studies but they also present a unique opportunity to study molecular evolution. The differences between the nuclear "fossil" and the authentic mitochondrial sequence reflect the differences in mutational environment to which the two sequences have been exposed. The relative rates of change, codon positional bias and transition/transversion ratios can all be investigated to answer questions about the molecular evolution of both genomes.

Data from the mitochondrial genome must be analysed with more caution than was originally suspected but mitochondrial DNA is continuing to be a useful marker, especially for tracing recent population changes. Its transmission genetics leaves it with a very small effective population size, one quarter that of the nuclear genome, making it especially vulnerable to population changes such as bottlenecks (Wilson *et al.* 1985; Harrison 1989). The effects of a bottleneck may be evident in the mitochondrial genome whilst the nuclear genome appears unchanged. The advent of PCR has made it possible to study useful sections of the nuclear genome with relative ease, so it is now possible to gather data from both genomes and use the useful features of both to form a clear and reliable picture of the phylogeny of interest.

1.5.1.2 Mitochondrial genome evolution in animals

The most useful feature of the mitochondrial genome in animals is its rapid rate of evolution. It was first reported by Brown *et al.* (1979) when they compared variation in mitochondrial genomes among four primate species. They showed that the mitochondrial genome of these primates evolved between five and ten times faster than the nuclear genomes. Accelerated evolution is a feature of animal mitochondrial DNA but it is not necessarily the case for other eukaryotes; plants, for example, have very slowly evolving mitochondrial genomes.

Many reasons have been suggested to explain this accelerated evolution in animals. Factors known to influence mutation rates include the metabolic rate of the organism, its generation time, differential fixing of slightly deleterious mutations, the efficiency of DNA repair mechanisms, and the nucleotide composition of the sequence (Lopez *et al.* 1997). For example, the shorter generation time of rodents is the most likely explanation for the rodent nuclear genome evolving faster than the human genome (Wu and Li 1985); the mitochondrial genome also evolves slightly faster in rodents than in man (Moritz *et al.* 1987). However, the high rate of mutation experienced by the mitochondrial genome is likely to be the result of a combination of factors, including (Gray 1989; Li and Graur 1991):

1. Greater exposure of the genome to mutagens, such as free radicals, created by the metabolic functions of the mitochondria,
2. A more error-prone replication system, for example in dNTP selection,
3. Absence or deficiency in the DNA repair systems,
4. Lack of a recombinational mechanism whereby natural selection could eliminate mildly deleterious mutations,
5. The high turnover rate of the mitochondrial DNA.

Of these, an inefficient repair system is probably the most important. The rate of mitochondrial evolution is less variable than the rate seen in the nucleus, suggesting that nuclear genome evolution is slowed down rather than the mitochondrial rate speeded up (by increased oxidative damage for example) (Moritz *et al.* 1987).

Mammalian mitochondrial DNA appears to lack both excision and recombination repair capacities, but some of the enzymes implicated in the repair mechanisms of other systems are present, so the importance of this observation is unclear (Gray 1989). The pattern of mutation accumulation, with a high incidence of length mutations and transitional changes (see below), is consistent with a lack of repair as these are the most common replication errors. Bacterial mutants lacking in repair mechanisms show the same mutational profile as the mitochondrial genome (Wilson *et al.* 1985). There are many reasons for there to be a more tolerant system in mitochondria; errors in replication of the genome will only affect it indirectly because the mitochondrial genome does not code for the proteins involved in its own replication and less accuracy may be more acceptable in a system that only codes for 13 polypeptides (Wilson *et al.* 1985).

Each gene and region of the genome has its own mutation rate and the relative rates of change are consistent across the mammals at least (Gray 1989). Overall, the control region evolves most rapidly, although the central domain within the control region is relatively conserved, and "hot-spots", where base substitution is very frequent, are evident (Moritz *et al.* 1987). In the rest of the genome, the protein genes evolve faster than the RNA genes (Gray 1989). Lopez *et al.* (1997) compared the rates of divergence for all the genes (except the tRNAs) and the control region domains in several mammals. The peripheral domains of the control region were by far the most variable regions and the 12S and 16S RNA genes were the most conserved. The COI and COIII genes were also highly conserved and ND2, ND6 and ND5 were amongst the most rapidly evolving sequences. The central domain of the control region was, on average, amongst the most conserved, but it also varied most in its position within the rankings of divergence.

The actual rate of evolution of each region is generally pretty consistent amongst mammals, with the occasional exception. For example, along the lineage leading to higher primates, the COII gene shows a rate of evolution five times greater than the same gene on other lineages. The cytochrome *c* gene in the nuclear genome shows the same accelerated evolution and codes for a protein that interacts directly with this subunit of cytochrome oxidase in the electron-transport chain (Wilson *et al.* 1985). The reason for the increase in the rate of evolution is unknown but it is an important consideration for phylogeneticists using mitochondrial genes to compare distantly related lineages.

The rate of sequence divergence between two lineages is fastest in the initial stages of divergence. When it reaches about 15% diverged, the rate decreases until, by about 30% diverged, the actual rate of divergence is one order of magnitude less that it was initially (Moritz *et al.* 1987). Of course, mitochondrial evolution is still occurring but once a site has mutated, further changes at this site are not detectable by comparison with the other lineage. This is referred to as "saturation". As more sites become saturated, the rate of detectable sequence divergence also slows down.

The mitochondrial genome shows a rapid accumulation of both point and length mutations (Wilson *et al.* 1985) and the patterns seen in mutation accumulation are thoroughly reviewed by Moritz *et al.* (1987); the mitochondrial genome is one of the best understood macromolecules. The pattern of base substitution is largely determined by the degeneracy of the genetic code. Each amino acid is coded for by more than one codon, so not all base changes cause an amino acid change in the resultant protein. Silent changes, which do not affect the protein, are referred to as "synonymous" and it is these nucleotides that evolve fastest due to the lack of functional constraint (Gray 1989). For this reason, changes at the third codon position are most common (Moritz *et al.* 1987). A comparison of primate, cow and mouse sequences showed that 51% and 61% of the first and second codon positions respectively were conserved compared with only 16% of third positions (Aquadro *et al.* 1984). Selective constraints, due to the functional requirements of the proteins, are likely to affect most of the genome, reducing the amount of tolerable mutation. For example, constraints relating to the secondary structures of the tRNA molecules are reflected as consistency in the location of conserved regions in the sequences for the molecules (Aquadro *et al.* 1984). Base substitutions that are transitions (changes from purine to purine (A↔G) or pyrimidine to pyrimidine (C↔T)) are much more common than transversions (from purine to pyrimidine or vice versa), with ratios of 10:1 being reported (Gray 1989), but this pattern disappears with increasing divergence time.

Animal mitochondrial genomes show little variation in size compared with the organelles of other groups, and mammals, particularly, have genomes that are relatively uniform in size (Moritz *et al.* 1987). Minor length variation, however, is common and is most usually found in the control region where deletions or duplications of sequence frequently occur (Moritz *et al.* 1987). Some species contain lengths of tandemly repeated sequence in the control region which varies in repeat length due to replication slippage (for example, in shrews (Fumagalli *et al.* 1996), rabbits (Dufresne *et al.* 1996; Casane *et al.* 1997) and the loggerhead shrike (Mundy *et al.* 1996)).

The original popularity of the mitochondrial genome as a molecular marker stemmed from the ease with which it could be isolated and purified. With the advent of PCR, that is not such an important consideration as virtually any section of a genome can now be isolated by amplification. However, the mitochondrial genome remains a useful marker, largely due to the knowledge we have of the genome as a result of the many studies that have used or investigated it. The different sections of the genome continue to be a popular source of information for all levels of phylogenetic inference. The control region, in particular, has proved very useful in population genetics because of its high rate of evolution. Maternal inheritance makes it especially useful for detecting a reduction in population size as it will feel the effects far more than the nuclear genome. It is also useful in that it provides an alternative source of genetic information that can be used in combination with rapidly evolving nuclear markers, such as microsatellites. They will be affected by population genetics processes in different ways and so the comparison and interpretation of both data sets should lead to greater accuracy and resolution in the conclusions drawn.

1.5.2 Microsatellites

The pioneers of population genetics, the theorists Wright, Fisher and Haldane, developed models of the behaviour of gene frequencies in populations. The development of methods to analyse allozyme markers, where proteins of differing composition are separated by applying an electric current, revolutionised population genetics by providing a method for quantifying gene frequencies in natural populations. Population genetic theories could be tested and applied. The allozyme studies provided data about discrete loci with co-dominant alleles that were inherited in a Mendelian fashion. Microsatellite loci also fit these criteria and have the added advantages of having more alleles per locus and an increased likelihood of neutrality.

The existence of repetitive elements or tandem repeats in the genomes of eukaryotes has been known about since the 1970's (Bruford and Wayne 1993) and they were named "satellites" before they were known to consist of repetitive DNA. In a density gradient genomic DNA forms a band depending on the guanine and cytosine (GC) content of the DNA, prokaryotic genomic DNA forms a single peak whilst eukaryotic genomic DNA either forms a very broad peak or multiple peaks (Li and Graur 1991; Tautz 1993). The extra peaks formed by eukaryotic genomes were called "satellites" and it was later shown that they consisted of millions of tandem repeats of a short sequence motif. Their sequence simplicity means that their GC content often deviates from the average and they form satellite peaks in a density gradient.

These large repeats are too big to be utilised as molecular markers but shorter regions of repeats or variable number tandem repeat (VNTR) loci have proved to be very useful. Minisatellites are only repeated up to several hundred times at each locus but they consist of relatively large repeat units, 10-60 bp, so allele sizes may reach 50kb in length (Bruford and Wayne 1993). These loci are extremely polymorphic, varying in the number of tandem repeats, with heterozygosity values often greater than 90% and mutation rates around 10^{-2} per generation. For this reason the technique of DNA fingerprinting (Jeffreys *et al.* 1985), which employs Southern blotting techniques to reveal minisatellite variation as a banding pattern, is extremely productive for individual recognition and in parentage studies.

Microsatellites (originally known as simple sequences and sometimes referred to as short tandem repeats (STRs)) are shorter, up to only 150bp in length. They consist of repeat units of 1-6bp which may be repeated up to 100 times, but the most useful loci contain 15–30 repeats. Like minisatellites, they are highly variable in the number of tandem repeats found at the loci. The large number and wide distribution of microsatellites was described by Hamada *et al.* in 1982 but it was not until 1986 that Tautz *et al.* showed that many different simple sequence motifs existed in eukaryotes and that they were five to ten times more common than equivalent random motifs. They are found in all higher organisms but their abundance in plants is about a fifth of that in animals, although the variation is just as high (Zhivotovsky *et al.* 1997). They are less variable than minisatellites, with mutation rates of between 10^{-4} and 10^{-5} being typical (Ashley and Dow 1994) giving them a broader spectrum of potential applications.

It is thought that there is some clustering of microsatellite loci in the genome on a small scale but over the whole genome they are distributed randomly and evenly (Jarne and Lagoda 1996) as indicated by hybridisation studies. It has been estimated that there is about one every 6kb in the human genome (Ashley and Dow 1994). Microsatellite loci can be found within expressed regions of the genome where they may feel effects of selection by hitch-hiking and by size limitation. However the frequency of loci in such regions relative to those in neutral regions is assumed to be low enough to be of little concern to population geneticists (Jarne and Lagoda 1996).

Microsatellite sequences can be grouped into three different types: pure, where just one repeat unit is replicated precisely throughout the locus; compound, where two or more repeat units are replicated next to each other; and interrupted, where the replication of the repeat unit is interrupted by imperfections. Examples of these different types are given in figure 1.9, these examples show dinucleotide repeat units but any combination of these types is possible with any size of repeat units. Di-, tri- and tetranucleotide repeats are the most commonly used. The abundance of different types of motifs is highly skewed, for example $[CA/GT]_n$ repeats are by far the most common in the mammalian genomes whereas $[GC/CG]_n$ is very rare; the most common in the genomes of higher plants is $[AT/TA]_n$. These patterns should be considered when isolating loci for the development of new primers.

Pure	CACACACACACACACACACACA	$(CA)_{11}$
Compound	CACACACACACACA**GAGAGAGA**	$(CA)_7(GA)_4$
Interrupted	CACACA**TT**CACACACA**TT**CACA	$(CA)_3TT(CA)_4TT(CA)_2$

Figure 1.9: Examples of the three types of microsatellite sequences found in the eukaryotic genome (based on Jarne and Lagoda (1996)).

No use was made of microsatellites until the advent of the polymerase chain reaction (PCR) (Saiki *et al.* 1988) which facilitated the repeated replication of short stretches of DNA in an automated reaction. This allows the relatively short microsatellite loci to be amplified, and so visualised by electrophoresis, with great ease; this process is much faster and easier to apply than the protocols for DNA fingerprinting. In 1989 several successful attempts were made to amplify microsatellites using PCR and they were showed to be variable (Litt and Luty 1989; Tautz 1989; Weber and May 1989). The short length of the loci and the use of PCR means that microsatellites can be amplified from poor quality DNA extracted from sources such as hair roots, feathers, faeces and bones (Schlötterer and Pemberton 1994). Unfortunately, these advantages are partially offset by the need, in most cases, to develop new primers for each new species studied and this can be a long laborious process.

Microsatellites have shot to popularity in the 1990's, especially for mammalian geneticists, for three main reasons:

1. They are central to the efforts to construct whole genome maps as they occur evenly throughout eukaryotic genomes (Valdes *et al.* 1993), although this use has largely been superseded, in humans at least, by efforts to sequence entire genomes.
2. Instability in repeats is implicated in at least three human diseases, Fragile X, Myotonic dystrophy and Huntington's disease (Rubinsztein *et al.* 1995). These diseases all involve an abnormally expanded trinucleotide repeat so the mutation processes of microsatellites is a matter of interest for medical geneticists.
3. Their high levels of polymorphism and the ease of scoring the alleles has made them extremely popular with population geneticists.

1.5.2.1 Microsatellite evolution

A change in the number of repeats in a tandem array could occur by two mechanisms, slip-strand mispairing or recombination. For population and medical genetics, an understanding of the mutational processes of microsatellites is crucial to the correct interpretation of results. Recombination or crossing-over involves an exchange of DNA between homologous chromosomes. It can easily occur between misaligned tandem arrays resulting in the uneven exchange of material so that one array gains some repeat units whilst the other loses some (Li and Graur 1991). This process could lead to relatively large gains or losses of repeat units. Mutation in microsatellite loci rarely seems to involve recombination as most mutations involve the gain or loss of a single repeat unit suggesting slip-strand mispairing (Weber and Wong 1993; Primmer *et al.* 1996a; Primmer *et al.* 1998) and mutations in *Escherichia coli* and yeast that reduce or eliminate most types of recombination do not affect microsatellite stability (Wierdl *et al.* 1997).

It is generally accepted that most microsatellite mutations occur by slip-strand mispairing (Levinson and Gutman 1987; Schlötterer 1998a). This process was described by Levinson and Gutman (1987) as the local denaturation and displacement of the strands of a DNA duplex followed by mispairing of complementary bases at the site of an existing short tandem repeat. In essence the two strands of a molecule misalign in the repeated region and replication leads to insertions or deletions to correct the mispairing. This process is illustrated in figure 1.10. It can be seen that where the mismatch loop forms in the strand being replicated the continued synthesis results in the addition of a repeat unit, whereas where the mismatch loop is in the template strand, the result is the deletion of a repeat in the replicated strand (Wierdl *et al.* 1997).

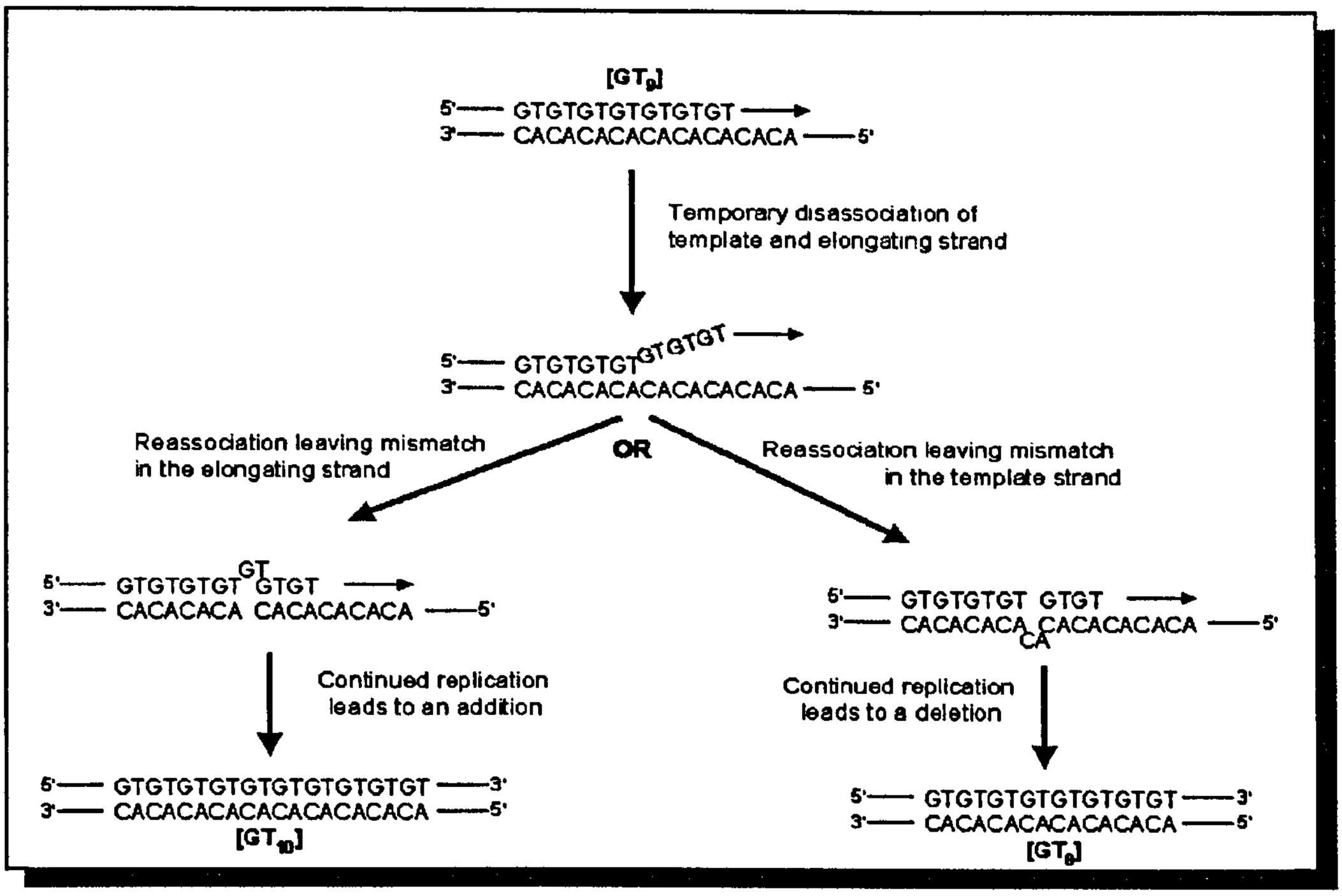


Figure 1.10: The process of slip-strand mispairing where the misalignment of the two DNA strands within a tract of short tandem repeat units leads to the addition or deletion of one repeat unit. (Based on Wierdl *et al.* (1997)).

In vitro studies and model building of the double helical DNA structure have shown that the bulges that result from DNA slippage can be accommodated into the DNA structure (Levinson and Gutman 1987) and slippage reactions similar to those described can be carried out *in vitro* in a laboratory to extend sections of repeated DNA (Schlötterer and Tautz 1992). In theory, slip-strand mispairing could occur any time that unpaired loops are formed, during repair as well as replication, so it could potentially be a frequent event. Indeed slippage mutation rates are estimated ranging from 10^{-5} to 10^{-2} for microsatellite loci. These values are orders of magnitude greater than the rates of nucleotide substitutions in the nuclear genome (Rose and Falush 1998). It would be expected that shorter slips that distort the DNA helical structure less should be more frequent (Levinson and Gutman 1987) and this is the pattern observed as the majority of mutations are gains or losses of only one repeat unit.

It appears that only a short array of repeats is needed for slip-strand mispairing to occur, leading to the rapid expansion of the repeated array. Rose and Falush (1998) examined a large number of repeat arrays in the genome of *Saccharomyces cerevisiae* and proposed a threshold size of eight nucleotides beyond which microsatellite expansion occurs. They found that this minimum size of repeat array for expansion to begin was not dependent on the number of repeats, but on the length of the repeat region in nucleotides. This is supported by

Messler *et al.* (1996) who described the "birth" of two separate microsatellites in the same place in the genome of different primates. One involved a point mutation which created the motif $[ATGT]_2$ which was then expanded and the other involved a substitution changing the sequence $GTAT[GT]_2$ into $[GT]_4$ which was also subsequently expanded. In both cases the original motif was eight nucleotides in length. Weber (1990) found that the information content of dinucleotide repeats in the human genome, which is a factor of the variation seen in repeat length, dramatically increased when the number of repeats exceeds about 10, further confirming this trend.

The original short motifs that are the raw material for microsatellite loci would frequently be created by chance in the genomes by nucleotide substitution, as occurred in the primates described. Groups of related species can be described as sharing a "library" of conserved short satellite sequences, some of which expand into long loci in particular lineages (Mestrovic *et al.* 1998). Families of species can be seen to share many loci, but often it is only in a few of the species that particular loci have been expanded to large variable tracts of sequence. Mestrovic *et al.* (1998) looked at four insect species from the genus *Palorus* and found the same set of loci present in all the species but in each species only one of the loci, a different one in each case, had expanded to a large size.

Primmer and Ellegren (1998) witnessed both the "birth" and "death" of microsatellite loci when looking at loci in 39 different bird species. They postulated an ancestral repeat of $[GA]_4$ for the locus *FhU2* which had been expanded by slippage in some species of the phylogeny but in others it had been reduced or interrupted and in two species totally lost. Microsatellite loci can be lost by deletions removing repeat units or by point mutations interrupting them and stabilising them. Slippage is less likely if there are imperfections in the repeat (Weber 1990). In the two species that had in effect lost the *FhU2* locus, a point mutation had occurred in the middle of the short array reducing the length of the longest perfect repeat to two, too short for slip strand mispairing.

1.5.2.2 Microsatellite mutation models

There are two principle models of mutation, the "infinite alleles model" (IAM) and the "step-wise mutation model" (SMM). The IAM was originally described by Fisher and by Wright in the 1930s (Nei 1987) but it was not until the age of allozyme electrophoresis that it was considered in more detail by Kimura and Crow (1964). The IAM recognised that the mutation rate at each nucleotide site is so low that almost all mutations occur at previously monomorphic sites. The model holds that each mutation gives rise to a new allele not

previously found in the population and that there are an infinite number of possible allelic states (Nei 1987; Shriver *et al.* 1993).

The SMM was proposed by Ohta and Kimura (1973) in response to the recognition in previous studies that there were regularities in the distribution of alleles that could be distinguished by protein electrophoresis (Valdes *et al.* 1993). These distributions seemed to be consistent with a mutation model where the net charge of the allele changed by one unit when it mutated. This is the step-wise mutation model and is illustrated in figure 1.11. When applied to allozymes it assumes that only the net charge of the molecule is responsible for its electrophoretic mobility and most mutations affect the charge in single units (Nei 1987), but interest in the model as applied to allozymes quickly waned when it was discovered that adjacent electrophoretic alleles did not, in reality, usually differ by a single charge state (Valdes *et al.* 1993).

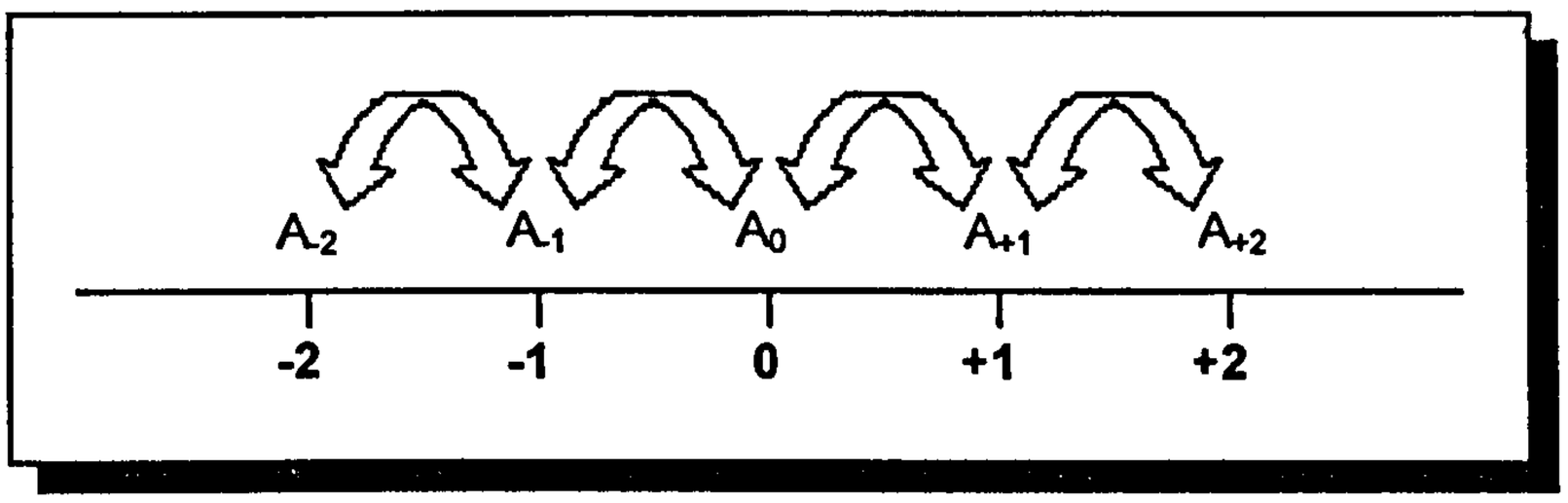


Figure 1.11: The step-wise mutation model. The pattern of mutations under this model can be visualised as a series of integer points on a line and a mutation results in the conversion of one allele into one on the left or right of it (SHRIVER *et al.* 1993). Based on Nei (1987).

More recently Valdes *et al.* (1993) and Shriver *et al.* (1993) rediscovered the model when studies of allele distributions and mutation patterns indicated that it may be applicable to microsatellites. They tested the model with computer simulations to generate data for alleles experiencing a step-wise mutational process and compared the data to that expected by the formulae described by Ohta and Kimura. They found the data generated fitted the SMM model better than the IAM. Since then many studies have looked at natural populations and directly tested the SMM model (Weber and Wong 1993; Amos *et al.* 1996; Primmer *et al.* 1996a; Primmer *et. al.*, 1998) and found that the general pattern of microsatellite mutations is indeed single step changes in repeat number as described by the model.

Unfortunately, some of these studies and others have also indicated that, whilst the SMM may broadly reflect observed patterns of microsatellite mutations, it is an oversimplification and inadequate to fully describe microsatellite evolution.

1.5.2.3 Problems with the step-wise mutation model

Many of the assumptions made by the step-wise mutation model have been challenged in the last six years by studies examining microsatellite mutations and allele frequency distributions more closely. The assumptions are that:

1. All mutations at microsatellite loci involve the loss or gain of a single repeat unit;
2. Mutations are unbiased in direction, gains in repeats are as common as losses;
3. There is no constraint on allele size at either end of the scale;
4. The mutation rate at a particular locus is constant.

Firstly, some of the more informative studies will be reviewed and then their impact on the above assumptions will be discussed.

One of the first studies of microsatellite mutation was Weber and Wong (1993), who looked at mutations in human microsatellites by parent-offspring comparisons within 40 reference families from the CEPH (Centre d'Étude du Polymorphisme Humain). These are a group of human families, established as a collection in 1984, to provide a consistent set of reference pedigrees (NIH/CEPH Collaborative Mapping Group 1992). Weber and Wong found that, although most of the mutations (91%) involved the gain or loss of a single repeat unit (as predicted by the SMM) 9% of the mutations involved larger changes than one unit and the majority of mutations were gains rather than losses. They also found that most of the mutations occurred in the male germline. This sex bias has also been seen in the barn swallow (*Hirundo rustica*) (Primmer *et al.* 1998) and is presumably due to there being over ten times more cell divisions in the process of sperm formation than in egg formation (Weber and Wong 1993).

The directional bias in microsatellite evolution, the tendency for the repeat array length to increase, has since been recognised in humans (Amos *et al.* 1996), barn swallows (Primmer *et al.* 1996a; Primmer *et al.* 1998) and strains of yeast (Wierdl *et al.* 1997). If mutations have a strong tendency towards an increase in array size then, unless there is a constraint on size, each locus would be expected to expand indefinitely. This is clearly implausible, so there must be some process preventing them from expanding out of control. For this reason the discussions about directional bias in mutation and constraints have become inextricably linked.

Rubinsztein *et al.* (1995) and Garza *et al.* (1995) fired up the debate when they amplified loci isolated from the human genome in primates. Both studies found that the alleles were consistently longer and more variable in humans than in primates which, along with the

positive skews visible in the allelic distributions, they took as evidence for directional mutation. They hypothesised that the mutation rate in humans is higher so directional mutation favouring increase in allele size leads to humans having longer alleles. Rubinsztein *et al.* (1995) proposed three possible reasons that humans may have a higher mutation rate: (1) a less efficient mismatch repair mechanism; (2) human males are much older than primate males when they reproduce so mutations may be more likely to occur during sperm production; and (3) human populations are larger and so are likely to experience more mutation events. It also seems possible that if an allele size ceiling exists then it could simply be higher in humans than in primates. However Rubinsztein *et al.*, whilst not rejecting that possibility directly, argued against the existence of a size limit on alleles. At the time there was no evidence for any constraint on allele size and their allelic frequency distributions had a positive skew when a negative one would be expected if there was an upper limit on size.

A lack of size constraint contributed to their arguments against the pattern being due to an ascertainment bias. Ellegren *et al.* (1995) described this, explaining that during the process of isolating microsatellite markers, longer, and usually more polymorphic, loci are selected. Since the divergence of two species, homologous loci will have evolved independently, so when loci from the upper end of the size distribution in one species (the "focal" species, humans in Rubinsztein *et al.*'s study) are chosen they are not likely to also be the longest loci in the other species. Therefore mean allele length will usually be longer in the focal species than in other species from which it is amplified.

Rubinsztein *et al.* (1995) did not think that this bias was sufficient to explain the highly significant patterns they observed. They thought that the lack of a size ceiling meant that there could not be a real "size range" from which loci could be selected (the possible presence of size constraints will be discussed later) and they felt that the extent to which the allelic distributions overlapped in both species indicated that any ascertainment bias was minimal. The loci compared were free to mutate and were polymorphic in both species so it may be expected that at least some of the loci would be longer in other primates than in humans if there was no directional mutation. Ironically, the small differences found between allele size distributions in different species has also been used as evidence for the existence of allele size constraints (Garza *et al.* 1995).

In an attempt to determine whether the observed patterns could be solely due to ascertainment bias or whether they are indeed due to mutation bias, Ellegren *et al.* (1997) made a reciprocal comparison of loci isolated from cattle and sheep. They applied 13 bovine and 14 ovine markers to both species and found that 11 of the bovine loci were longer in

cattle than in sheep and 12 of the ovine markers were longer in sheep than in cattle. In addition to this, several of the bovine loci were monomorphic in sheep and vice versa, and the heterozygosities were higher in the focal species.

This obviously supported the ascertainment bias explanation but, as the authors pointed out, did not rule out directional mutation itself. However, Crawford *et al.* (1998) made a similar comparison of loci in sheep and cattle found that loci were longer in sheep than cattle regardless of their derivation, not supporting the ascertainment bias hypothesis. As Hutter *et al.* (1998) pointed out, these studies were confounded by using microsatellite loci isolated by different methods, so potentially exposed to different biases, and applied to domestic populations that were likely to have had similar histories. Hutter *et al.* (1998) carried out a similar reciprocal comparison on natural populations of two species of *Drosophila* but found no difference in PCR product length between the species, although the heterozygosity was generally higher in the focal species. The lack of length difference may be due to the low mutation rates and the generally shorter loci lengths in *Drosophila* microsatellites (Schlötterer *et al.* 1998). Without a difference in length it is hard to draw conclusions about what may be causing differences, so the debate remains unsettled.

Other evidence for directional mutation exists however, so resolution of the debate about the potential ascertainment bias is not vital for understanding microsatellite evolution. A study of spontaneous germline mutations in the barn swallow, detected in a large scale paternity assessment carried out by Primmer *et al.* (1996a), revealed many interesting features of microsatellite evolution, at least for the evolution of the particular highly polymorphic locus concerned. 34 mutations were observed and, as expected, most involved single step changes, but six events involved larger changes of up to five repeat units. The step-wise mutation model was therefore not strictly applicable to this locus. They also found that the number of changes involving gains significantly exceeded losses, adding to the evidence for directional mutation. The alleles showing mutational events had a significant tendency to be the longer alleles, so they hypothesised that mutation rate increases with repeat array size. They extended this study and in 1998 again reported a positive relationship between allele size and mutation rate (Primmer *et al.* 1998) indicating that slippage is more likely for longer repeat arrays. Schlötterer *et al.* (1998) also found this relationship to be true for a locus in *Drosophila melanogaster* where one hypermutable allele was amongst the longest reported for the species. Other studies have also found this pattern for loci in *E. coli*, yeast and human genomes (Weber 1990; Wierdl *et al.* 1997).

Perhaps the most potentially useful study to date is that of Wierdl *et al.* (1997) which looked at the instability of $[GT]_n$ repeats in yeast (*Saccharomyces cerevisiae*). They saw many of the patterns already described but they also drew some interesting conclusions. They found that the rate of increase in mutability with repeat array length was more than linear. Weber (1990) had plotted informativeness (variability) against repeat length and showed that an S-shaped curve best fit the data. If this increase in mutation rate was simply due to an increased probability of misalignment with increasing array length then the increase would be expected to be linear, directly relating to repeat array length. They suggested that the more than linear relationship was due to increased probability of misalignment combined with a decrease in the efficiency of the mismatch repair mechanisms (which correct misalignments before they are replicated) with increasing array length.

When looking at the pattern of observed mutation, Wierdl *et al.* (1997) confirmed that most mutations were small changes, again additions exceeded deletions, but they also witnessed some large deletions. They found that deletions of more than eight repeats were more common than those involving between two and seven repeat units. No other study has observed such large changes in repeat length but as Primmer *et al.* (1998) pointed out, large deletions may represent a very small proportion of all mutations and as mutational events are quite rare anyway, other studies may not have looked at enough mutations. Some large changes may have been missed in previous studies as they all assume that the parental allele of a new variant in the allele that is closest to it in size. This might not always be true (Primmer *et al.* 1998).

It was proposed that the sudden occurrence of large deletions could be due to a change in DNA conformation when the repeat array becomes large. These could be the result of slippage events involving the formation of large loops, or be due to recombination. Wierdl *et al.* (1997) argued that the deletions were more likely to be due to large slippage events because recombination would result in more large deletions occurring during meiosis than mitosis, as that is when most recombination occurs, but the rates of large deletions during both types of cell division were found to be similar. They also looked at the occurrence of large deletions in a mutant strain of yeast lacking the RAD52 gene product which is necessary for recombination and found the same rate of large deletions. Recombination therefore seems unlikely to be the cause of large deletions, although it cannot be ruled out.

The excess of additions led Wierdl *et al.* (1997) to suppose that there may be another mutation mechanism in addition to slip-strand mispairing which results in the gain of a repeat unit. Alternatively, the mismatch repair system may be biased so that it corrects mismatches

on the elongating strand, so preventing additions, less efficiently as the array gets longer or corrects mismatches on the template strand with increasing efficiency as the repeat gets longer, leaving more additions relative to deletions. It was not suggested by the authors, but the imbalance may simply be due to mismatches occurring more often on the elongating strand; this bias may not become noticeable until the tract reaches a reasonable length.

Clearly there is evidence for many violations of the step-wise mutation model, perhaps the most fundamental of which is the occurrence of mutations larger than the loss or gain of a single repeat. Primmer *et al.* (1996a), Weber and Wong (1993) and Wierdl *et al.* (1997) all witnessed small repeat array length changes of more than one repeat unit but only Wierdl et al. (1997) saw large changes, all deletions, which most commonly affected the longer arrays. Most mutations however do involve single step changes and the rate of larger mutations may be low enough not to affect the model significantly. This would need to be tested when more data about the frequency of larger mutations are available.

Regardless of whether some of the conclusions from studies implicating directional evolution are valid given the possibility of an ascertainment bias (Garza *et al.* 1995; Rubinsztein *et al.* 1995), it seems quite clear from direct observation that the mutation process is biased in favour of additions (Amos *et al.* 1996; Primmer *et al.* 1996a; Wierdl *et al.* 1997; Primmer *et al.* 1998). This bias is strong enough to lead some authors to suggest that there may be another mutation mechanism involved that only causes additions (Wierdl *et al.* 1997). This pattern is also important in that it implies a need for a size constraint mechanism to prevent the loci growing out of control.

The possibility of such a size constraint was suggested by other authors who noted that the difference in allele distributions between species and populations was not as much as may be expected from rapidly evolving loci in two diverged groups of organisms (Garza *et al.* 1995; Falush and Iwasa 1999). They implicated some sort of size limiting mechanism. Both Garza *et al.* and Falush and Iwasa suggested possible constraint mechanisms; Garza *et al.* favoured a biased mutation explanation where small alleles tend to increase in size and large alleles decrease. This seems increasingly unlikely as more evidence accumulates indicating that mutability increases as allele size increases, whilst single step additions remain the most common mutation. Falush and Iwasa implicated selection; as they pointed out, biologically there has to be a limit on the array size that can be tolerated so large arrays exceeding such tolerance limits would be selected against. The strength of selection on a locus would depend on the strength of the upward bias.

Another fundamental assumption of the SMM is that the mutation rate at each locus is constant. Many studies have in fact found that the rate of mutation increases with repeat array size (Weber 1990; Primmer *et al.* 1996a; Wierdl *et al.* 1997; Schlötterer *et al.* 1998; Primmer *et al.* 1998). This is intuitively logical as it makes sense that longer repeat arrays are more likely to misalign than shorter ones.

The process of mutation at microsatellite loci has proved to be much more complicated than that assumed by the SMM, several authors have described the SMM as simplistic and unrealistic (Garza *et al.* 1995; Schlötterer 1998a; Nielsen and Palsbøll 1999). The processes of microsatellite evolution appear to be much more complex than originally thought and may vary in different loci. Bearing this in mind Nielsen and Palsbøll warned "against strong interpretations of the results of ecological or demographic studies that rely heavily on specific models of microsatellite evolution."

1.5.2.4 Alternative models

A strict one step mutation model such as the SMM is clearly not adequate to describe fully the mutational processes experienced by microsatellite loci. In 1994, Di Rienzo *et al.* proposed an alternative model that they felt better fit the observed frequency distributions seen in human loci. The SMM made no allowance for the occasional mutational change larger than just single repeat unit additions or deletions. Their model, the "two-phase mutation model" (TPM), assumes that most mutations are single step changes, but also includes rare but important changes of a larger magnitude. They found that for all the loci they examined, a relatively low probability of multistep mutations was sufficient to explain the distributions seen.

This model still does not take into account the observed directional biases, or the presumed resulting limitation on allele size. The pattern of directional mutation and the more common occurrence of large deletions in longer arrays, as described by Wierdl *et al.* (1997), suggests the possibility that most mutations are step-wise additions until a locus reaches a large size when, perhaps due to resulting changes in conformation of the DNA, some large deletions also occur to moderate the allele sizes. Such a process of cyclical evolution was modelled by Falush and Iwasa (1999). They set up a model whereby mutability rapidly increases with array size and, despite the upward mutational bias, their model showed many features of constrained evolution. When the array reached a large size it always quickly returned to a small size again. This model seems to take most of the observed trends in mutational behaviour of microsatellite loci into account. It models a directional mutation process with the rate increasing with array length. It includes the possibility of occasional multistep mutations,

although the mechanism for large changes in the model is recombination which Wierdl *et al.* (1997) thought was less likely to have a role than large slippage events. It is the actions of recombination, as they modelled them, that moderates array size.

This model appears to be the most accurate portrayal of microsatellite mutation available to date, but it should be noted that each locus probably evolves slightly differently. Every model has its limitations and the use of a model involves making large assumptions about the mutational processes experienced by a locus. Nielsen and Palsbøll's warning, that specific models of microsatellite evolution cannot be relied on, should still be remembered.

Overall, microsatellites have many advantages over other molecular markers. Their level of variability makes them appropriate for phylogenetic studies on a wide-range of scales from the individual level (such as issues of parentage and gender) to comparisons between species; this has lead them to be described as the "master of all trades" (McDonald and Potts 1997), although they are perhaps most applicable to population genetics. Their ease of amplification and visualisation by PCR is probably their greatest attraction, particularly in population studies when a large number of samples is being analysed or DNA is only available from a low quality source.

However, there are limitations on their usefulness. It is often necessary to develop a new set of primers for each new species to be studied, as loci isolated from a related species are unlikely to be the most informative in the study species. This can be time consuming and laborious, but as techniques continue to improve it will become easier and as more and more primers are developed, fewer new ones will need to be found. Perhaps the most significant problem is the limited understanding of the mutational processes experienced by microsatellites. Interpretation of microsatellite data to determine the genetic history of populations relies upon understanding the marker evolution. Inaccurate assumptions about microsatellite evolution could lead to the misinterpretation of results and false conclusions being drawn. Microsatellites are not just a modern type of allozyme or an easier alternative to DNA fingerprinting, they are a different type of marker with different characteristics and must be treated as such. New statistical analysis methods and computer packages need to be developed to take the differences in mutational processes into account and inaccuracies in mutation models must be considered when conclusions are drawn. As McDonald and Potts (1997) stated: "Microsatellites are not simply glorified allozymes, and new models will be required to deal with all the ramifications of mutation process and rate that they entail."

1.6 SUMMARY

The aim of this project is to investigate the genetics of a group of red squirrel populations near Antwerp, northern Belgium, that have been affected by habitat fragmentation. Eight fragment populations within a local area were investigated and comparisons were made to larger populations, two also located in northern Belgium and one in Bavaria, another region of northern Europe.

Two different types of molecular marker were examined, the control region of the mitochondrial genome and microsatellite loci found in the nuclear genome. Both markers are assumed to be neutral but both may have been subjected to selective pressures indirectly by hitch-hiking. By analysing and comparing data from markers located in both genomes, it should be possible to generate a more complete picture of the genetic history of these populations than would be possible by the use only one type of marker. The use of microsatellites allows the analysis of several loci which also increases the reliability of the conclusions.

In Chapter 2, the full sequence of the mitochondrial control region in *S. vulgaris* L. is presented and compared to the sequences of other mammal species. Chapter 3 reports the results of PCR-SSCP analysis of control region sequence variation in the study populations. Chapter 4 describes the isolation and testing of a set of microsatellite markers, and chapter 5 presents the results of the application of these markers to the populations. Finally, chapter 6 brings the results of the mitochondrial and microsatellite analyses together and draws conclusions about the genetic history of these populations and the effects of habitat fragmentation.

CHAPTER TWO:

THE RED SQUIRREL MITOCHONDRIAL CONTROL REGION SEQUENCE: A COMPARISON WITH OTHER MAMMALS

2.1 INTRODUCTION

The control region is the only non-coding section of the vertebrate mitochondrial genome. It is referred to as the control region as it contains the main regulatory elements of the genome, controlling replication and translation. It is also known as the D-loop because of the presence of a short section of DNA of heavy (H) strand sequence extending from the origin of H-strand replication, this displaces the original H-strand forming a displacement loop or D-loop. In mammals the control region is bounded by the $tRNA^{pro}$ and $tRNA^{phe}$ genes and ranges in length from 880 to 1400 bp, excluding repeated sequences that can greatly extend its length (Sbisà *et al.* 1997). This nascent H-strand is usually around 600 bp in length but varies even within the same species, for example, in humans there are three different D-loop strands ranging from ~570 – 655 bp, varying in length at their 5' ends (Clayton 1982; Doda *et al.* 1981). The D-loop strands are repeatedly synthesised and degraded (Bibb *et al.* 1981) and at any one time the frequency of the D-loop form found in the population of mitochondrial genomes ranges from <1% in drosophila to >75% in some mouse cells (Clayton 1982).

The mitochondrial control region has been sequenced in many invertebrates and vertebrates and comparisons of the sequences have revealed some conserved features within an otherwise highly variable region. It seems likely that these sequences and structures are in some way involved in the regulation of the replication and transcription of the genome but the details of these mechanisms remain elusive. This chapter describes the characterisation of the red squirrel control region sequence and compares its structure with other mammals for which the control region sequence is known.

2.1.1 The replication of the mitochondrial genome

The mitochondrial genome has a unique asynchronous mode of replication where the two strands have separate and distantly located origins of replication. The processes involved with the replication of the animal mitochondrial genome were reviewed by Clayton (1982) and are summarised here. The mitochondrial genome exists in two forms in animal cells, a superhelical form and the D-loop form which is a covalently closed circle without the superhelical turns but with the displacement loop structure. Of the two DNA strands, the replication of the H-strand proceeds first, primed with a short piece of RNA. It is initiated at the same point in the control region as the short D-loop strands but continues on in the same direction beyond their termination point. The origin of L-strand replication is a distance from the control region, located between the CO1 and ND2 genes, amongst 5 tRNA genes

(Anderson *et al.* 1981). When H-strand synthesis is 67% complete the origin of the L-strand synthesis is exposed and replication is initiated, continuing in the opposite direction to H-strand synthesis.

Necessarily, H-strand synthesis is completed first so new molecules consisting of original L-strands and new H-strands are completed whilst molecules of new L-strands with original H-strands are still being completed. Approximately 100 superhelical turns introduced to the daughter molecules completes the process forming the standard superhelical molecules which may then open up to allow the formation of the D-loop strand and the D-loop form of the molecule. The full cycle takes about 2 hours to complete with the synthesis of a full length daughter strand taking approximately 1 hour, which is a relatively slow rate for DNA replication.

There is no evidence to suggest that existing D-loop strands actually prime H-strand replication; there is a very high turnover of D-loop strands, most are synthesised and then rapidly lost (Anderson *et al.* 1981) so most D-loop strands do not participate in replication. Why so many D-loop strands are maintained when not acting as primers is unclear, there may be a function for the D-loop structure more complex than merely priming replication (Walberg and Clayton 1981). Perhaps the existence of the displacement loop leaves the original H-strand exposed and accessible to the enzymes and proteins involved in transcription and replication.

2.1.2 The structure of the mitochondrial control region

The vertebrate mitochondrial control region has a structure of three regions or domains, originally identified by Walberg and Clayton (1981). The central domain is highly conserved and easily aligned in mammals (Saccone *et al.* 1991). It is surrounded by two peripheral domains that are very variable, showing little similarity when comparisons are made between species. Both these domains contain some short stretches of conserved sequence and preserved secondary structures that may be important in the function of the control region. The two variable domains have been described as the left domain (adjacent to the $tRNA^{pro}$ gene) and the right domain (next to the $tRNA^{phe}$ gene) but it is less confusing to refer to them as the TAS (termination associated sequences) domain and the CSB (conserved sequence block) domain respectively, after the conserved sequences located within them. These domains evolve very rapidly, as much as 5 times faster than the rest of the genome (Ishida *et al.* 1994), and show variation in both length and base composition, the length variation being due to an unusual tendancy to insertion and deletion events and the presence of tandem repeats(Saccone *et al.* 1991).

A generalised structure of the mitochondrial control region is given in figure 2.1. It has been noted that the structure and features of the region resemble other non-coding regions in both prokaryotes and eukaryotes (Sbisà *et al.* 1997). In particular, it is very similar to the origin of bacterial chromosomal replication, OriC which also consists of interspersed blocks of conserved sequence (Gemmell *et al.* 1996). This may reflect the hypothesised prokaryotic origin of the mitochondrial genome or it may simply be that this is an efficient structure for a genomic replication origin.

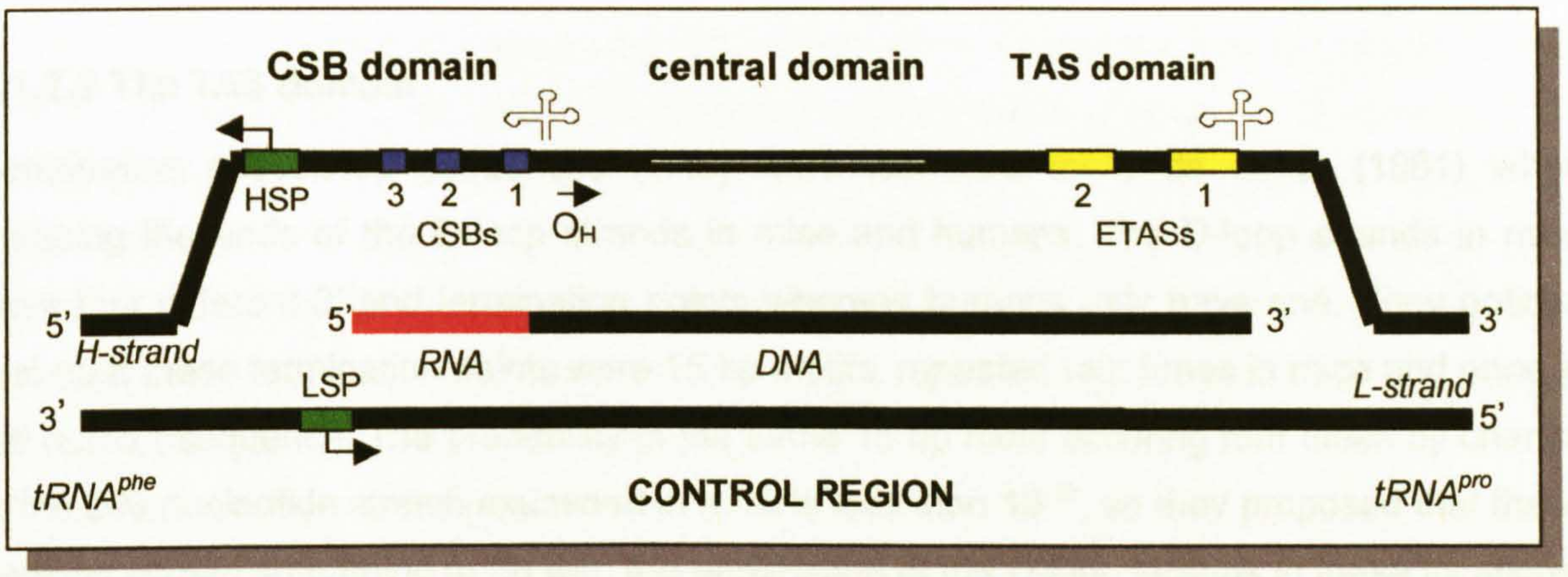


Figure 2.1: The structure of the mitochondrial control region whilst in the D-loop formation showing the conserved sequence features found in mammals including the heavy strand replication promoter (HSP) and light strand promoter (LSP). The locations of two hairpin loop structures are indicated; see text for further explanations.

2.1.2.1 The central domain

Walberg and Clayton (1981) identified a large block of conserved sequence in the middle of the control region when comparing the diverged genomes of mouse and human. The conservation of this domain, which is approximately 300 bp in length, is comparable to the functional genes found in the rest of the genome (Walberg and Clayton 1981; Brown *et al.* 1986). Brown *et al.* (1986), when looking at the base composition of the control region, found that it could be divided up into three domains on the basis of adenine content. The central domain is characterised by a low adenine content, whereas the peripheral domains are relatively adenine rich; this is perhaps associated with reduced evolutionary constraint due to the bias towards the use of adenine in redundant positions. Generally, the L-strand of the whole mitochondrial genome has a reduced guanine content but this is especially true of the fast evolving peripheral domains of control region and again, in contrast, the central domain has a relatively high guanine content (Saccone *et al.* 1987; Saccone *et al.* 1991).

The high level of sequence conservation generally found in the central domain has led to the suggestion that it must be under some functional constraint but as yet no functions have been suggested. Rather, the suggested functions of the control region all implicate the more variable left and right domain sequences rather than the central domain. Open reading frames have been identified (Walberg and Clayton 1981; Saccone *et al.* 1987) but they are short and only a 7 amino acid motif is conserved, and only within the eutherian mammals (Gemmell *et al.* 1996). It seems unlikely that any resulting protein would have an important function. It remains a mystery why this stretch of sequence is so conserved in mammals.

2.1.2.2 The TAS domain

Termination associated sequences (TAS) were identified by Doda *et al.* (1981) when mapping the ends of the D-loop strands in mice and humans. The D-loop strands in mice have four different 3' end termination points whereas humans only have one. They noticed that near these termination points were 15 bp motifs, repeated four times in mice and once in the human sequence. The probability of the same 15 bp motif occuring four times by chance in the 200 nucleotide stretch examined in mice is less than 10^{-15}, so they proposed that these sequences had something to do with the termination of the D-loop strands in these locations. As the motifs are found upstream of the termination points and elongation proceeds past them before stopping they could not suggest any functional mechanism.

It has since been found that some TAS elements form part of a stable hairpin loop structure in this region which is conserved in mammals and could be involved in the arrest of D-loop strand synthesis (Saccone *et al.* 1991). Yet Sbisà *et al.* (1997) was more sceptical about their role, comparisons of postulated TAS elements in many different species show that they are not very well conserved and that they exist at variable distances upstream from the termination sites. Given this, it is difficult to see how any function could be preserved. They went on to describe the existence of "extended TAS" sequences or ETASs which are conserved blocks of sequence about 60 bp in length containing TAS motifs.

2.1.2.3 The CSB domain

The CSB (conserved sequence block) domain contains the origin of H-strand replication (O_H) and the H- and L- strand promoters (HSP and LSP), and is probably the most functionally important domain. Despite its importance it is also the most diverged domain when comparisons are made between species and the most variable in length, ranging in mammals from 221bp in sheep to 763bp in the hedgehog (Sbisà *et al.* 1997). Where they

have been characterised the locations of the HSP and LSP have been reasonably consistent, both being found at the $tRNA^{phe}$ end of the region. O_H cannot be placed with confidence by sequence homology alone, but, as with the HSP and LSP, it has been found consistently to be just beside CSB1.

Three CSBs, between 15 and 20 bp in length, were first described by Walberg and Clayton (1981) along with the central domain. Their proximity to the 5' ends of the D-loop strands led them to hypothesise a role for these sequences in the initiation of H-strand replication. Since then the control regions of many species have been sequenced and only CSB1 has been found in all the vertebrates examined. The most extensive comparison of mammals was carried out by Sbisà *et al.* (1997) who looked at the control regions of 27 species. This revealed that CSB2 is only partially present in five species, although previous studies had stated that this block was absent from some of the same species, presumably because the sequence similarity was so low (Gemmell *et al.* 1996). CSB3 was found to be absent or partially deleted in seven species. There is no apparent correlation between the relatedness of the species and the presence of the CSBs.

These results suggest that only CSB1 is essential although enough of CSB2 may be preserved in all the species examined to maintain function. It seems plausible then that the CSBs have different functions that vary in importance, but, as with the other sequence motifs described, their functions remain elusive.

2.1.2.4 Secondary structures

Brown *et al.* (1986) found sequences that could form thermodynamically favourable and stable cloverleaf-like structures at similar locations in the control region sequences of the four species they examined. One included CSB1 and the other, in the TAS domain, included at least one TAS. A similar stem-loop structure can also be found in the region surrounding the origin of L-strand replication and is essential for the initiation of L-strand synthesis (Brown *et al.* 1986), so it seems likely that these two structures in the control region are functionally important in H-strand synthesis. CSB1 has already been implicated in intiating H-strand synthesis, so the cloverleaf structure including CSB1 may have a similar role to the stem-loop near O_L in initiating strand synthesis. Gemmell *et al.* (1996) found some variation in the conformation of the secondary structures of different mammals, the significance of which is unclear.

Stem-loop configurations are associated with the origins of replication in many different systems. For example, in plasmid ColE1 mutations disrupting the folding of a sequence into a stem-loop lead to replication failure. It seems that the structure is necessary for the association of a primer-precursor RNA with the DNA template prior to cleavage by an RNase to form the RNA primer necessary for DNA synthesis (Saccone *et al.* 1987). The same process forms the RNA primer for H-strand replication so it seems plausible that the stem-loop in the CSB domain functions in the same way as in the ColE1 plasmid.

2.1.2.5 Repetitive sequences

Regions of tandem repeats have been found in many vertebrate and invertebrate species and are largely responsible for the extensive size variation of the control region, differences in repeat length have been implicated in some cases of heteroplasmy within individuals. The most likely mechanism for generating these repeats is strand slippage and mispairing during replication. The repeats are either located in the TAS domain or in the CSB domain between CSB3 and $tRNA^{phe}$, but mostly between CSB1 and CSB2 (Fumagalli *et al.* 1996). Some of the repeats have been reported to form stable secondary structures leading to some suggestions that they may have a function, but no role has yet been found for them.

2.2 METHODS

2.2.1 Sample collection

Whole carcasses of two red squirrels were acquired, one from the Merodese bosen population in northern Belgium (collected by Luc Wauters) and one from Kielder Forest in England (collected by Peter Lurz), the causes of death are unknown. Dissections were performed and a portion of liver tissue removed. The tissue was stored frozen at –20°C.

2.2.2 DNA extraction

Extractions were carried out using the following standard DNA extraction protocol. As carry-over of DNA between samples can be a serious problem, contaminating PCR reactions, great care was taken to clean tools between setting up extractions of different samples.

1. Finely chop a small piece of tissue, about 5mm^2 in size, using a fresh sterile scalpel blade. Homogenise the tissue overnight at 55°C, in a solution of 650µl of SET buffer (0.15M NaCl, 50mMTris-HCl, 1mM Na_2EDTA pH 8.0) with 0.005% SDS, 0.0015% Triton X-100 (final concentrations) and 200µg proteinase K.

DNA was extracted from the digested tissue using the phenol/chloroform method:

2. Add 0.5ml freshly prepared saturated phenol solution (phenol crystals dissolved in an equal volume of 1M Tris-HCl, pH 8.0) and mix by gentle rotation on a rotary mixer for up to 30 minutes. Spin sample at 13,000rpm in a microfuge for 10 minutes and then transfer the aqueous (upper) layer to a fresh eppendorf using a cut-off eppendorf tip to prevent shearing of high molecular weight DNA.

3. Repeat step 2 twice, using phenol/chloroform/iso-amyl alcohol (25:24:1v/v) and then chloroform/iso-amyl alcohol (24:1v/v) in place of phenol, each time transferring the aqueous layer to a fresh eppendorf.

The DNA was ethanol precipitated:

4. Add 2 volumes of cold (-20°C) absolute ethanol and 1/10th volume 3M Sodium acetate, incubate at –80°C for 30 minutes. Spin in a microfuge at 13,000rpm for 10 minutes to pellet the precipitated DNA, then carefully pour off the ethanol.

5. Rinse the DNA pellet with 75% ethanol and spin again for 5 minutes.

6. Remove the ethanol and leave the pellet to air-dry or dry in an oven at 55°C.

7. Resuspend pellet in 50µl TE (1mM Na_2EDTA, 10mM Tris-HCl pH 8.0) by overnight incubation at 55°C.

The DNA samples were stored frozen at –20°C.

2.2.3 Visualisation of DNA by electrophoresis and Ethidium bromide staining

DNA has a negative charge so DNA molecules can be separated by pulling them through an appropriate matrix, such as agarose, by applying an electric current. Smaller DNA molecules move faster through the gel, leaving the larger molecules behind them. The molecules then appear as bands or smears when stained and visualised.

DNA from extractions, PCR reactions etc. can be visualised quickly and effectively in agarose gels at various concentrations. For a 0.8% gel (appropriate for a whole genome DNA extraction) 0.8g of Seakem LE agarose (Flowgen) was dissolved in 100ml of 1xTAE buffer (40mM Tris, 20mM Sodium acetate, 1mM Na_2EDTA pH 8) by bringing it briefly to the boil in a microwave. It was allowed to cool to 55°C in a waterbath and then poured into a perspex gel base containing an appropriate well comb. Once set, the gel was transferred to a wide mini sub-cell horizontal electrophoresis tank (Biorad) containing enough 1xTAE buffer to cover the gel. Small volumes (just a few μl) of the DNA samples were mixed with an equal volume of 2x loading buffer (4% Ficoll (R) 400, 40mM Na_2EDTA ph 8, 0.05% bromophenol blue, 0.05% xylene cyanol FF) and loaded into the wells of the gel. The gel was exposed to 100V for approximately 30 minutes, or for as long as necessary.

The TAE buffer contained 0.5μg/ml Ethidium bromide (EtBr) which binds to the DNA as it runs through the gel. Alternatively, EtBr can be added to the agarose when it is dissolved or the gel can be stained in a solution containing EtBr after being run. The DNA was then visualised in the gel by exposing it to UV light which causes the EtBr to fluoresce and reveal the positions of DNA strands as bands in the gel. An appropriate quantity of a size marker ladder was usually run in one of the wells next to the samples under examination, such as the 1 kb DNA ladder (Gibco/BRL). An estimation of the amount of DNA in each band could be made from the intensity of the band under UV light when compared to the known amount of the size marker run. The gels were photographed using an EASY RH enhanced analysis system (Herolab).

2.2.4 Amplification of the control region

The control region of the mitochondrial genome was amplified using the primers t^{pro} and t^{phe} which bind to the $tRNA^{pro}$ and $tRNA^{phe}$ genes respectively. The primer sequences are:

t^{pro} (5'- CTACCTCCAGCTCCCAAAGC -3')

and t^{phe} (5'- ATCTAAGCATTTTCAGTGCT -3')

Primer t^{pro} had been developed previously in the laboratory for use in birds, but as the tRNA genes are very highly conserved it was applicable to the red squirrel. The primer t^{phe} was developed by aligning the $tRNA^{phe}$ genes of the mouse, rat, bank vole and human, and selecting an appropriate 20bp portion.

The $tRNA^{phe}$ gene sequences were downloaded from the GenBank database, release 99, with the accession numbers: mouse (*Mus musculus*) - J01420, rat (*Rattus norvegicus*) - X14848, bank vole (*Clethrionomys rufocanus*) - D42091 and human (*Homo sapiens*) - V00662. They were aligned using the program ClustalW (section 2.2.9).

The PCR reaction was optimised by trial and error, adjusting the annealing temperature and the final concentrations of Magnesium chloride, nucleotides, primers and template. The amplification of the control regions was carried out using the following reaction mixture given in table 2.1.

Reagent	Conc^{n}.	Quantity	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	5 µl	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	4 µl	0.1 mM
Magnesium chloride (Advanced Biotechnologies)	25mM	2 µl	1 mM
primer t^{pro}	10µM	4 µl	0.8 µM
primer t^{phe}	10µM	4 µl	0.8 µM
Taq DNA polymerase (Promega)	5 U/ µl	0.4 µl	2 U
template	~ 2-6 ng/µl	8 µl	~ 0.3-1 ng/µl
sterile distilled water		22.6 µl	

Table 2.1: The PCR reaction mix used to amplify the mitochondrial control region of the red squirrel.

The reactions were run on a programmable thermocycler (PTC-100, MJ Research, Inc.) using the following cycling program:

Step	Temperature (°C)	Time (minutes)
1	94	2
2	94	1
3	49	1
4	72	1 1/2
5	Go to step 2	29 more times
6	72	5

The PCR products were visualised in agarose gels as in section 2.2.3, using 1.5% agarose. The whole product was run through the gel and clean bands were excised using a scalpel blade. The product DNA was extracted from the agarose matrix using the Qiagen gel extraction kit.

2.2.5 The cloning of the control region

The amplified control region sequence was ligated into the pGEM-T vector (Promega, following the supplied protocols) which was then used to transform some competent bacterial cells of *E. coli,* strain JM101, using protocols based on Sambrook *et. al.* (1989) and those developed by researchers in the Department of Genetics, University of Nottingham. Competent cells were made using the Calcium chloride method and transformed using the heat shock technique. Successfully transformed colonies were selected using the *lac Z* selection system.

2.2.5.1 The ligation reaction

The ligation was carried out using the pGEM-T vector system (Promega). 1μl each of 10x T4 ligase buffer, pGEM-T vector (50ng) and T4 DNA ligase (3 U/μl) were mixed with 7μl of the cleaned PCR product and incubated at 15°C for 3 hours. A positive control reaction was also set-up (using the control insert DNA supplied with the kit) and the completed reactions were stored at –20°C.

2.2.5.2 Preparation of the competent cells

1. Inoculate 5ml of LB medium (1% bacto-tryptone, 0.5% bacto-yeast extract, 1% Sodium chloride pH 7) with *E.coli* JM101 (from glycerol stock) and grow the culture at 37°C overnight.

2. Inoculate 10ml of fresh LB with 100µl of the overnight culture and incubate at 37°C, agitating constantly, for a few hours, until the optical density (wavelength of 650nm) is between 0.3 and 0.6. The cells should be in the log phase of growth at this stage, care should be taken not to grow the culture beyond this phase.

3. Transfer the culture to a large Falcon tube, cool the cultures on ice for 5 minutes and centrifuge the cells in a benchtop centrifuge at 6000rpm, 4°C for 5 minutes. Pour off the medium and leave the tubes upside down for a few minutes to drain off any remaining medium.

4. Resuspend the cells in 1/2 volume (5ml) cold 50mM Calcium chloride and incubate on ice for 20 minutes.

5. Centrifuge the cells again at 6000rpm, 4°C for 5 minutes and pour off the supernatant.

6. Resuspend the cells in 1/10 volume (1ml) cold 50mM Calcium chloride. Incubate the cells on ice for between 1 and 24 hours, the cells will only be good for use during this period.

2.2.5.3 Transformation of competent cells

200µl of competent cells were added to the defrosted ligation reaction in an eppendorf and the following standard method to induce transformation was employed.

1. Incubate the cells on ice for 20-30 minutes.

2. Heat shock the cells at 42°C, by incubation in a waterbath, for 2 minutes exactly.

3. Chill the cells on ice again for 5 minutes.

4. Add 200µl LB broth and incubate at 37°C for 1 hour.

2.2.5.4 Selection of recombinant cells

After 1 hours growth the cells have recovered sufficiently to be expressing the ampicillin resistance gene and to grow when plated out on agar plates. The T-vector contains an ampicillin resistance gene so when plated on agar containing ampicillin, only cells successfully transformed and containing the T-vector will grow. The insertion site within the T-vector is within the *lac Z* operon. If the ligation is successful and the target DNA is successfully incorporated into the vector then the *lac Z* operon should be disrupted. This prevents the production of the β-galactosidase enzyme which breaks down X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside). If X-gal is broken down the colonies turn a dark blue colour, if not, they remain yellow/white. This form of selection is known as blue-white screening.

1. Melt 300ml agar (LB agar plus bacto-agar (15 g/L)) in the microwave, allow to cool until hand hot and add 15mg ampicillin (0.005%, final concentration). This should be enough agar to pour at least 12 plates. Once set, dry the plates thoroughly in an oven and allow to return to room temperature. The plates can be stored by refrigeration until required.

2. Add 50μl IPTG (isopropylthio-β-D-galactoside) and 50μl X-gal to 100μl of transformed culture and spread on an agar plate using sterile technique. Incubate plate upside-down overnight at 37°C.

White colonies that grow on the selective plates should contain the vector carrying the target DNA.

A few white colonies were used to inoculate some LB broth containing ampicillin (0.005%, final concentration) and incubated at 37°C. After a few hours growth, 8μl of the culture was used as a template in a PCR reaction to amplify the mitochondrial control region as in section 2.2.4. Successful amplifications indicated the presence of the control region DNA within the vector or plasmid DNA.

2.2.5.5 Culture storage

Cultures of bacterial cells shown to contain plasmids carrying the red squirrel control region were stored as glycerol stocks. 1.5ml of 80% glycerol was added to 2.5ml culture and stored frozen at –20°C. These stocks should never be allowed to thaw completely.

2.2.6 Plasmid preparations

Large volumes of very clean plasmid templates containing the control region DNA were required for repeated sequencing reactions, so a protocol developed by Nick Harvey in the laboratory to meet these requirements (based on the methods described in Sambrook *et al.* 1989) was used. It employs many different cleaning methods, each of which lead to the loss of a small amount of plasmid template, but they ensure that the plasmid template is of the highest quality.

Harvesting plasmid containing bacteria

1. Inoculate 15ml of LB media containing 0.75mg ampicillin with the bacterial culture carrying the vector from glycerol stock. Grow the culture overnight at 37°C agitating constantly.

2. Transfer the culture to a large Falcon tube and spin in a benchtop centrifuge at 4000rpm for 10minutes at 4°C. Pour off the supernatant broth to leave a large pellet of culture.

Alkaline lysis

The following steps burst open the cells and separate the plasmid DNA (Sambrook *et al.* 1989).

3. Resuspend the bacterial pellet in 600µl of solution 1 (50mM glucose, 25mM Tris.Cl pH 8, 10mM EDTA pH 8) and leave at room temperature for 5 minutes.

4. Add 1.2ml of solution 2 (0.2M NaOH, 1% SDS), shake gently and incubate on ice for 5 minutes.

5. Add 900µl of solution 3 (3M potassium and 5M acetate solution made from 60ml 5M potassium acetate, 11.5ml glacial acetic acid, 28.5ml H_2O) shake well and leave on ice for 15 minutes.

6. Spin solution in a benchtop centrifuge at 4000rpm for 15 minutes at 4°C and transfer the supernatant to a fresh Falcon tube.

7. Precipitate the DNA by adding 2.5 volumes of 100% ethanol and incubating at –80°C for 30 minutes. Spin the solution at 4000 rpm for 10 minutes at 4°C to produce a pellet. Rinse the pellet in 70% ethanol and then air -dry before resuspending in 500µl TE buffer. Transfer the solution to a 1.5ml eppendorf.

Purification of plasmid DNA

The following steps remove the RNA and any impurities by further precipitation steps.

8. Add 500µl of cold 5M Lithium chloride, mix well and leave on ice for 10 minutes. (Lithium chloride precipitates high molecular weight RNA.) Spin the solution in a microfuge at 13000 rpm for 10 minutes and transfer the supernatant to a clean tube.

9. Precipitate the DNA with an equal volume of isopropanol, mix well and spin at 13000 rpm for 15 minutes. Remove the supernatant and rinse the pellet with 70% ethanol. Dry the pellet and resuspend in 250µl TE buffer.

10. Add 1µg RNase and incubate at 37°C for at least 30 minutes.

11. Add an equal volume (250µl) of 1.6M NaCl containing 13% (w/v) polyethylene glycol (PEG 8000), mix well and spin for 15 minutes at 13000 rpm. This differentially precipitates the plasmid DNA. Remove all traces of supernatant and resuspend the pellet in 400µl TE.

12. Perform three extractions, one with an equal volume of phenol, one with phenol:chloroform and one with just chloroform (as in steps 2 & 3 of the DNA extraction protocol, section 2.2.2).

13. Precipitate the clean plasmid DNA by adding 1/10th volume 3M Sodium acetate and 2x volume cold 100% ethanol, incubate at –80°C for 30 minutes and spin at 13000rpm for 15 minutes. Remove the supernatant, rinse the pellet with 70% volume, spin again briefly, remove supernatant and dry the pellet thoroughly. Resuspend the DNA in 50µl TE.

2.2.7 Manual sequencing of plasmid template DNA

Primers SP6 (5'-TATTTAGGTGACACTATAG-3')
and T7 (5'-TAATACGACTCACTATAGGG-3') (Promega)

are designed to bind to T-vector DNA either side of the insertion site. These primers were used along with t^{phe} and t^{pro} to sequence the inserted DNA. Using primers that work in both directions on the same template gives the sequence of each strand separately, allowing more of the insert to be sequenced and the sequences to be compared and checked for accuracy.

The following protocol for manual sequencing was found to only produce around 200 - 300 bp of readable sequence from each primer. To cover the full length of the insert three more primers were designed as required, binding further into the control region. These primers were RSCR1 (5'-CACATAGTACATAGACATTAGG-3'),
RSCR2 (5'-GACTTGAGAGAGTGGAGCTCTG-3')
and RSCR3 (5'-TATGCTTCCACTCACCATTGG-3')

They bind to the control region 210bp, 794bp and 640bp into the sequence respectively (from the 5' end of the L-strand, figure 2.2). RSCR1 and 3 bind to the H-strand and RSCR2 binds to the L-strand.

Sequencing was carried out using T7 sequencing mixes (Pharmacia Biotech) and the accompanying protocols, which are broadly based on the Sanger sequencing method (Sanger *et al.* 1977).

Preparation of the plasmid template

1. Denature the template by mixing 8µl of template DNA (in TE buffer) with 24µl sterile distilled water (SDW) and 8µl 2M Sodium hydroxide and left at room temperature for 20 minutes.

2. The denatured DNA is then precipitated by the addition of 4µl SDW, 7µl 3M Sodium acetate and 120µl cold 100% ethanol. Incubate at –80°C for 15 minutes and spin at 13000rpm for 15 minutes to pellet the DNA. Discard all the supernatant, rince the pellet (carefully as it is likely to be too small to be visible) with 70% ethanol, spin again briefly and discard supernatant. Oven dry the pellet completely.

Primer Annealing

3. Resuspend the pellet in a mixture of 10µl SDW, 2µl annealing buffer and 2µl primer. Incubate at 65°C for 5 minutes and 37°C for 20 minutes then allow to stand at room temperature for at least 10 minutes.

Sequencing reactions

Preliminaries (per reaction)

4. In a 0.5ml eppendorf labelled "E" dilute 0.4µl T7 DNA polymerase (9.8 U/µl, 4 units per reaction) with 2.1µl dillution buffer. Store on ice.

5. Label 4 eppendorfs "G", "A", "T" and "C" and pipette 2.5µl of each of the four terminating mixes into the appropriate tube.

6. Into an eppendorf labelled "L" pipette 3µl labelling mix A (appropriate for use with the radioactive isotope [α-^{35}S]dATP), 2µl of the diluted enzyme from tube "E" and 1µl (0.37Mbq) [α-^{35}S]dATP. Mix by gentle pipetting.

Labelling reaction

7. Add the 6µl mix from tube "L" to the primer and template . Leave at room temperature for 5 minutes.

Termination reactions

8. Take 4.5µl aliquots from the labelling reaction and add to each of the terminating mixes in the tubes labelled "G", "A", "T" and "C". Incubate each for 5 minutes at 37°C.

9. Add 5µl stop solution (including a loading dye) to each reaction tube. Spin the tube briefly in a microfuge to ensure all the reaction mix is at the bottom of the eppendorf.

The stopped reactions were stored at 4°C until run out on a polyacrylamide gel.

2.2.8 Running polyacrylamide gels

Polyacrylamide gels were run on Sequi-Gen II gel rigs (BioRad) according to the manufacturer's instructions and using the following protocol:

1. Make up 500ml of a 6% polyacrylamide solution according to the following recipe:

		final conc^{n}
40% (w/v) bis-acrylamide	75ml	6%.
1xTBE (0.09M Tris-borate, 0.002M EDTA (pH 8))	250ml	0.5x
urea	210g	42%

Dissolve the urea in the TBE and acrylamide by incubating in a 55 C waterbath and mixing regularly. Make up to 500ml with SDW, filter through (Whatman) filter papers and store at 5°C. Refrigerate the solution for at least 1 hour before use.

2. Pour a thin polyacrylamide gel according to the gel rig manufacturer's instructions with a mix of 100ml 6% polyacrylamide solution, 250µl 10% Ammonium persulphate (GibcoBRL) and 100µl Temed (Sigma). Allow the gel to set.

3. Connect the gel to a power pack according to the instructions and fill the rig with 0.5xTBE buffer.

4. Remove the gel comb and clean the top of the gel by blowing buffer along the surface with a Pasteur pipette. This is important for removing any remaining loose polyacrylamide. Replace comb and clean out wells. Load 2-5µl of loading buffer into a few wells to ensure the comb is tight fitting and the wells do not leak.

5. Pre-warm the gel by running at a constant 80W for 30-60 minutes. Before loading the samples blow out any urea crystals that may have precipitated in the wells using a Pasteur pipette.

6. Denature samples by heating to 95°C for 5 minutes, move samples onto ice and immediately load 2-5µl (depending on the well size) of each sample into the wells of the gel. The four termination reactions of each sequencing reaction were loaded in adjacent wells in the order "G,A,T,C".

7. Run the gel for 1 1/2 - 5 hours, as appropriate for the size of product to be visualised, at either a constant 80W or a constant 50°C.

8. Transfer the gel onto a sheet of 3MM chromatography paper (Whatman), cover with Saranwrap and dry on a vacuum gel drier at no higher than 80°C until dry (1-2 hours).

9. Put the dried gel in an autoradiograph cassette (when ^{35}S has been used as the radioactive label, the Saranwrap must be removed) and place a sheet of medical x-ray film (Fuji) on top in a dark room. Leave the film to be exposed for 1 day to several weeks, as required.

10. Develop the x-ray film in LX24 developer (Kodak) for 4 minutes, rinse in water, then fix in Hypam rapid fixer (Ilford) for 4 minutes. Leave to dry, after which the gel can be read.

2.2.9 Sequence analysis

DNA sequences were analysed using the DNA analysis programs of the Computer and Molecular Evolution Laboratory at the University of Nottingham (CMELUN). Control region sequences for the mouse and rat were obtained from GenBank, release 99, under the accessions numbers J01420 (mouse, *Mus musculus*) and X14848 (rat, *Rattus norvegicus*). Comparisons were made between sequences by alignment using the programs ALIGN (Bill Pearson) and CLUSTAL-W (Thompson *et al.* 1994). Base composition analysis was carried out using the CODONS program. The alignments of the central domain, CSB and ETAS sequences carried out by Sbisà *et al.* (1997) were downloaded from the web site http://www.ba.cnr.it/dloop.html and aligned with the red squirrel sequences by hand.

2.3 RESULTS

2.3.1 The red squirrel control region sequence

The full control region sequences of a Belgium and a British red squirrel were obtained. The sequence from the Belgium individual is given in figure 2.2, it is 1056 bp in length. 9 sequence differences were observed between the two sequences: 8 transition-type changes (6 involving C and T and 2 involving A and G) and a single base pair deletion; most of the sequence differences were found in the TAS domain near the central domain. This bias towards transitions is typical in molecular evolution (Li and Graur 1991). The conserved central domain is indicated and was clearly identified when compared with the alignment of mammalian central domains carried out by Sbisà *et al.* (1997). The lengths of the three domains are compared with the average for mammals (calculated from the data given in Sbisà *et al.* 1997's study of 27 mammalian species) in table 2.2. The red squirrel domain lengths are close to the average for mammals. There are no repetitive sequences in the control region of the red squirrel mitochondrial genome.

	average	range	*S. vulgaris*
TAS domain	322	209 – 412	388
Central	315	300 – 328	318
CSB domain	389	221 - 763	350

Table 2.2: A table comparing the lengths of the red squirrel control region domains (bp) to the average lengths and the range in lengths of the three domains found in 27 other mammalian mitochondrial control regions (calculated from the data in Sbisà *et al.* (1997), excluding regions of tandem repeats where they exist).

Alignments of the complete control region sequences of red squirrel, mouse and rat (mouse and rat sequences obtained from GenBank) were carried out and percentage homologies are given in table 2.3. When aligning the domains, the differences in conservation can be clearly seen. The central domain shows a high level of conservation, whereas the TAS and CSB domains show much less homology. This illustrates the differences previously seen between the conserved central domain and the highly variable peripheral domains (Walberg and Clayton 1981).

S.vulgaris
10 20 30 40
GTACTTACTTGACCAATCCCTCACTAAACTGACTCTCATG

S.vulgaris
50 60 70 80
TACCTATCATTAATTCCTCTCACATTGATGTCTATGTAAT

S.vulgaris
90 100 110 120
TCGTGCATTAATGCATGTCACATTAATTAAATGGTACAGT

S.vulgaris
130 140 150 160
ACATAATAATAAGAAAGTACATAGAACATATCATGTTTAA

S.vulgaris
170 180 190 200
TCAACATTAAAACCTCCACCACATGCTTATAAGCATGCAC

S.vulgaris
210 220 230 240
ATTAATACTCACATAGTACATAGACATTAGGATTTAACCC

S.vulgaris
250 260 270 280
TCCACATTTAAAGTCCTACAACATGGATATTCTTTACCCC

S.vulgaris
290 300 310 320
ATTACATTCTTTATATTGCATAGCACATAACATTCACTGG

S.vulgaris
330 340 350 360
CGGCACATACCCCATTTAAGTCATAAACCTTCCTCGTCCA

S.vulgaris
370 380 390 400
AATGACTATCCCCTTCCAACGGTGGTCTCTTAATCTACCT

S.vulgaris
410 420 430 440
ACCTCCGTGAAATCATCAACCCGCCCGATACGTGTCCTCT

S.vulgaris
450 460 470 480
TCTCGCCTGGATCCCATTTAACTTGGGGGTGACTAACTAT

S.vulgaris
490 500 510 520
GCTCTTTGACAGGGCATCTGGTTCCTACCTCAGGGCCATG

S.vulgaris
530 540 550 560
TAATGCGTTATCGCCCATACGTTCCCCTTAAATAAGACAT

S.vulgaris
570 580 590 600
CACGATGGATTAGTTCCATTCTAGCCCGTGACCCAACATA

S.vulgaris
610 620 630 640
ACTGCACTGTCATGCCTTTAGTGGTTTTTATTTTTGGGGT

S.vulgaris
650 660 670 680
ATGCTTCCACTCGCCATTGGCCGTCAGAGGCCCCGACGCA

S.vulgaris
690 700 710 720
GTCAATTCAATTGTAGCTGGACTTTAGGTCATTATTCTTT

S.vulgaris
730 740 750 760
ACTCGCATATACATCCAAGGTGTTCATAATATTCATGCTT

S.vulgaris
770 780 790 800
GATGGACATATTAATTTTTAATTCACAATTTAACAAGAGC

Figure 2.2: The complete sequence of the red squirrel mitochondrial control region.

Figure 2.2: The complete sequence of the red squirrel control region (continued) showing the heavy (H) strand sequence.Two possible ETAS1 sequences are shown in red ink (GATC), the proposed ETAS2 sequence is in blue (GATC) and several possible TAS sequences are identified by a black line above the sequence (——). The central domain is in bold font (**GATC**), and the conserved blocks are indicated with a coloured line above: **A** with a green line (——), **B** a red line (——) and **C** a blue line (——). In the CSB domain, CSB1 is highlighted in yellow (GATC) and CSB2 in green (GATC).

comparison of:	control region	TAS domain	central domain	CSB domain
red squirrel & mouse	59.1%	55.5%	71.7%	50.4%
red squirrel & rat	58.4%	55.5%	69.6%	50.5%
mouse & rat	77.0%	80.7%	85.1%	67.5%

Table 2.3: The percentage homology seen when pairwise comparisons are made of the whole control region and the three domain sequences of red squirrel, mouse and rat.

2.3.2 Base composition

An analysis of the base composition of the red squirrel sequence is given in table 2.4, the composition of each domain is considered as well as the region as a whole. The control region of mammals is usually A/T rich (Ishida *et al.* 1994) and, although the red squirrel sequence does show the expected high AT and low GC content, it also has a high TC content reflecting the overall T rich nature of the sequence.

BASE	control region	TAS domain	central domain	CSB domain
T	33%	31%	30%	39%
A	28%	32%	20%	31%
C	27%	27%	30%	22%
G	12%	10%	20%	8%
A+T	61%	62%	51%	77%
G+C	39%	38%	49%	33%
A+G	40%	42%	41%	39%
T+C	60%	58%	59%	61%

Table 2.4: The base composition analysis of the control region sequence of the red squirrel, *S. vulgaris*, showing the composition of the entire region and each of the domains.

The most notable feature of the base composition is the very low G content, only 12% overall but higher in the central domain than in the peripheral domains. This pattern is also seen in other mammal sequences, the low G content of the L-strand and the corresponding low C content of the H-strand is a feature of mammalian mitochondrial genomes (Saccone *et al.* 1987; Saccone *et al.* 1991).

The nucleotide thymidine (T) dominates the CSB domain in red squirrels which is unusual in mammals. Sbisà *et al.* (1997) found that in most mammals A is at the highest frequency in both the peripheral domains, except in the primates where C is the highest frequency base in the CSB domain. In three species (pygmy sperm whale, bowhead whale and sheep) T is marginally the most common base in the CSB domain, but only in the platypus is T at a comparable frequency to red squirrels at 42%. These differences in distribution in the peripheral domains are probably simply the result of chance during uninhibited molecular evolution.

2.3.3 The conserved features of the control region

With only the sequence to examine it is not possible to identify the origin of H-strand replication or the H- and L- strand promoters, but other sequence features can be identified by alignment with other mammalian sequences.

2.3.3.1 The central domain

Gemmell *et al.* (1996) identified three subsequences, labelled "A", "B" and "C", within the conserved central domain which showed a homology of >70% in the ten mammals he compared. The three blocks are marked on the red squirrel sequence given in figure 2.2. These blocks are obvious within the alignment carried out by Sbisà *et al.* (1997) but other highly conserved blocks are also visible and it is not clear that the three blocks identified by Gemmel *et al.* (1996) really justify distinction from within a generally highly conserved region.

2.3.3.2 The TAS domain

Two putative TAS sequences can be found and are indicated in figure 2.2. The alignment of these two sequences to the two original TAS sequences identified by Doda *et al.* (1981) is shown in figure 2.3. Both motifs show a high level of homology to the human and mouse TAS sequences, which is further increased if the second block can be either of the pyrimidines, cytosine or guanine, as in the mouse consensus.

human	TAACCCAAAAATACA
mouse	TAAYY AAATTACA
red squirrel	TAAG AAAGTACA
	TA CCCC ATTACA

Figure 2.3: The alignment of two proposed TAS sequences located in the red squirrel control region with the original sequences identified by Doda *et al.* (1981). The mouse sequence is a consensus, Y indicates the presence of either pyrimidine.

Neither of these TAS sequences are located within ETAS sequences, as was also found to be the case in the mammals examined by Sbisà *et al.* (1997). In the red squirrel, two possible ETAS1 sequences were identified (given in red ink in figure 2.2) and one possible ETAS2 sequence (blue ink in figure 2.2). The alignments of these sequences with those of the 27 mammal species included in Sbisà *et al.* (1997) are shown in figure 2.4 and 2.5 respectively. A list of the species included in this comparison, giving both scientific and common names, is given in table 2.5.

Both sequence blocks have an average length of 59 bp in the 27 mammals compared. ETAS1 is the most conserved amongst the species considered, this is illustrated by the number of determined nucleotides in the consensus sequence, the ETAS2 consensus has more undetermined nucleotides than ETAS1. The figures at the end of each species' sequence are a count of the number of nucleotides each sequence has in common with the

consensus, the average for ETAS1 is 47 whereas the average for ETAS2 is only 39. The proposed red squirrel ETAS2 sequence has a length of 60bp and shares 36 nucleotides with the consensus sequence, which is slightly below average but higher than several of the other mammal species. For example, the fat dormouse (*Glis glis*) only shares 29 nucleotides out of a sequence 55 bp in length, with the consensus. If this sequence can be accepted as conserved, then the proposed ETAS2 for red squirrels must also be accepted.

Figure 2.4 shows two possible sequences for ETAS1 in red squirrels, one 62 bp in length and the other 60 bp. The second sequence is marginally more similar to the consensus sequence than the first one (sharing 44 nucleotides as opposed to 40) but this difference is too little to really be able to determine which is most likely to be an ETAS1 sequence. If a stable cloverleaf-like secondary structure that overlapped with one of these two sequences could be detected in this region of the red squirrel, that could be taken to indicate which sequence represents the ETAS1 as Sbisà *et al.* (1997) found that some of the ETAS1 sequence was included in a cloverleaf structure in all the mammals considered. This would require further analysis employing specific computer programs designed to detect potential secondary structures in DNA sequences.

Organism	Common Name	Organism	Common Name
Hylobates lar	gibbon	*Ovis aries*	sheep
Homo sapiens	human	*Diceros bicornis*	black rhinoceros
Pan troglodytes	common chimpanzee	*Equus asinus*	donkey
Pan paniscus	pygmy chimpanzee	*Equus caballus*	horse
Gorilla gorilla	western lowland gorilla	*Crocidura russula*	white toothed shrew
Pongo pigmaeus	orangutan	*Sorex araneus*	Eurasian common shrew
Mus musculus	house mouse	*Erinaceus europaeus*	European hedgehog
Rattus norvegicus	Norway rat	*Halichoerus grypus*	gray seal
Glis glis	fat dormouse	*Phoca vitulina*	harbor seal
Cephalorinchus commersonii	Commerson dolphin	*Mirounga angustirostris*	elephant seal
Kogia breviceps	pygmy sperm whale	*Felis catus*	domestic cat
Balaena mysticetus	bowhead whale	*Didelphis virginiana*	North American opossum
Bos taurus	cow	*Sciurus vulgaris*	Eurasian red squirrel

Table 2.5: The latin and common names of the 27 mammals included in the analysis of the mitochondrial control region carried out by Sbisà *et al.* (1997) and included in the alignments shown in figures 2.4, 2.5 and 2.6.

```
                          10                  30                  50                  73
h.lar                AC...TCTA.TGTAC.TTCGTACATTA..CTG.CCAGTCCCCATGC.ATA..T.TGT.ACA...GTACTATAA   42
h.sapiens            AC.C.GCTA.TGTAT.TTCGTACATTA..CTG.CCAGCCACCATGA.ATA..T.TGT.ACG...GTACCATAA   42
p.troglodytes        AC.C.GCTA.TGTAT.TTCGTACATTA..CTG.CCAGCCACCATGA.ATA..T.CGT.ACA...GTACCATAT   41
p.paniscus           AC.C.GCTA.TGTAT.TTCGTACATTA..CTG.CCAGCCACCATGA.ATA..T.TA...CATA.GTACTATAA   45
g.gorilla            ATTA.TC.A.TGTAT.GTCGTGCATTA..CTG.CCAGACACCATGA.ATAA.TGTA...CA...GTACCATAA   40
p.pigmaeus           GCGG..CTA.TGTAT.TTCGTACATTC..CTG.CCAGCCAACATGA.ATA..T.CAC.CCA.....ACA.CAA   37
m.musculus           ACAT.T.TA.TGTAT.ATCGTACATTAAACT.ATTTTCC.CCAAGC.ATA..T..AA.GC.TA.GTACATTAA   53
r.norvegicus         ACAT.T.TA.TGTAT.ATCGTACATTAAATT.ATTTTCC.CCAAGC.ATA....TAA.GCAT..GTA.ATATA   52
g.glis               CCAT.AATA.TGTAATATCC.ACATGAAACT.ATTTACC.CCATGA.ATA..T.CAA.TCA.A.GTACAT.AT   45
o.cuniculus          ACAT.ACTA.TGTTTAATCGTGCAT.AAATTCCTCATCCCCCATGA.ATAA...TAA.GC.TA.GTACATT.A   50
c.commersonii        ACAT.GCTA.TGTATTATTGTGCATTCATTT.ATTT....CCATACGATAAGTTAAA.GCCC..GTA..TTAA   46
k.breviceps          ACAT.GCTA.TGTATAATAGTGCATTCAATT.ATTTT...CCACACGAGAAGTTAAA.GCCC..GTA..TTAA   45
b.mysticetus         ACAT.GCTATTGTATAATCGTGCATTCAATT.ATTTT...CACTACGGGAAGTTAAA.GC.TC.GTA.ATGAA   45
b.taurus             ACAT.AATA.TGTAT.ATAGTACATTAAATT.ATATGCC.CCATGC.ATA....TAA.GCA.A.GTACATGAC   51
o.aries              ACAT.TATA.TGTAT.AAAGTACATTAAAT.GATTTACC.TCATGC.ATA....TAA.GCAC..GTACATAAT   50
d.bicornis           CCGG..GTA.TGTAT.ATCGTGCATTAAATTG.TTTGCC.CCATGC.ATA....TAA.GCATATGTACTACAT   49
e.asinus             ATAC.CCTA.TGTAC.ATCGTGCATTAAATTG.TTCACC.CCATGA.ATAA...TAA.GCAT..GTACATAAT   50
e.caballus           ATGG.CCTA.TGTAC.GTCGTGCATTAAATTG.TCTGCC.CCATGA.ATAA...TAA.GCAT..GTACA.TAA   49
c.russula            ACAT.ACTA.TGTAT.ATCGTACATTATATTC.TCGTCC.CCATA.CATA....TAA.GCA.A.GTACTATTA   48
s.araneus            ACAT.TTTA.TGTAT.ATCGTACATTAAGTTC.TATTCC.ACATGC.ATA....TAA.GCAT..GTACAT.AC   51
e.europaeus          ACAT.TAAATTATAT.TTT.TACTATATATTTATGTAAT.TCTAGC.ATA....TAA.GCAT..GTACATTAA   42
h.grypus             ACCT.CCTA.TGTAT.ATCGTGCATTAA..TGGTTTGCC.CCATGC.ATA....TAA.GCAT..GTACATGAA   53
p.vitulina           ACCC.CCTA.TGTAT.ATCGTGCATTAA..TGGTTTGCC.CCATGC.ATA....TAA.GCAT..GTACATGAA   52
m.angustirostris     GC...CCTA.TGTAT.ATCGTGCATTAA..CGGTTTGCC.CCATGC.ATA....TAA.GCAT..GTACATAAA   50
f.catus              ACAT.ACTA.TGTAT.ATCGTGCATTAA.TTGCTAGTCC.CCATGA.ATA..T.TAA.GCAT..GTACAGTAG   52
d.virginiana         TTGT.CCTA.TGTAT.ATAGTACA.TGGATTTATTTACC.CCTAGC.ATA..TATTA.TAAT..RTACATTAA   45
o.anatinus           ACAT.TATA.TGTAT.ATAGTACATTAA.TTGCATGTCC.GCATGG.ATA..T.TAA.GAACCAATA...TAA   46

CONSENSUS            ACAT..CTA.TGTAT.ATCGTACATTAAATTG.TTT.CC.CCATGC.ATA..T.TAA.GCAT..GTACATTAA

s.vulgaris           .GATGTCTA.TGTA.ATTCGTGCATTAA.T.GCATGTC..ACATT..A.A..TTAAATGGTACAGTACAT.AA   40
s.vulgaris           ACATATC.A.TGTTTAATCA.ACATTAAAA..CCTCCACCACATGC..T...TATAA.GCAT..GCACATTAA   44
```

Figure 2.4: The alignment of 27 mammalian ETAS1 sequences with two possible ETAS1 sequences from the red squirrel mitochondrial control region. The consensus sequence is shown and conserved nucleotides are shaded in grey. The number of nucleotides in common with the consensus sequence in given at the end of each sequence, the consensus sequence length is 59 bp and the average ETAS1 length is 59.9 bp.

```
                          10                  30                  50                  72
h.lar                CCATCTAAAGGGCATG.A.TGCACTC.ATTCATTCAC...CGCACATACAAACTCCCTACCACACTCAACTC   37
h.sapiens            CCCTTAACAGTACATA..GTACATAA.AGCCATTTAC...CGTACATAGCACAT...TAC...AGTCAAATC   44
p.troglodytes        CCCTTGACAGAACATA..GTACATACAACC.ATACAC...CGTACATAGCACAT...TAC...AGTCAAACC   41
p.paniscus           TCCTTAACAGTACATA..GCACATACAATT.ATATAC...CGTACATAGCACAT...TAC...AGTCAAATC   45
g.gorilla            TCACAAAAAGTACATA.AC.ACATAAGATC.ATTTA.T..CGCACATAGCACAT.CC..C...AGTTAAATC   42
p.pigmaeus           AGCTTTAAAGTACATA..GCACATAACACCCCT..AC...CGTACATAGCACAT...TTC..TACT.AACTC   38
m.musculus           TGGTTCA.GGT.CATA.A..A.ATA..ATC.ATCAA....CATAAAT.CAATATATATAC..CA.TGA.ATA   35
r.norvegicus         TGATTT..AGGACATACATTTAAACTCAACTATAAATT..C..ACA.ACAACATGTCTAT..T.CTCAAATA   34
g.glis               TAGTCAACCGTACATACA..ACAT.....T.A..ATCT..CA.ATACCC.ACATAAC.AG..TCCTCAACAG   29
o.cuniculus          TCCACT.TAATACAT.CAC.ACATA..ATCCAACAAAAAATTGACCCA.AACATGAATATTCTCACCAAAAA   30
c.commersonii        TCATTAATTTTACATA..TTACATGATATGTAT.AA.T..CTTACATATTATAT..ATCCCCTAA.CAATTT   40
k.breviceps          TTATTAATCTTACATA..TTACATAATATT.ATT.GGT..CGTACATAAGACAT....AC.CT..TTAAATC   45
b.mysticetus         TTATTTATTTTACATA.CGTACATAATAATCATTGA.T..CGTGCATGGTATATG..TCC.....TCAAATC   43
b.taurus             CTATAG.CAGTACATA.A.TACATATAATT.ATTGAC...TGTACATAGTACAT...TAT....GTCAAATT   42
o.aries              TGCTTGACCGTACATA..GTACATG..A...A.........GT.CAAA.TCCAT...T.C..TAGTCA.A.C   33
d.bicornis           CTGTTGATTTTACATA.A.TACATATTATT.ATTGA.T..CGTACATACGCCAT..C..C..AAGTCAAATC   45
e.asinus             TTATTTATCTTACATG.AGTACATCATATT.ATTGA.T..CGTACATACCCCAT..C..C..AAGTCAAATC   45
e.caballus           TCATTTATCTTACATA.AGTACATTATATT.ATTGA.T..CGTGCATACCCCAT..C..C..AAGTCAAATC   45
c.russula            TTATACATAATACATTAATTCCTTA..A.........T..CGGACATAGCACAT.....C..TAGTGAAATA   36
s.araneus            TTATA.ACAGTACATA.A.TACATTTTATT.ATTTA.T..CGTACATAGGACATA.C.AGTTATATCAATTC   44
e.europaeus          TTATTAATATTACATA..GTACATAATATT.ATTGA.T..CTTACATAGCGCAT.CCTAT..TAATAAACTT   44
h.grypus             TGGTTGATTTTACATGTACGGCATACAGTT.GT.AA....CG.CCAAACTTCAGA.GTAT..AACT.ACCTG   32
p.vitulina           TGGTTGATTTTACATATATGGCATATAATT.GT.AA....CA.CCA....AGTT..CTAA...AG.CA..TA   32
m.angustirostris     TGGTTGATTTTACATA.ATGGCATACGATT.GT.AA....CA.CCA....ATTTTTGGATAAAACT.ACCTA   31
f.catus              TTATATATATTACATA.A.GACATAATAGT.GCTTAAT..CGTGCATT.CACCTTAATTCTA.GGACAG.TC   38
d.virginiana         GCATA.ATCTGACATA.A.TACATAT.AT..ATTAAGA..CGTACATAT.ACATTCTTTC......CA..TG   39
o.anatinus           ACATT.ATATTAAAAATCCT.CATATCATG.ATT.A.T..CC.ACAT....CAAGTAAGCC....TCAACGC   35

CONSENSUS            T.ATT.ATA.TACATA.A.TACATA..ATT.ATT.A.T..CGTACATA..ACAT...TAC...A.TCAAATC

s.vulgaris           TCTTT.ATATTGCATA..GCACATAACATTCACTG.G..CGGCACATACCCCATT..TA....AGTCATAAA   36
```

Figure 2.5: The alignment of 27 mammalian ETAS2 sequences with the one found in the red squirrel mitochondrial control region. The conserved nucleotides are shaded in grey and the number of nucleotides in common with the consensus sequence is given at the end of each sequence. The consensus sequence length is 51bp and the average ETAS2 sequence length is 59.1 bp.

CSB1

```
                    1                   25
h.lar               AATTAATTCATGCTTGTTGGACATA
h.sapiens           AATTAATTAATGCTTGTAGGACATA
p.troglodytes       GATTAATTAATGCTTGCAGGACATA
p.paniscus          AATTAATTAATGCTTGTAGGACATA
g.gorilla           AATTAATTCATGCTTGTTGGACATA
p.pigmaeus          ATTTAATTTATGCTAG.AGGACATG
m.musculus          AATT.ATTCATGCTTGTTAGACATA
r.norvegicus        ATTTTATTCATGTTTGTAAGACATA
g.glis              AT.AAATCCATGCTTGTTGGACATA
o.cuniculus         ATTCTTTTAATGCTTGTCGGACATA
c.commersonii       ATTTAATTAATGGTTACAGGACATA
k.breviceps         ATT.AATTAATGGTCTCAGGACATA
k.breviceps         CTTAAATTAATGATCGTAAGACAT.
b.mysticetus        ATTAAATTAATGGTTACAGGACATA
b.taurus            TTACAGTCAATGGTCACAGGACATA
o.aries             ATATAATTAATGGTCACAGGACATA
d.bicornis          ATTCAGTCAATGGTTACAGGACATA
e.asinus            ATTCAGTCCATGGTAACAGGACATA
e.caballus          ATTCAGTCCATGGTAGCGGGACATA
c.russula           TATGGTTTAATGATCTGAAGACATA
c.russula           ATTCAGTCAATGGAATCAGGACATA
s.araneus           ATACGTTTAATGGTTACAGGACATA
s.araneus           AAATAATCAATGGTCACAGGACATA
e.europaeus         AATCTTTTAATGGTTACAGGACATA
e.europaeus         ATATGCTTAATGCTAGATGGACATA
h.grypus            TTTTAGTCAATGGTAGCGGGACATA
p.vitulina          TTTTAGTCAATGGTAGCGGGACATA
m.angustirostris    TTTTAATCCATGGTAGCGGGACATA
f.catus             ATTCAGTCAATGGTCACAGGACATA
d.virginiana        ATTTATTTAATATACGAAGGACAAT
d.virginiana        ...AAATCAATGATCTAAAGACATA
o.anatinus          TATAGATTAATGCTTGACGGACATA

CONSENSUS           ATT.AATTAATG.T.GCAGGACATA

s.vulgaris          ATAATATTCATGCTTGATGGACATA
```

CSB2

```
                    1                   18
h.lar               CAAACCCCCCCA.CACCC
h.sapiens           CAAACCCCCCCT.CCCCC
p.troglodytes       CAAACCCCCCCT.TCCCC
p.paniscus          CAAACCCCCCCT.CCCCC
g.gorilla           CCAACCCCCCCC.CCCGC
p.pigmaeus          CAAACCCTCGAC.AGCTC
m.musculus          CAAACCCCCC.A.CCCCC
r.norvegicus        CAAACCCCCCCA.CCCCC
g.glis              TAAACATTGTTT.ACTTT
o.cuniculus         CAAACCCCCCCT.ACCCC
c.commersonii       TAAACCCCCCTT.CCCCC
k.breviceps         CTTTTTTCCCCT.TACCC
k.breviceps         CACAAGCCCCAT.GACAC
b.mysticetus        CAAACCCCCC.T.CCTTC
b.taurus            TATTATATA..T.CCCCC
b.taurus            CAGGCCCCCC..CCCCCC
o.aries             TTTTCCCCCCTT.CCCCT
d.bicornis          CAAACCCCCCCA.CCCCC
e.asinus            CAAACCCCCCCC.ACCCC
e.caballus          CAAACCCCCCCT.ACCCC
c.russula           CAAACCCCC.TT.ACCCC
s.araneus           CAAACCCCCC.TACCCCC
e.europaeus         CAAACCCCCCTT.ACCCC
h.grypus            CAAACCCCCCTT.ACCCC
p.vitulina          CAAACCCCCCTT.ACCCC
m.angustirostris    CAAACCCCCCTT..CCCC
f.catus             CAAACCCCCCTT.ACCCC
d.virginiana        CTAAACCCCCTT.ACCCC
o.anatinus          TTCCCCCCCCTT.CCCCC

CONSENSUS           CAAACCCCCC.T..CCCC

s.vulgaris          CTTTACCCCTTAAACCCC
```

CSB3

```
                    1                   18
h.lar               TGCCAAACCCCAAAAACA
h.sapiens           TGCCAAACCCCAAAAACA
p.troglodytes       TGCCAAACCCCAAAAACA
p.paniscus          TGCCAAACCCCAAAAACA
g.gorilla           CACCAAACTCCAAATTTC
g.gorilla           AACCAAACCCCAAAGACA
p.pigmaeus          ATATCAACCC.AACAGAA
m.musculus          TGCCAAACCCAAAAAACA
r.norvegicus        TGCCAAACCCCAAAAACA
g.glis              CAAGAAA...TAAAAACT
o.cuniculus         TGCCAAACCCCAAAAACC
c.commersonii       ..................
k.breviceps         ..................
b.mysticetus        ..................
b.taurus            ..................
o.aries             ..................
d.bicornis          TGCCAAACCCCAAAAACA
e.asinus            TGCCAAACCCCGAAAACA
e.caballus          TGCCAAACCCCAAAAACA
c.russula           TGCCAAACCCCTAAAACA
s.araneus           TGCCAAACCCCTAAAACA
e.europaeus         TGACAAACCCCAAAACCA
h.grypus            TGCCAAACCCCAAAAACA
p.vitulina          TGCCAAACCCCAAAAACA
m.angustirostris    TGCCAAACCCCAAAAACA
f.catus             TGCCAAACCCCACAAACA
d.virginiana        CGTCAAACCCCAAAACCG
o.anatinus          ..................

CONSENSUS           TGCCAAACCCCAAAAACA

s.vulgaris          ..................
```

Figure 2.6: The alignment of the CSB sequences found in 27 mammals (from Sbisà 1998) with the proposed CSB1 and CSB2 sequences in the red squirrel.

2.3.3.3 The CSB domain

Both the CSB1 and CSB2 are present in the mitochondrial control region sequence of red squirrels, showing significant homology with the other mammal CSB sequences. The alignment of the CSB sequences produced by Sbisà *et al.* (1997) is shown in figure 2.6, aligned by hand to the proposed red squirrel CSB1 and CSB2 sequences.

CSB1 is the most preserved of the three blocks; in some species, namely the sperm whale, two species of shrew, hedgehog and opossum, it has been found to be duplicated (both sequences are included in figure 2.6). CSB2 and CSB3 are not so highly maintained. All of the species examined by Sbisà *et al.* (1997) carry some form of CSB2 but in fat dormouse, sperm whale, cow, sheep and platypus, it is only partially present and CSB3 is missing from the sequences of several species.

Two possible CSB2 sequences are given for cow and sperm whale and, whilst neither of the sperm whale sequences show a high degree of homology with the other CSB2 sequences, one of the cow sequences, although reduced to a long run of Cs, is 80% homologous to the consensus sequence making it a likely candidate for a CSB2. In red squirrel the proposed CSB2 sequence is only 60% homologous to the consensus sequence but it retains the structure of two separated blocks of C which may mean it retains its function. CSB3 is not apparent in the control region sequence of red squirrels. This is not unusual in mammals, it is also missing from the genomes of the dolphin, sperm whale, bowhead whale, cow, sheep and platypus, and is only partially present in the gorilla and orangutan. Whatever function CSB3 has, it is obviously not essential.

2.4 DISCUSSION

The control region of the mammalian mitochondrial genome has a three domain structure, defined by the adenine content of the three sections (Brown *et al.* 1986) and their propensity towards sequence change (Walberg and Clayton 1981). The red squirrel control region shows this three domain structure, defined by aligning the red squirrel sequence with the other mammalian sequences in which the domains have been defined. The central domain has the expected lower A and higher GC content in the L-strand and the much higher sequence homology with the mouse and rat sequences than the other two domains. The three domains in red squirrel are typical in length for mammals, all being close to the average for the mammals so far examined (table 2.2). Although, it is more common in mammals for the CSB domain to be longer than the TAS domain, this is not the case in red squirrels. Here, the TAS domain is the longest of the domains, but this deviation from the usual mammalian pattern is not great and this feature is unlikely to be of importance.

The central domain ranges from 300 – 328 bp in length in the mammals so far studied, the red squirrel central domain falls well within this range at 318 bp. In mammals, this domain seems to be structured with long stretches of conserved sequence spaced by short more variable motifs. The three sequence blocks, A, B and C, identified by Gemmell *et al.* (1996), are also obvious in the alignment carried out by Sbisà *et al.* (1997) but they seem to be somewhat arbitrary as there are other blocks of highly conserved sequence within the domain. Gemmell *et al.* (1996) suggested that these regions may represent a functional minimum, the most important sections of sequence in this domain for function to be maintained. It has been found that these sequences are capable of forming stable hairpin secondary stuctures which perhaps adds to the case for the sequences having some importance, but until the function of the central domain and reason for the evolutionary constraint it experiences is discovered then the importance of these sequence blocks will not be certain.

In the CSB domain, CSB1 is the least well conserved of the three conserved sequence blocks (figure 2.6). CSB2, except in the few species such as the cow, fat dormouse and the sperm whale, is very well conserved and CSB3 shows even greater sequence conservation. CSB1 may be the most variable of the three domains, but it is also the only domain found complete in all the mammals so far studied so may be the most important functionally. Only the end portion of CSB1, a motif of GGACATA, is found almost perfectly in all the species, perhaps this is the only region absolutely essential for function. The CSB1 sequence located

in the red squirrel sequence shows a reasonable homology to the other mammal sequences and includes the conserved end motif. The proposed CSB2 sequence in the red squirrel genome is not as well conserved as most of the other CSB2 sequences but it retains the basic structure of two separated blocks of cytosine which may be essential for function. This pattern is not seen in the proposed fat dormouse CSB2 sequence and in the cow, the CSB2 sequence is just one long undivided run of Cs. The CSB3 sequence in the dormouse is also only partial and is entirely missing from the cow so it is difficult to see how a function can be maintained, if one exists. The other species missing CSB3 have a CSB2 containing two separate blocks of C. The absence of CSB3 from the red squirrel and several other mammal species indicates that CSB3 itself cannot be essential for the functioning of the control region.

It is very unlikely that these conserved sequence blocks would occur in so many species just by chance. There must be a reason for their conservation and it seems reasonable to assume that the reason is one of function. Walberg and Clayton (1981) suggested that the locations of CSB2 and CSB3 just downstream from the 5' ends of the D-loop H-strands in mice may mean they have a role in the initiation of H-strand synthesis. Saccone *et al.* (1991) reported that a site-specific endoribonuclease "RNase MRP" specifically recognises CSB2 and CSB3, they act as signals for cleavage by the enzyme, so forming the RNA primer. In cows, RNase MRP cleaves the substrate at the short run of C residues that resemble CSB2 but in rat, cleavage occurs near CSB1 (Sbisà *et al.* 1997). This implicates these sequences in processes involved in mitochondrial DNA replication, but it does not seem that CSB2 and CSB3 are always essential.

CSB1 forms part of one of the cloverleaf structures found in the control region and is located near the point of transition from RNA primer to DNA synthesis in the replication of the control region, so early on it was suggested it could be a signal for this switch (Brown *et al.* 1986). This block has also been found to be a binding site for the mitochondrial transcription factor A (mtTFA) in humans. CSB2 is a binding site for the mitochondrial single-stranded DNA binding protein (mtSSB) (Gemmell *et al.* 1996) and three endonucleases that will cut either DNA or RNA have been identified that show some specificity for CSB2. As Sbisà *et al.* (1997) suggested, these sequences could be involved in several different processes involving RNase MRP, mtTFA and mtSSB but their actual functional roles remain elusive; all these functions would put limits on the evolution of the control region. Sbisà *et al.* (1997) thought that the spacing of the CSBs within the region could also be important for function, further limiting the region's potential for sequence change.

The TAS domain is also potentially very important as it is probably here that it is determined whether H-strand synthesis should be terminated, leaving the D-loop strand, or continued, initiating the replication cycle. It also seems reasonable to suggest that the conserved features of this domain are involved in this process in some way. When Doda *et al.* (1981) first described the TAS sequences, they suggested that they may be involved in signalling the termination of D-loop strand synthesis. But recently, the validity of the TAS sequences has been put in doubt because of their lack of conservation in a range of different mammalian species. Instead Sbisà *et al.* (1997) proposed that two ETAS sequences may be involved in these processes. The presence in this region of a cloverleaf-like secondary structure is likely to be significant as such structures have previously been found to be involved in the signalling processes surrounding replication.

The red squirrel TAS domain sequence contains two possible TAS sequences (both of which may be valid), an ETAS2 sequence and two possible ETAS1 sequences. The cloverleaf-like structure found in other mammals incorporates a portion of the ETAS1 sequence, so identifying this structure in the red squirrel would determine which of the two possible ETAS1 sequences is valid. The functions of these sequences, if indeed they have any, remain unknown. Suggestions have been made that ETAS1 could act as a recognition signal for the termination of the D-loop strand synthesis and ETAS2 may contain the binding site for the termination factor (Sbisà *et al.* 1997), but there is no real evidence to support these hypotheses. Alternatively, it may simply be that these sequences were once important in some way but that those functions have since been lost, leaving sequences that still show some residual conservation to be degraded in a random way by mutation.

The situation remains that there is a control region for the mitochondrial genome that presumably controls the replication and transcription of the genome yet the mechanisms of control are completely unknown. It seems probable that the conserved sequences and structures described in this chapter are somehow involved, as it is the most likely explanation for their existance and maintenance.

The control region of the mitochondrial genome is a very useful tool for population and evolutionary studies as it contains both highly variable and conserved portions of DNA. The variable regions are useful for population studies and the conserved central domain may be useful in higher level species comparisons and phylogenetic analyses. However, it is important to bear in mind the possible constraints on evolution the functional roles of the region may impose.

CHAPTER THREE:

PCR-SSCP ANALYSIS OF THE MITOCHONDRIAL CONTROL REGION

3.1 INTRODUCTION

Soon after the discovery of the mitochondrial genome in the early 1970s, evolutionary biologists began to investigate its usefulness as a potential molecular marker. For the first time, homologous sections of DNA could be isolated with relative ease from many different individuals. Restriction enzymes, isolated from bacteria, were used to cut the DNA wherever a specific sequence was found and could be used to roughly assay variation in DNA sequence. Where a change had occurred in the sequence identified by a restriction enzyme, that enzyme would no longer cleave the DNA; this led to a change in the patterns of DNA sections visualised on an agarose gel. By assaying with many different enzymes, a large amount of sequence could be covered and an estimate of genetic variation could be made. It was using this method that Avise *et al.* (1979) distinguished eastern and western populations of pocket gopher, *Geomys pinotis*, in the south-eastern United States. This grouping had been roughly identified previously using protein electrophoresis, but now local subpopulations could also be identified within the two populations. Mitochondrial DNA proved to have much greater resolving power than any previously used genetic marker and quickly became the marker of choice in population studies.

More recently, with the development of PCR techniques and the ease with which sections of DNA can be sequenced, restriction enzyme analysis of the whole mitochondrial genome has largely been superceded by methods with a greater resolution in assaying genetic variation, although it is still sometimes used as it is a quick and effective method. For example, Taylor *et al.* (1997) used restriction analysis of the whole mitochondrial genome to successfully investigate the genetic structure of koala (*Phascolarctos cinerus*) populations is southern Australia. They found some structuring not previously revealed by nuclear markers, presumably because of the predominantly male-mediated dispersal of koalas. When only males disperse, mitochondrial genes are not spread.

The mitochondrial genome can often prove to be more useful than the nuclear genome because it is uncomplicated by recombination and is usually maternally inherited, resulting in a greatly reduced effective population size. This makes it much more sensitive than the nuclear genome to genetic drift and more likely to show signs of population changes. It has a high mutation rate which means that recent population divergence is more likely to be detected with mitochondrial than with nuclear genes. As Moritz (1994) stated: "So long as variation exists, differences between populations will be more readily detected with mitochondrial DNA than with nuclear genes."

The control region of the mitochondrial genome is the fastest evolving part of the genome and has proved to be especially useful for population genetic investigations. Encalada *et al.* (1996), in their study of the population structure of the green turtle (*Chelonia mydas*), found that control region sequence data provided much higher resolution than had been achieved in previous studies using restriction enzymes and facilitated the accurate determination of phylogenetic relationships within the Atlantic region. This marker region has been used in a great number of studies to investigate levels of genetic variation, gene flow and population structuring in species ranging from the greenfinch (Merila *et al.* 1997) to the bank vole (Stacy *et al.* 1997; Aars *et al.* 1998) to man (Comas *et al.* 1996).

Gonzalez *et al.* (1998b) used this region to study populations of the endangered Pampas deer intending to reveal the effects of habitat fragmentation. There used to be millions of the species roaming the open habitats of South America, but today there are fewer than 80,000 individuals and their habitat has been greatly reduced by agriculture and urbanisation. Even though they now exist in low numbers in small and isolated populations, the variability of the mitochondrial control region was one of the highest of any mammal. Similarly high levels of variation had been found previously in African bovids (Arctander *et al.* 1996) so they concluded that, if the dynamics of the control region sequences are similar in these deer and the African bovids, then the high levels of variation reflect the large numbers that are thought to have existed previously. If the bottlenecks the populations experienced in the process of habitat fragmentation were not very extreme then much of the diversity could have survived the process. It is yet too early to clearly see the effects of drift on the smaller populations, although some differentiation between the Pampas deer populations is already apparent.

More commonly, in this age of habitat destruction, population genetic studies are revealing a lack of genetic diversity. This is particularly the case when the control region of the mitochondrial genome is used as a marker as it is particularly sensitive to the effects of population size reductions. Many studies have concluded that low levels of variation indicate that bottlenecks have occurred in the histories of the populations concerned, for example, the control region has been useful in detecting a bottleneck in sperm whales (Lyrholm *et al.* 1996). The diversity levels in this species were found to be extremely low, even when compared to other marine mammals which are thought to have a slower rate of evolution, and this did not appear to be due to an especially low mutation rate in this species. The diversity levels were found to be similar to those found in the northern elephant seal which was reduced to just a few individuals by hunting in the 19th century and so experienced a very dramatic bottleneck. It was concluded that the sperm whales must also have experienced a bottleneck, perhaps (as is concluded for so many species) during the late

Pleistocene glaciations. It has since been limited in the acquisition of diversity by long generation times, population structuring and changes in ocean temperature affecting survival and food abundance.

Recently, however, concern has been growing that the effects of bottlenecks alone are too easily blamed for reductions in variation levels that may in fact be due to other causes, or may just reflect a low level of variation that is natural in the species concerned. The role of population structuring in reducing effective population sizes, and therefore the levels of variation in a species, may by underestimated (Pimm *et al.* 1989; Gilpin 1991; Hedrick 1996).

Population studies require the examination of a large number of samples, often over several independent loci. This can be expensive and time consuming, so methods are preferred that are quick and inexpensive but still retain the necessary sensitivity for meaningful results; the identification of "single stranded conformational polymorphisms" (SSCPs) in PCR products is such a method. Known as "PCR-SSCP", it is used to survey short sections of DNA for the presence of mutations and so can be used to quantify the number of alleles present in a population sample. It retains almost all the resolving power of sequencing without having to actually sequence all the samples which would be both time consuming and expensive. It is often useful for the exact sequences of the different alleles to be determined by sequencing after PCR-SSCP has determined the frequencies of the different alleles. PCR-SSCP can then be used to produce data on the allele frequencies in the sample and so on the variation present, the basic data of classical population genetics. It also identifies the samples worth sequencing; knowledge of the sequences of the different alleles can then be used for phylogenetic investigations at higher taxonomic levels (Lessa and Applebaum 1993).

3.1.1 Single-Strand Conformational Polymorphism (SSCP)

Single-strand conformational polymorphism (SSCP) refers to differences in structures formed by single-stranded DNA under the appropriate conditions (Orita *et al.* 1989). In theory, it can be used to distinguish between DNA strands that differ in sequence by only one or two base pairs. It is based on the principles that the molecular conformation of single-stranded DNA is dependent on its nucleotide sequence and that even minor conformational changes in the single-stranded DNA can change the mobility of the molecule through a polyacrylamide gel (Lessa and Applebaum 1993). DNA strands with only minor differences in sequence will have different single-strand conformations and so will migrate through the gel at different rates. When combined with PCR, the method can be applied to many samples and the resulting patterns of DNA easily visualised on a polyaccryamide gel. PCR-SSCP is a cheap and effective technique used to survey short sections of DNA for sequence variation.

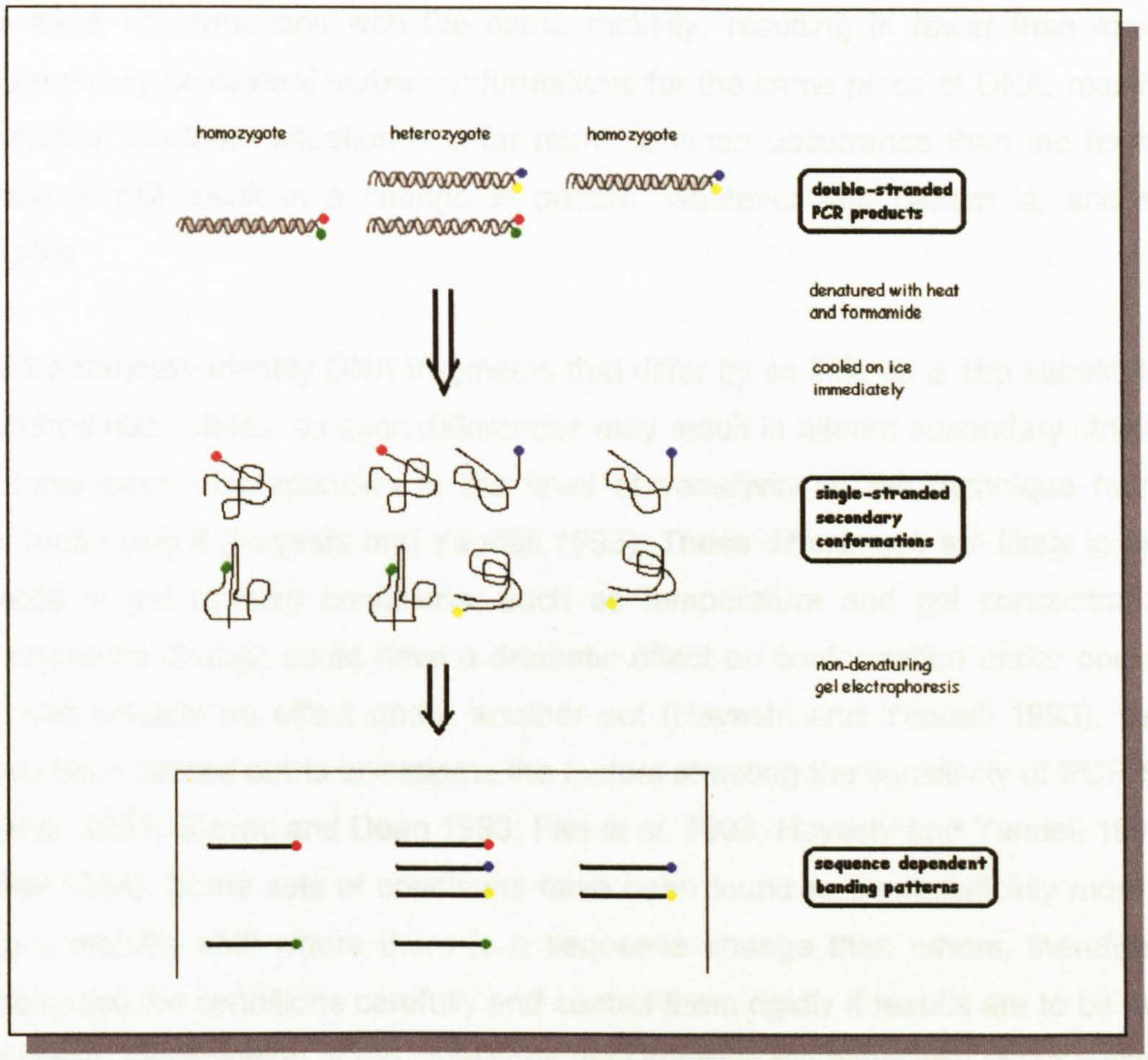


Figure 3.1: The principles of SSCP analysis used to detect differences in sequence between individuals.

The technique was first described in 1989 by Orita *et al.* and has since been used widely for detecting mutations associated with cancer and genetic diseases as well as in population studies. Double-stranded PCR products are denatured using heat and formamide, then quickly cooled on ice so the DNA strands do not reanneal to each other but form individual conformation products or secondary structures, stabilized by intramolecular interactions (figure 3.1). As these structures are dependent on many variables such as temperature and ionic concentration they are difficult to predict theoretically. Indeed there may be several stable conformations for one individual DNA strand (Hayashi and Yandell 1993).

The DNA strands in their secondary conformations are then run through a non-denaturing polyacrylamide gel. PCR products of differing sequence are visualised on the gel as bands in distinctive positions due to variation in the mobility of their various secondary structures. The banding patterns are referred to as haplotypes. It is expected that the two denatured DNA strands will migrate at different speeds (due to differences in their base compositions and secondary structures) and separate on the gel; a homozygote would display two bands and a heterozygote four bands, but this is not necessarily the case. Two different DNA strands may

by chance have conformations with the same mobility, resulting in fewer than expected bands, or there may be several stable conformations for the same piece of DNA, resulting in additional bands; the latter situation is a far more common occurrence than the former. A mutation should still result in a change in pattern, whatever that pattern is, and so be distinguishable.

SSCP can theoretically identify DNA fragments that differ by as little as a 1bp substitution in several hundred nucleotides, as such differences may result in altered secondary structures, but there have been discrepancies in the level of sensitivity of the technique found by different groups using it (Hayashi and Yandell 1993). These differences are likely to be due to differences in gel running conditions, such as temperature and gel concentration. A particular sequence change could have a dramatic effect on conformation under one set of conditions and virtually no effect under another set (Hayashi and Yandell 1993). Several studies have been carried out to investigate the factors affecting the sensitivity of PCR-SSCP (Spinardi *et al.* 1991; Glavac and Dean 1993; Fan *et al.* 1993; Hayashi and Yandell 1993; Liu and Sommer 1994). Some sets of conditions have been found to be statistically more likely to result in a mobility shift where there is a sequence change than others, therefore it is important to chose the conditions carefully and control them rigidly if results are to be reliable and reproducible. Optimisation of the conditions may improve the sensitivity of the technique.

Changes in the running conditions result in different conformations. As Dean and Milligan (1998) pointed out, the number of possible conformations for each strand of DNA is so large that there is no theoretical basis for choice of conditions; the aim is to find the conditions which maximise the chances of detecting a sequence difference. Environmental factors affecting the sensitivity of the technique by affecting the conformations of the molecules include temperature, ionic strength of the buffer, the gel matrix concentration and the inclusion of additives such as glycerol and sucrose. By varying these conditions between gels it should be possible to determine the conditions under which a gel will distinguish different sequences more clearly than others, reducing the probability of a false negative.

It has been shown by most groups that cooler conditions generally improved the sensitivity of the technique (Glavac and Dean 1993; Liu and Sommer 1994). Orita *et al.* (1989) suggested that some of the semi-stable conformations may be destroyed at higher temperatures making them less discriminating. A cold room provides a constant cool environment in which to run gels, but whatever temperature the gels are run at, it should be constant to ensure consistent results. The ionic strength of the buffer must also be consistent for comparisons to be possible between gels. Poly-acrylamide gels are traditionally run in TBE at a 1x

concentration but Spinardi *et al.* (1991) found that using a 0.5x TBE buffer lead to clearer, sharper bands on the gels. Originally ordinary polyacrylamide gels were used to separate the DNA strands, but there are now other polyacrylamide based matrices available developed specifically for PCR-SSCP. For example the MDE™ (Mutation Detection Enhancement) gel solution produced by Flowgen has been shown to be superior to polyacrylamide in detecting mutations, and to have a more consistent efficiency when used in varied conditions (Liu and Sommer 1994).

The length of the section of DNA being examined seems to have an effect on the efficiency of mutation detection. PCR-SSCP is most efficient for detecting changes in short lengths of DNA, only a few hundred base pairs in length. Dean and Milligan (1998) suggested that the optimum length for analysis is only 200-300 bp but Liu and Sommer 1994 found that MDE gels were more efficient than polyacrylamide so longer fragments could be used, although they started to find some decrease in efficiency when products between 241 and 295 bp were examined. Many studies have found a negative correlation exists between sequence length and detection efficiency (Hayashi and Yandell 1993) but Fan *et al.* (1993) felt that this relationship was an oversimplification. Some mutations produce a larger mobility shift when contained within a larger fragment making the detection of the mutation more likely. He felt that careful selection of the operating conditions was a much more important consideration than the length of section to be examined. It is, of course, possible to examine longer sections of DNA with SSCP by digesting them with a restriction enzyme before carrying out the electrophoresis.

Generally, it seems that shorter fragments are preferable as SSCP appears to be more reliable in mutation detection when used on shorter lengths of DNA. This tendency must be balanced with the need for clearly visible bands on the gel which can be easily distinguished from one another. The bands of larger fragments are generally easier to resolve as they are more easily and quickly separated when electrophoresed through a matrix such as polyacrylamide. In a population study such as this, the extent of variation present in the sample must also be taken into account. A small fragment may not have enough potential variability to be detected and quantified in a meaningful way, larger fragments may be needed to display a useful level of variation.

3.1.2 Statistical analysis methods

Statistical tests are used to assess whether the data resulting from an experiment support a null hypothesis. The null hypothesis (H_o) is the hypothesis that the data are being used to test, for example, when comparing the genetic profile of several populations, the data can be used to test the hypothesis that there is no genetic differentiation between the populations. A statistical test is used to assess the probability of getting the results observed if the null hypothesis is correct, or more strictly, they calculate the probability of falsely rejecting a null hypothesis, on the basis of the data, that is actually correct. For example, if the probability value assigned to the data is 60% then that data could lead to the false rejection of the hypothesis more than half of the time, so it is most likely that the null hypothesis is correct and it is accepted. Whereas, if the probability is only 1% then the null hypothesis would be rejected incorrectly on only 1 in 100 occasions, so it is concluded that the null hypothesis is not supported by the data and it is rejected.

The level at which the null hypothesis is rejected is the significance level, it is the level at which the probability of falsely rejecting a null hypothesis is low. This level is generally accepted as 5% or during 1 in 20 tests, but there is no reason for this particular level, it is just the consensus of opinion in the scientific research community. Therefore, if the probability is more than 5%, the null hypothesis is accepted, but if the probability is less than 5%, the null hypothesis is rejected. If a correct hypothesis is falsely rejected it is called a type 1 error, so the probability calculated in the analysis can also be called the type 1 error probability.

This probability is assigned by a variety of methods: some calculate a statistic (value) describing the data and then relate it to a probability distribution appropriate for that statistic. Other methods directly calculate the probability of that data set occurring given the null hypothesis. Originally, probability distributions were calculated laboriously by the statistician who developed it by repetitive testing and calculating the actual probability of each outcome. These methods rely on the data being appropriate to compare to the probability distribution and tests may first be carried out to assess whether this is the case. For example, for parametric analysis methods such as t-tests and Anova, the data must be normally distributed, this can be tested by a Kolmogorov-Smirnov normality test. Probability distributions have many constraints on their usage and if these requirements are not met by the data then the distribution is not applicable. For example, the χ^2 test is not reliable when expected numbers are less than five as the χ^2 value tends to infinity, so another test, an exact test, may be used instead (Bailey 1981). It is for this reason that exact tests have been used in this study.

3.1.2.1 Pairwise Exact Tests

The exact test is a statistical method employing contingency tables and is used as an alternative to the χ^2 distribution; it calculates the actual probability of the set of results occurring. All the possible contingency tables that still result in the same category totals are worked out and their conditional probabilities are calculated using the following formula (Bailey 1981):

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{N!\,a!\,b!\,c!\,d!}$$

for the table:

a	b	a+b
c	d	c+d
a+c	b+d	N

The type 1 error probability of the contingency table under investigation is calculated by summing the conditional probabilities of all the tables with a probability less than or equal to that of the table under consideration. The hypothesis can then be accepted or rejected on the basis of this probability.

Unfortunately, the exact test is impossible to calculate manually for most population data sets as the number of possible contingency tables for a set of results is often too large. For example, the number of tables possible for two populations with four equifrequent alleles and 160 individuals is 10^7. An alternative approach involves estimating the *p*-value by considering a random sub-set of the possible tables. This can be done using the Markov Chain Monte Carlo algorithm (Rousset and Raymond 1997). The principle of this method is to explore the "space" of all the possible contingency tables; the program moves from table to table but whether it moves between the tables is determined by certain conditions. As it moves, the time spent on each table encountered with a probability less than or equal to that of the table under consideration is recorded and the proportion of time spent on these tables represents an unbiased estimate of the type 1 error probability (Raymond and Rousset 1995b). Details of the algorithm used in the Markov Chain are given in Raymond and Rousset (1995b).

Exact tests were carried out to compare each population with each of the other populations in a pairwise manner and determine whether there was any significant difference between them. Pairwise exact tests (PETs) were carried out using the Markov Chain Monte Carlo algorithm using the computer program GENEPOP (v. 3.1) (Raymond and Rousset 1995a).

3.1.2.2 The Bonferroni correction using the Dunn-Šidák method

The comparisons made by the PET analysis are not independent as they involve repeated use of each set of data in different comparisons. This means that if the outcome of one comparison is significant then the outcomes of further comparisons are more likely to be significant and hence increase the probability of incurring a type 1 error. This type of error arises when a true hypothesis is falsely rejected, so when the null hypothesis is that samples have been drawn from the same population a type 1 error occurs when samples from that same population are deemed to be significantly different to the others and the null hypothesis is rejected. By the intrinsic nature of significance testing there is always a risk of a type 1 error, if the significance level is taken to be 5% then samples from within the same population will be falsely found to be significantly different 5% of the time (Sokal and Rohlf 1981).

When the comparisons are not independent, a correction can be performed using the Dunn-Šidák method to allow for the resulting increase in experiment-wide error rate and keep the error rate at an acceptable level (Sokal and Rohlf 1981). This involves recalculating the significance level for each comparison according to the formula:

$1-(1-\alpha)^{1/k}$ where α is the original significance level (usually 0.05). and k is the number of comparisons.

In this study there are 66 comparisons to be made. Each comparison is made and the resulting probabilities are ordered from the lowest to the highest. The lowest probability is compared to the significance level as calculated by the formula $1-(1-0.05)^{1/66}$ and the conclusion as to whether to reject or accept the null hypothesis is made. The next lowest probability is compared to the significance level calculated as $1-(1-0.05)^{1/65}$ as there are now only 65 comparisons, and so on for each of the comparisons.

3.1.2.3 Genetic diversity measures

The simplest measure of genetic variation in a population is the **allelic diversity** which is a straight count of the number of alleles present in the population. This can only be measured accurately when all the individuals in the population have been sampled and this is rarely possible. However, the number of alleles in a sample from that population can give an idea of the level of variation present but as allelic diversity is also dependant on sample size a comparison cannot be made between samples of differing size.

The **nucleotide diversity** (Nei 1987; Hoelzel and Bancroft 1992) is a measure that incorporates the number of differences between two sequences, the sequence length compared and the number of comparisons made when every pairwise comparison is made between samples; it overcomes the problem of sample size differences between the populations being compared. It is calculated as:

$$\sum\left(\frac{\text{no. of differences between two sequences}}{\text{sequence length under comparison}}\right)\Big/\text{no. of pairwise comparisons}$$

The number of pairwise comparisons is given by $n(n-1)/2$ where n is the number of samples to be compared.

The number of differences between two sequences divided by the sequence length is the sequence divergence or genetic distance (Hoelzel and Bancroft 1992; Arctander *et al.* 1996). It is usually quoted as a percentage and given as a range from the smallest to the largest within each group considered. It facilitates comparison of the range of variability found within populations. The mean number of pairwise differences is the average number of differences found between the sequences when all the pairwise comparisons are carried out. It can also be used to calculate the nucleotide diversity by dividing it by the sequence length.

A variable site is a position in the sequence found to vary in nucleotide content within the samples studied. A more variable population can be expected to have a higher **number of variable sites** than a less variable one. This is highly dependent on sequence length so the sequences under comparison should be the same length.

3.2 METHODS

3.2.1 Tissue sampling

Tissue samples were collected from the German red squirrels by Sibylle Münch and from the Belgian populations of Peerdsbos and Merodese Bossen by Luc Wauters in the early 1990s. Samples from the Belgian fragment populations were collected by Goedele Verbeyen. The squirrels were caught in traps (figure 3.2), then weighed, measured and electronically tagged. Individuals were periodically radio labelled so their movements could be tracked and their home ranges plotted. A small portion of tissue was removed from the ears of the squirrels (figure 3.3) and stored at –20°C in a saturated salt solution containing 10% Dimethyl sulfoxide (DMSO). As the fragment populations were so small and being monitored continuously, it was possible to be reasonably certain that every individual present in these areas was trapped and sampled in this way.

3.2.2 DNA extraction

DNA had been extracted from the Peerdsbos and Merodese Bossen samples by Iwona Hutchinson and Heike Knothe during work previously carried out in the same laboratory. These samples were further cleaned by reprecipitation with isopropanol, as outlined below, to remove excess salt that may have inhibited the PCR reactions.

DNA from the other samples was extracted using the phenol/chloroform method given in section 2.2.2. The precipitation of the DNA, step 4, was carried out using IPA by adding $1/7^{th}$ the volume of IPA and centrifuging the samples in a microfuge at 13000rpm for 30 minutes. This was to avoid the further addition of salt to samples that already had a high salt content from the storage solution. A 100% ethanol rinse in addition to the 75% rinse was also carried out to further ensure that any remaining salt was removed before the pellets were oven dried. The pellets were resuspended in 50µl of TE. 2µl of each extraction was run through a 0.8% agarose gel and visualised by EtBr staining (as described in section 2.2.3) to check the extraction was successful and acceptably clean. They were then diluted 1µl in 10µl of TE ready for PCR amplification.

photograph: Rebecca Todd

Figure 3.2: A red squirrel in a trap

photograph: Rebecca Todd

Figure 3.3: A small piece of tissue is removed from the ear of each squirrel

3.2.3 PCR amplification of the control region

A section of the red squirrel mitochondrial control region 367bp in length was amplified using the primers RSCR1 (5'-CACATAGTACATAGACATTAGG-3')

and RSCR4 (5'-GGAACTAATCCATCGTGATG-3')

RSCR1 had been designed for the sequencing of this region (section 2.2.7) and RSCR4 was designed to bind to the sequence an appropriate distance from RSCR1. The amplifed section of the control region included all the variable sites identified in the comparison of Belgian and English sequences (section 2.3.1).

The PCR reaction was optimised by altering the annealing temperature and the final concentrations of Magnesium chloride, nucleotides and primers. The reaction was carried out with the following reaction mix:

Reagent	Conc^{n}.	Quantity	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	2.5 µl	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	1 µl	0.05 mM
Magnesium chloride (Advanced Biotechnologies)	25mM	1 µl	1 mM
primer RSCR1	10µM	3 µl	1.2 µM
primer RCSR4	10µM	3 µl	1.2 µM
Red Hot *Taq* (Advance Biotechnologies)	5 U/ µl	0.2 µl	1 U
template	5-50 ng/µl	2 µl	0.4-4 ng/µl
sterile distilled water		12.3 µl	

The reactions were run on a programmable thermocycler (PTC100 or PTC200, MJ Research, Inc.) according to the following program:

Step	Temperature (C)	Time (minutes)
1	94	2
2	94	1
3	47	1
4	72	1 1/2
5	Go to step 2	29 more times
6	72	5

The PCR products were visualised in agarose gels as described in section 2.2.3, to ensure that each reaction had worked. Only clean products of the correct size were used in the SSCP analysis. 5μl of each product was loaded into the wells of a 1.5% agarose gel in 1xTAE buffer containing 0.5μg/ml EtBr to be electrophoresed at 100 V for approximately 20 minutes and visualised under UV light.

3.2.4 Single-Strand Conformational Polymorphism gels

The16cm Protean II xi cell system (Biorad) was used to run MDE gels according to the manufacturers instructions. Gels were run under constant conditions in a cold room maintained at 4°C in order to give consistent results.

1. 0.5 x non-denaturing MDE gel solutions were made to the following recipe (enough for one gel):

2x MDE	(Flowgen)	12.5ml
1x TBE	(0.09mM Tris-borate, ImM Na_2EDTA, pH 8.0)	30ml
distilled water		7.5ml
temed	(Sigma)	100μl
10% ammonium persulfate	(GibcoBRL)	200μl

2. Gels were poured by syringing the solution between the two glass plates of the gel rig, separated by 0.4mm spacers and a lane forming comb was pushed into the top of each gel.

3. Once set, the combs were removed and the lanes were rinsed with distilled water. The gels were set-up in the rig with 0.6x TBE buffer and equilibriated by running a current through them for about an hour in the cold room before loading.

4. Meanwhile, the DNA from the PCR reactions was precipitated by the addition of 0.6 volumes of isopropanol, mixed by vortexing and centrifuged in a microfuge at 13,000 rpm for 10 minutes. The supernatant was removed and the pellets of DNA were dried in an oven. The precipitated DNA was resuspended in 12μl of the loading buffer (250μl formamide, 10μl 0.5M Na_2EDTA pH 8.0, 12.5μl 1% bromophenol blue).

5. The DNA was denatured by heating to 95°C in a PTC-200 Thermal Cycler for 3 minutes and then transferred immediately to ice.

6. 6μl of each denatured sample was loaded into the wells of the gel. The gels were run for 16 hours at 200V, in the cold room.

3.2.5 Silver Staining of SSCP gels

Silver staining is a more sensitive visualisation technique than staining with Ethidium bromide (Lessa and Applebaum 1993) and is less hazardous than either labelling with radioisotopes and Ethidium bromide staining. The following protocol, adapted from standard protocols by Neumann (1996), was used for all gels; 100ml of each solution was used for each gel:

1. Fixing: the glass plates were prised apart and the gel peeled off carefully into a tray containing a fixing solution (50% methanol, 10% acetic acid). The gel was soaked in this solution for 10 minutes with constant agitation.

2. Rinsing: twice in a wash solution (10% ethanol, 0.5% acetic acid) for 3 minutes.

3. Staining: with silver by soaking it in a 0.1% silver nitrate solution for 10 minutes with constant agitation.

4. Developing: the gel was rinsed briefly in developing solution (1.5% NaOH, 0.02% $NaBH_4$, 0.12% formaldehyde) to precipitate the excess silver, and then soaked in fresh developer until the bands were visible.

5. Finally, the gel was rinsed in distilled water and soaked in a 0.75% Na_2CO_3 solution for 20 minutes. The stained gel was then transferred to a piece of Whatman 3mm chromatography paper (Whatman International) and dried using a gel drier.

3.2.6 Sequencing

Selected PCR products were also sequenced to determine the exact nature of any sequence variation. Samples to be sequenced were again amplified by PCR using primers RSCR1 and RSCR4 and the full 25µl reactions were run through a 1.5% agarose gel (section 2.2.3). The bands of amplified DNA were cut out of the gel using a scalpel blade and extracted from the agarose matrix using the Qiagen gel extraction kit and protocols.

These products were sequenced using the ABI d'Rhodamine Terminator Cycle Sequencing system (Applied Biosystems Inc.), which is an automated system using a fluorescent label rather than radioactivity. The use of fluorescence-based technologies for automated DNA detection is reviewed by David and Menotti-Raymond (1998). The reactions were carried out using the supplied kit and protocol, as follows. The sequencing reactions were carried out to incorporate flourescently labelled ddNTPs and then taken to the Division of Immunology, University of Nottingham, were the reactions were run through an automated sequencer (ABI Prism 377 Automated Sequencer, Applied Biosystems Inc.). The labelled DNA is detected by a laser and recorded by a computer. The resulting sequence is recorded as a chromatograph.

1. Pipette the following components into 0.2ml tubes and mix well:

Reaction mix (Applied Biosystems Inc)	6µl
Primer	1µl
Template DNA	5µl
SDW	8µl

2. Cover the mixtures with a drop of mineral oil and run the reactions on a programmable thermal cycler (PTC-200, MJ Research, Inc.) with the following program:

Step	Temperature	Time
1	93°C	3 mins
2	94°C	30 secs
3	55°C	5 secs
4	60°C	4 mins
5	Go to step 2	23 more times
6	4°C	1 min

Transfer the reaction to a 0.5ml eppendorf precipitate the DNA by the addition of 2µl 3M Sodium acetate and 50µl cold 100% ethanol. Vortex the tubes briefly and incubate the tubes on ice for 10 minutes. Spin the tubes in a microfuge at 13000rpm for 30 minutes. Carefully remove the supernatant, rinse the pellets (too small to be visible) with 70% ethanol and spin again briefly. Carefully remove all the supernatant and oven dry the pellets thoroughly.

These reactions were then taken to the automated sequencer through which each resuspended reaction was run by a technician. The computer produced a chromatograph which could be printed out to create a permanent record. The computer files could be accessed and edited by eye, the computer translations of each chromatograph to DNA sequence were checked by eye before being analysed.

3.2.7 Data analysis

The sequences were examined for differences by alignment using the program CLUSTAL-W (Thompson *et al.* 1994). Pairwise Exact Tests (PETs) were carried out on the Belgian populations using GENEPOP (v. 3.1) (Raymond and Rousset 1995a) and a Bonferroni correction was made using the Dunn-Šidák method (section 3.1.2.2). The nucleotide diversity and the mean number of pairwise differences were calculated using ARLEQUIN (v 1.1) (Schneider *et al.* 1997). The Kolmogorov-Smirnov normality test, the t-test and the χ^2 calculation were carried out using MINITAB (release 10.5, Minitab Inc.).

3.3 RESULTS

129 red squirrel samples from Belgium, comprising 39 samples from 2 large populations and 90 samples from 8 small fragment populations, were amplified and analysed using PCR-SSCP. A further 25 samples from the large German population of Waldhäuser were also analysed. Haplotypes were assigned to the samples based on the banding pattern seen in the SSCP analysis. Sequencing was undertaken to determine the differences between the alleles and to confirm that the haplotype assignment was an accurate representation of the allelic diversity found at the locus. 314 bp of clean sequence was obtained in all the samples sequenced (from base 230 to 544, counting from the 5' end of the H-strand sequence (figure 2.2)), so this sequence fragment was taken to be the locus length in the following analyses.

The banding patterns seen in the SSCP analysis contained more than the two bands that may have been expected from a product of the haploid mitochondrial genome. Most of the patterns contained 3 – 6 bands, some of which were stronger than others (figure 3.4). Multiple bands arise when there is more than one stable conformation for the DNA strands (Hayashi and Yandell 1993).

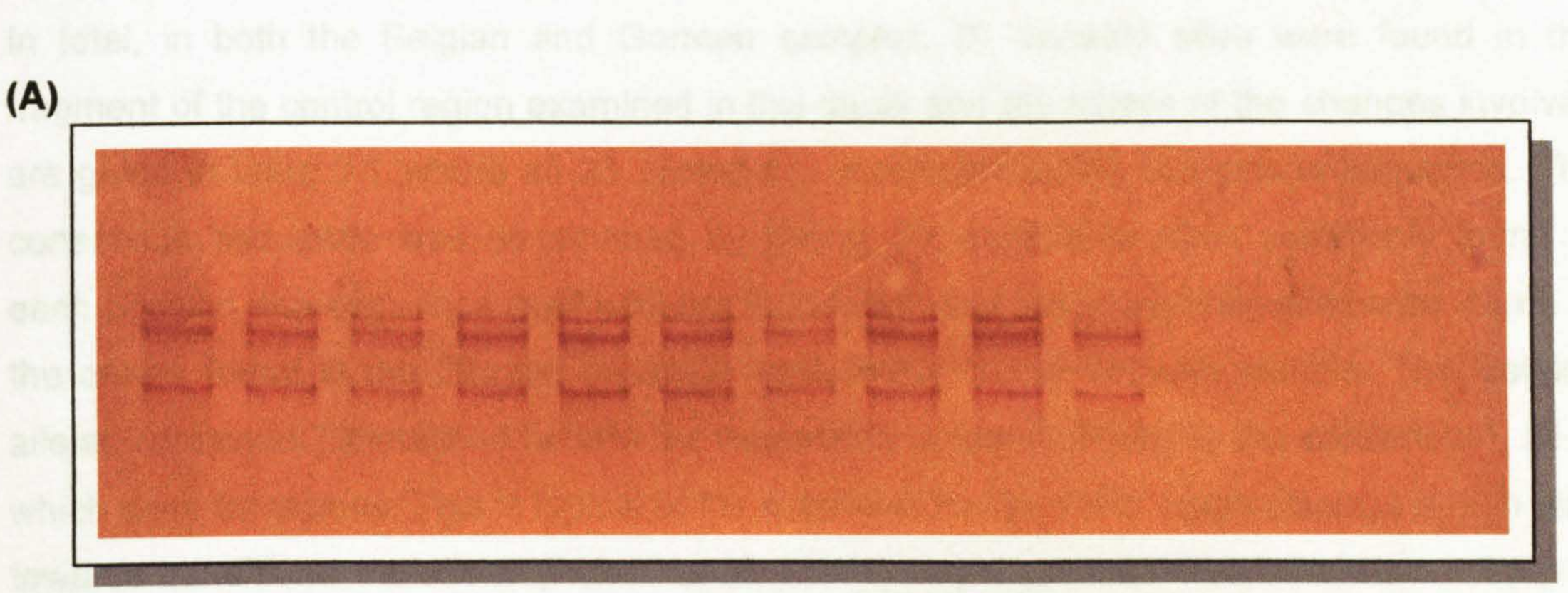


Figure 3.4: SSCP gels showing the haplotypes of individuals from **(A)** the Belgian population of Tallaarthof, and **(B)** the German population of Waldhäuser.

3.3.1 The sequence variation detected using PCR-SSCP

In all the 129 Belgian samples, only 3 alleles were detected; all but four individuals shared the same allele. This was confirmed by sequencing 33 (25%) of the samples, which did not result in the detection of any further alleles. By contrast, the German sample of 25 individuals contained 18 different haplotypes. Sequencing these samples revealed a further two alleles not identified by PCR-SSCP making a total of 20 alleles in the sample of 25 individuals.

As SSCP can vary dramatically in reliability, it was important to be sure that the technique, with the conditions used here, was accurately detecting the variation present. The German sample was very variable, so by sequencing all these individuals and comparing the level of variation detected by PCR-SSCP with that detected by sequencing it was possible to assess the reliability of PCR-SSCP as used in this study. Sequencing the German samples revealed 20 different alleles, compared with 18 haplotypes distinguished by the SSCP analysis, so PCR-SSCP revealed 90% of the variation present. Of the two alleles missed by the SSCP analysis, one differed from the others of the same SSCP haplotype by only 1 bp and the other differed by three single base changes. By chance, all three of the changes did not affect the conformational structure of the DNA strand under the conditions used in this study.

In total, in both the Belgian and German samples, 37 variable sites were found in the fragment of the control region examined in this study and the details of the changes involved are given in table 3.1 where all 23 alleles are compared to the consensus sequence. The consensus sequence was constructed by taking the nucleotide most commonly found in each position; the sequence itself was not found in any of the individuals examined. None of the alleles found in the Belgian squirrels were found in the German sample. The Belgian alleles contained 7 mutations (where the nucleotide content differed to the consensus), all of which were transitions. This is typical of the mitochondrial genome which displays a high bias towards transitional mutations (Wilson *et al.* 1985). Only 2 of these mutations were unique to the Belgian populations, the other changes were also found in the German sample. At the 35 German variable sites 37 changes were observed. Two different mutations were found at each of two sites: 255 and 324; at these sites both a transition and a transversion had occurred. Of the 37 changes, 30 were transitions, 6 were transversions and one insertion of a single base pair had occurred at base 449. Again, these figures reflect the high transition to transversion ratio found in the mitochondrial genome when comparing closely related taxa. Of the 32 transitions found in total in both populations, only six did not involve thymine and cytosine (pyrimidines). This reflects the bias for pyrimidines, specifically for thymine, found in the H-strand sequence (see table 2.4).

position	234	235	237	239	240	245	249	254	255	256	257	258	259	261	271	273	275	279	288	289	292	293	294	298	304	311	319	324	348	352	366	388	395	427	449	497	498
consensus	T	T	A	C	C	C	T	T	C	C	T	A	C	A	T	T	T	C	T	C	T	A	T	G	C	T	A	T	C	C	C	T	C	G	-	T	T
0											C																										
1								C																A													
2			G								C		T																	T				A			
3				T							C																										
4								C					T											A			G										
5								C																										A			
6					T			C																			G	C			A		A				
7						T		C																A	T		G										
8								C										T		T						C	G		T								
9								C			C						C																				
10								C																										A		G	
11									T					G																			C		A			
12											C																							A				
13								C										T		T						C	G		T					A				
14				T							C																	G			A		A				
15							C	C																A													
16	C				T			C			C							T	C					A			G										
17								C	A	T												G										C		A			C
18															C																						
19											C					C					A	G	C											A			
A																										C	G	C								C	
B																		T									G	C									
C		C										G																									

Table 3.1: A table showing the sequence differences between the 23 different alleles found in the red squirrel control region. The consensus sequence is given at the top and the differences between the alleles and the consensus sequence are shown below. The 20 alleles found in the German population are labelled 0 to 19 and the Belgian alleles are labelled A, B and C. The position refers to the base number in from the 5' end of the amplified product sense strand.

A total of 91 mutations were detected in the 154 squirrel samples examined (table 3.1); their distribution along the length of the PCR product is illustrated in figure 3.5. The PCR product spanned the left domain and the central conserved region of the control region, half of it (bases 1 – 156 of the product) in the left domain and the rest in the central domain. There are more variable sites in the left domain than in the central region, with 31 (20% of the 156 nucleotide positions) in the left domain and 6 (4%) in the central region. The left domain also contains most of the mutations, 84% of the total. Clearly the left domain is far more variable than the central domain, as would be expected.

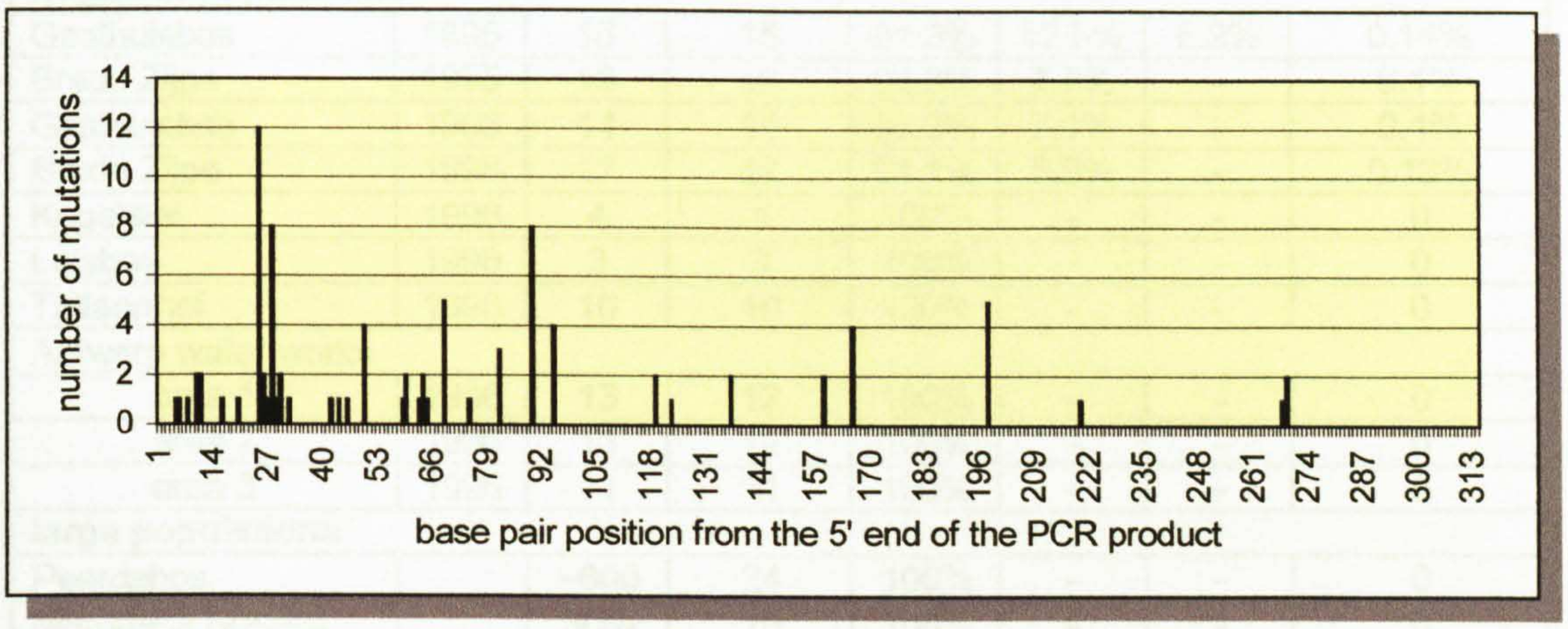


Figure 3.5: The distribution of mutations along the length of the control region segment. The graph shows the number of mutations found at each variable site.

3.3.2 The Belgian populations

If it is accepted that the SSCP analysis is likely to have distinguished all the alleles in the Belgian populations, then only three alleles were present in all ten of the populations. These alleles are named A, B and C and their frequencies are given in table 3.2. Most individuals (97%) carried allele A with alleles B and C found in just 2% and 1% of the samples respectively. The samples from both of the large populations contained only allele A. As these were just samples from larger populations, the results can only be taken as an indication of the level of allelic diversity present in the populations.

Most of the fragment populations (Tallaarthof, Kegelslei, Luisbos and the three areas around the Antwerp water works) were found to be fixed for allele A. Only Brede Zijpe and Gasthuisbos showed any variation. These two populations had been sampled over a two year period and the results for each year are shown in table 3.2. In 1995 Gasthuisbos contained 3 alleles at this locus, with one individual with allele C and two with allele B, but in

1996 only one of the allele B individuals remained. Brede Zijpe had one individual with allele B in both years. However, even this small amount of variation can be expected to be lost as both of the squirrels carrying allele B that remained in 1996 were male. The mitochondrial genome is maternally inherited, so these lineages will die with these two individuals. As these small populations were completely sampled, these results are actual measures of the allelic diversity of the populations.

population	year	size	sample size	haplotype frequencies			nucleotide diversity
				A	B	C	
fragment populations							
Gasthuisbos	1995	16	15	81.3%	12.5%	6.2%	0.14%
Brede Zijpe	1995	13	13	92.3%	7.7%	-	0.1%
Gasthuisbos	1996	14	14	92.9%	7.1%	-	0.4%
Brede Zijpe	1996	17	17	94.1%	5.9%	-	0.13%
Kegelslei	1996	4	4	100%	-	-	0
Luisbos	1996	3	3	100%	-	-	0
Tallaarthof	1996	10	10	100%	-	-	0
Antwerp water works							
area 1	1996	13	12	100%	-	-	0
area 2	1996	13	12	100%	-	-	0
area 3	1996	11	11	100%	-	-	0
large populations							
Peerdsbos		~600	24	100%	-	-	0
Merodese Bossen		300-400	15	100%	-	-	0

Table 3.2: The frequencies of the three control region alleles identified in the populations of red squirrels near Antwerp. The year in which sampling took place is given, along with the population size and the sample.

Pairwise Exact Tests (PETs) were carried out to compare each of the Belgian populations and the results are given in table 3.3. Where the populations being compared are both fixed for the same allele (for example Peerdsbos compared to Merodese Bossen), the populations are clearly identical and a PET is not possible. All the other comparisons have a probability of 1, or a probability that is clearly higher than the 0.05 significance level. The lowest probability is that comparing Gasthuisbos (1995) with Peerdsbos, but when a Bonferroni correction is carried out using the Dunn-Šidák method, the critical significance value for this comparison is reduced to 0.0007 (for a maintained error rate of 0.05) making this result also not significant. All the probabilities are far in excess of their corrected significance levels so it can be concluded that there is no genetic difference at this locus between any of the Belgian populations.

		BZ 95	BZ 96	GH 95	GH 96	T	KE	L	AWW 1	AWW 2	AWW 3	P	MB
Brede Zijpe 1995													
Brede Zijpe 1996		1											
Gasthuisbos 1995		1	0.38										
Gasthuisbos 1996		1	1	1									
Tallaarthof		1	1	0.69	1								
Kegelslei		1	1	1	1	-							
Luisbos		1	1	1	1	-	-						
Antwerp water works	area 1	1	1	0.48	1	-	-	-					
	area 2	1	1	0.49	1	-	-	-	-				
	area 3	1	1	0.48	1	-	-	-	-	-			
Peerdsbos		0.34	0.41	0.05	0.36	-	-	-	-	-	-		
Merodese Bossen		0.46	1	0.22	0.48	-	-	-	-	-	-	-	

Table 3.3: A table showing the probability values associated with each pairwise comparison of the Belgian populations using the Pairwise Exact Test.

3.3.3 A comparison of German and Belgian red squirrel populations

The Belgian and German samples have very different allelic distributions with no alleles in common; they represent distinct populations. The two groups also have completely different levels of intra-population diversity, as is illustrated by the results given in table 3.4 which gives some diversity statistics for both groups. Where possible, the significance of the differences was tested. The data for the mean number of pairwise differences in the German population was shown to be normally distributed using the Kolmogorov-Smirnov normality test (the Belgian sample involved too many comparisons for this test) and so could be examined using standard parametric tests. The mean number of pairwise differences were compared using a t-test and, as the nucleotide diversity is actually the mean sequence divergence, this could also be tested for significance using a t-test. A χ^2 test was carried out on the number of variable and non-variable sites to test for significant difference between the populations.

Diversity measure	Belgium	Germany	Significance test
Allelic diversity (sample size)	3 (129)	20 (25)	
Sequence divergence	0.95% - 1.91%	0.32% - 4.14%	
Mean number of pairwise differences	0.22	6.01	T= 35.25, p<0.0001
Nucleotide diversity	0.07%	1.9%	T= 35.25, p<0.0001
Number of variable sites (sequence length (bp))	7 (314)	35 (314)	χ^2= 20, p<0.0001

Table 3.4: A table showing the genetic variation present in the Belgian and German samples measured by a variety of diversity indices. Where a test for significance was possible the results are also given.

The allelic diversity present can only be compared with caution due to the large difference in sample size, but the nature of the results allow some conclusions to be drawn. The Belgian sample contains less than 1/6th the alleles present in the German sample, despite the German sample being over five times the size. It is quite clear that the Belgian populations contain dramatically fewer alleles than the German population.

The sequence divergence range is much larger in the German population than in the Belgian. The German sample contains the most and least divergent pair of alleles, so the Belgian range is contained within that of the German sample. The statistical tests all show that the mean number of pairwise differences, the nucleotide diversity and the number of variable sites are all significantly higher in the German population than in the Belgian. As the mean number of pairwise differences and the nucleotide diversity are different ways of measuring the same kind of variation, the t-test results are the same.

All these statistics show that the Belgian and German populations differed greatly in amount of sequence variation in the control region of the mitochondrial genome and these differences are highly significant. The Belgian populations have very low levels of variation, many of the populations have no variation at all at this normally highly variable locus.

3.4 DISCUSSION

PCR-SSCP was successfully used to detect genetic variation in a section of the mitochondrial control region of red squirrels from Belgium and Germany. The German sample showed a high level of allelic diversity, indicating that the control region of the mitochondrial genome in *S. vulgaris* has the potential to be highly variable and therefore a useful marker for population studies.

3.4.1 The reliability of PCR-SSCP

In the variable German sample 90% of the variation was detected using the PCR-SSCP technique under the conditions employed in this study. This is a reasonably high detection rate, detection rates of 90% and above have been considered to be very efficient in previous studies assessing the reliability of SSCP (Hayashi and Yandell 1993; Liu and Sommer 1994). The use of a long fragment (367 bp) in the analysis does not seem to have limited the efficiency of detection; the conditions used and the use of MDE gels may have contributed to the reliability of the technique. Fan *et al.* (1993) found that careful selection of operating conditions allows >90% detection of mutants in a sequence 354bp in length and Marklund *et al.* (1995) successfully used a PCR product 440 bp in length to detect sequence variants in different breeds of horses, although they did find it necessary to repeat some analyses. This was also found to be a necessity in this study where some gels were difficult to interpret. It was often hard to determine whether different samples on the same gel had the same banding pattern until they were run side-by-side. It was therefore necessary to re-run many of the samples next to each other to be sure of their haplotype.

It was also found that even the smallest difference should be taken as a possible mobility change due to mutation and not assumed to be variation due to running conditions. Many of the different alleles found in the German samples had extremely similar SSCP haplotypes. This made false positives (sequences with apparently variant banding patterns that do not actually vary in sequence content) more likely but by re-running the samples to allow closer examination, false positives were successfully eliminated. False negatives (sequences that do vary in nucleotide content but do not vary in banding pattern) can never be excluded, so the possibility must always be considered (Hayashi and Yandell 1993). In this study two alleles in the German population, one that differed by one base and the other differing by three bases from the consensus seuqence, failed to be detected under the conditions used. This illustrates the potential of this technique to be insensitive to sequence changes. There is always the chance, however carefully the conditions are optimised, that several mutations will coincidentally not affect the mobility of the fragment and therefore be missed.

Overall, PCR-SSCP was found to be a reliable technique when used with great care, the conditions had to be kept consistent between gels and samples were re-run if there was any doubt whatsoever as to their correct haplotype. Great care must be taken when reading gels so as not to miss any slight differences between samples. In this study, the technique was used successfully to detect allelic variation with almost as much efficiency as sequencing, but at a fraction of the cost and time required.

3.4.2 Variation within the German population

The German sample was taken from a large population not known to have experienced any population size reductions in its recent history, so typical levels of variation for this species may be expected in such a population. In fact, this population was found to be very variable with 20 different alleles identified in a sample of just 25 individuals giving a nucleotide diversity of 1.9%. An investigation into the structure of bank vole (*Clethrionomys glareolus*) populations in Norway found that this species in that area had a lower nucleotide diversity (0.9%) but a similar proportion of variable sites (11.1% in red squirrels compared to 11.5% in the vole) in the mitochondrial control region (Stacy *et al.* 1997). Worthington Wilmer *et al.* (1994) found varying levels of nucleotide diversity in the same region of the genome in the ghost bat (*Macroderma gigas*), the most variable population had a diversity measure of 2.13% which is higher than the bank vole or the red squirrel but similar to them. Stacy *et al.* (1997) felt that the bank vole sample contained levels of mitochondrial DNA variation typical for small rodents so it seems that the German population of red squirrels is also typical.

However, these patterns are probably very species or population specific as they are heavily dependent on the population histories, the life histories of the species and the geography of the areas inhabited. For this reason, generalising about such a big group of species such as the rodents may be misleading, although it can be useful to gain some indication of whether unusual amounts of variation have been found in a population. Rodents have been found to have a relatively fast evolving mitochondrial genome when compared to other mammals (Wu and Li 1985; Lopez *et al.* 1997) but divergences of as much as 12- 14% have been found in gazelles in East Africa (Arctander *et al.* 1996). Therefore, although the German red squirrels do have variation in the control region, mammals can in fact be much more variable.

3.4.3 Variation within the Belgian populations

The Belgian populations differ greatly from the German population in levels of variation at this locus. The Belgian populations are genetically impoverished, with only three alleles present in the large sample of 129 individuals, 97% of the samples contained the same allele and eight of the ten populations were fixed for it. The nucleotide diversity over the entire sample was only 0.07%, compared with 1.9% in the German sample. The sequences of the alleles were reasonably different, they diverged by 0.95%, 1.59% and 1.91% from each other. If the lineages had only recently diverged, only one or two differences would be expected between the sequences and lower values for the percentage divergence would be expected. Maybe this indicates a history of many divergent lineages in the Belgian red squirrels, most of which have been lost and this study has detected the few remaining. Alternatively, the rare alleles found in Brede Zijpe and Gasthuisbos are there due to the immigration of squirrels from outside the area carrying new alleles.

Now, due to the maternal inheritance of the mitochondrial genome, only one allele will remain in the fragment populations as even the two low frequency alleles detected in this study will be lost. This is a direct demonstration of how maternal inheritance and the reduced effective population size of the mitochondrial genome can result in a rapid loss of lineages from small populations. It is not unknown for small rodent populations to show no variation at the otherwise variable control region locus. The investigation into ghost bat population structuring found one variable population and several with very low levels of variation including one with no diversity (Worthington Wilmer *et al.* 1994) and a study looking at the effects of habitat fragmentation on rice rat (*Oryzomys palustris*) populations in Florida found one population that showed no variation in the control region, they concluded that this was due to a recent bottleneck (Gaines *et al.* 1997).

The most variable red squirrel populations in Belgium were two of the small fragment populations, Brede Zijpe and Gasthuisbos; these were the only populations in Belgium to show any variation. Only allele A was found in the larger populations of Peerdsbos and Merodese Bossen but, as these populations were only sampled it is possible that some alleles were missed. As there was no significant difference between the Belgian populations, no conclusions can be drawn about their relationships. Gene flow cannot be detected using a marker that shows no variation and relatedness cannot be quantified when the marker is the same in each population. As Moritz (1994) said, differences between populations would be more readily detected using mitochondrial DNA, but only if such variation exists.

It is possible that all of the Belgian populations, divided into small populations by habitat fragmentation, have, by chance, become fixed for the same allele but, if they were reasonably variable before becoming separated, then this is unlikely. It is more likely that they have a shared history; something caused a dramatic loss of variation, leaving allele A dominant in the whole population of squirrels in this area, before the current population structure was established. Once separated into small populations, the effects of drift removed any remaining alleles and left them all fixed for allele A. If they had been separated whilst there was still a reasonable level of variation at this locus, it would be expected that the populations would now be fixed for different alleles as the random effects of drift would affect each population differently.

3.4.4 The possible causes of reduced genetic variability

If it is assumed that at one time the Belgian squirrels displayed similar levels of variation to the large German population then the Belgian squirrels are clearly lacking in genetic variation in the mitochondrial control region. Several explanations for this loss of variation are plausible.

The mitochondrial genome does not experience recombination so it is one large totally linked locus. Most of the genome codes for important proteins so it is possible for any of them to become subject to selection. If an advantageous mutation occurs in one of these genes then that mutation may spread through the population by selection, this rapid spread of a mutation is known as a "selective sweep" (Ballard and Kreitman 1995) but is also sometimes referred to as "periodic selection" (Maruyama and Birky 1991). As the mitochondrial genome is linked then all of the alleles in that genome will also spread to dominate the population (Maruyama and Birky 1991; Ballard and Kreitman 1995). This is a process known as "hitch-hiking" and can lead to neutral and even slightly deleterious mutations spreading through the population by linkage to an advantageous mutation. Even though the control region is generally considered to be a neutral locus, it could have a history influenced by selection on other regions of the mitochondrial genome. In fact, selection on any maternally inherited factor, not just the mitochondrial genome, could influence the frequencies of different control region alleles (Ballard and Kreitman 1995). The influence of selection on a population is dependent on the population size. For selection to have had an effect on the population allele frequencies, the population involved would either have to be large or the force of selection would have to be extremely strong, otherwise it is drowned out by random drift (section 1.3.5).

Another explanation is that the squirrels in northern Belgium have suffered a bottleneck sometime in their recent evolutionary history. The effects of a bottleneck on the mitochondrial genome can be very dramatic, as illustrated by figure 3.6. However, it is not necessarily the case that the Belgian squirrels experienced a very severe bottleneck as some of the variation may have been maintained, until recently, in the populations of Brede Zijpe and Gasthuisbos. It is also possible that this variation is due to recent immigration into these populations from outside the area and a bottleneck may have, in fact, left the squirrels in this region totally devoid of mitochondrial variation. The metapopulation examined in this study is not a closed system, therefore it is impossible to distinguish between these alternative outcomes.

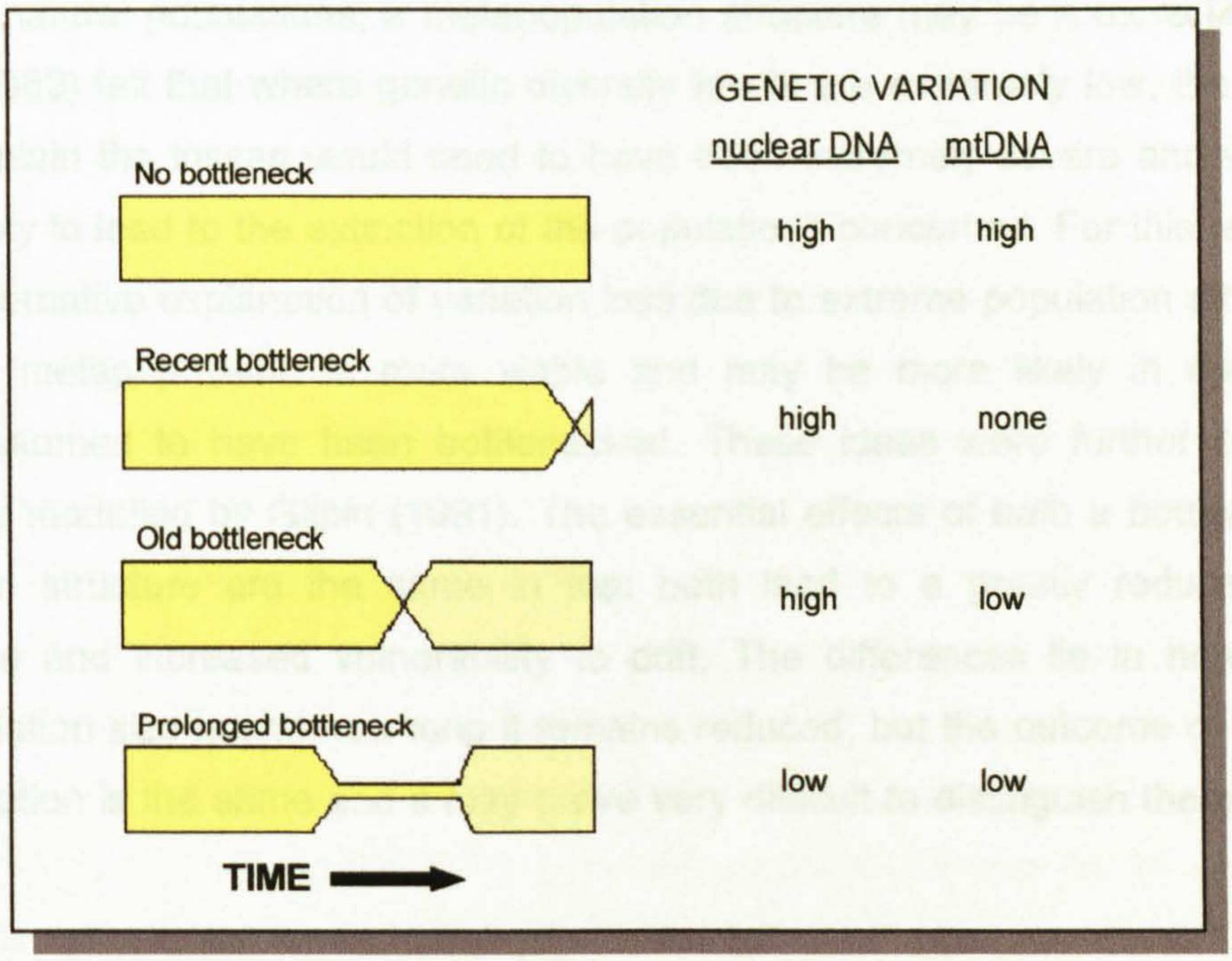


Figure 3.6: A diagram illustrating the expected loss of variation experienced by the nuclear and mitochondrial genomes as a result of different degrees of bottleneck, the depth of the coloured bar describes the population size (copied from Wilson *et al*. 1985).

Even if some diversity did remain after a bottleneck, the process of random lineage sorting would quickly have eliminated many lineages from small populations; the mitochondrial lineages of females that only have sons or simply do not reproduce are immediately lost (Avise *et al*. 1984; Harrison 1989). Indeed, diversity can be lost very quickly from a population without it necessarily suffering a sudden reduction in size as alleles will be lost faster than they can be replaced by mutation if a population is of a limited size for a prolonged period.

A metapopulation structure can result in large reductions in the amount of variation present in the metapopulation as each small population is greatly affected by drift. Small populations are also vulnerable to extinction causing all the variation unique to that population to be lost. If the area is then recolonised it can only be founded by individuals carrying variation present in other nearby populations. This kind of turnover will result in all the populations being closely related and similar in allele content. A metapopulation with a high rate of turnover will soon become low in genetic diversity (Gilpin 1991).

These theoretical ideas about the effects of a metapopulation structure on genetic variation have lead to the suggestion that where a bottleneck has been invoked to explain low levels of variation in natural populations, a metapopulation structure may be a more likely cause. Pimm *et al.* (1989) felt that where genetic diversity levels are extremely low, the bottleneck required to explain the losses would need to have been extremely severe and would have been more likely to lead to the extinction of the populations concerned. For this reason, they felt that the alternative explanation of variation loss due to extreme population structuring as is found in a metapopulation is more viable and may be more likely in many natural populations assumed to have been bottlenecked. These ideas were further iterated and mathematically modelled by Gilpin (1991). The essential effects of both a bottleneck and a metapopulation structure are the same in that both lead to a greatly reduced effective population size and increased vulnerability to drift. The differences lie in how small the effective population size is and how long it remains reduced, but the outcome of a reduction in genetic variation is the same and it may prove very difficult to distinguish the two possible causes.

It is perhaps surprising to find no variation in the larger Belgian populations of Peerdsbos and Merodese Bossen. Peerdsbos is around 30km away from Brede Zijpe and Gasthuisbos, but it is fixed for the same allele. This indicates that the reduction in variation experienced by these populations of red squirrels has affected a wide geographical area. It also perhaps suggests a bottleneck has been involved in removing variation at some time. If the loss of variation was simply due to population structuring with drift acting independently on each population then, over such a large area, it may be expected that distant populations may be fixed for different alleles. Metapopulation structuring alone may not have been enough to remove most, if not all, of the variation from the larger populations of Merodese Bossen and Peerdsbos, a bottleneck would probably have been required. However, if the variation in these squirrels was already reduced by a mild bottleneck, so that allele A dominated throughout the area before the metapopulation structure was established, then drift acting on the separate populations may have led to the fixation of the same allele in all the populations.

A severe bottleneck, on the other hand, would have eliminated most of the variation and a subsequent population expansion would also lead to the same allele spreading over a large area, the current metapopulation structure would then not have had any further effect on mitochondrial variation levels.

The dominance of only one allele over such a large geographical area and the apparent fixation of the same allele in the two larger Belgian populations seems to implicate a bottleneck of some degree during the evolutionary history of these populations. Recently, the metapopulation structure will have had a role in further reducing mitochondrial DNA diversity, if any remained, and maintaining it at low levels. The relative importance of a bottleneck and metapopulation structuring in determining the current genetics of these populations are unknown and indecipherable with these results.

3.4.5 The recent effects of habitat fragmentation

More recently, habitat fragmentation in this area of northern Belgium has led to the restriction of red squirrels to many separate small populations, isolated from each other. The populations of Brede Zijpe and Gasthuisbos were the only ones found to have any detectable variation in the form of three reasonably diverged alleles, two of which were present in both populations. These may be the remnants of ancestral diversity, in turn suggesting that these populations are remnants of a larger ancestral population, or they may be introduced alleles brought into the area by immigrants from outside the study area. The variation, whether it was ancestral or imported, is likely to have been lost completely, illustrating the effects of drift in removing low frequency alleles from small populations.

The other fragment populations of Tallaarthof, Luisbos, Kegelslei and the areas around Antwerp water works are all fixed for the same allele. They may either be the result of recent colonisation events or the fragments of a larger population, it is not possible to tell the origin of the populations from this data. These populations are smaller than Brede Zijpe and Gasthuisbos, so they would be expected to loose variation more quickly due to the effects of drift. With populations so small it is not surprising to find them totally devoid of variation at this locus. The on going effect of drift has eliminated all the variation from this loci in these populations and, whilst the populations remain small, they cannot be expected to gain and maintain any new polymorphisms.

3.4.6 Conclusions

The sample from the German population contains typical levels of genetic variation in the control region of the mitochondrial genome for a small rodent population. Comparison of the German and Belgian populations showed that the Belgian populations have remarkably low levels of variation at this locus, all the populations are fixed or nearly fixed for the same allele. It is likely that a bottleneck has contributed to this reduction in diversity at sometime in the history of the squirrels in this area and that it predates the establishment of the present population structure because all the current population fragments are fixed for the same allele indicating a shared history. If the bottleneck was recent then it is possible that it could have had a dramatic effect on the mitochondrial genome but barely been detectable in the nuclear genome, as illustrated by figure 3.6, therefore, analysis of microsatellite markers in these populations may shed more light on the history of the populations and the bottleneck they endured.

Recent habitat fragmentation has reduced the red squirrels in this area of Belgium to the occupation of small, isolated woodlots. The metapopulation structure they are now experiencing leaves them extremely susceptible to the impoverishing effects of genetic drift, which has already eliminated any mitochondrial variation that remained prior to the establishment of the current population structure. Now, these populations are probably completely devoid of variation in the mitochondrial genome and, even if immigrants introduce a new allele to a population, it has a low probability of being retained due to the powerful effects of random genetic drift.

CHAPTER FOUR:

THE ISOLATION OF MICROSATELLITE LOCI

4.1 INTRODUCTION

One of the many advantages of the use of microsatellites in ecological and evolutionary studies is the speed with which they can be applied and results gathered. This is due to the development of polymerase chain reaction (PCR) technology which allows the short loci to be amplified and visualised. However, the use of PCR requires the development of primers to use in the amplification reactions and this can be a time consuming process.

Usually it is necessary to develop primers for each species studied (Ashley and Dow 1994). In some cases primers have been shown to amplify polymorphic loci in species closely related to the one they were developed for, but often they are less consistent than in the original species (McDonald and Potts 1997). For the primers designed for one species to work in another a high degree of homology is required in the sequences flanking the microsatellite loci targeted by the primers. Most microsatellites are found within the non-coding regions of the genome which have high mutation rates. This leads to the rapid accumulation of mutations in the flanking sequences where the primers bind, increasingly inhibiting the successful annealing of PCR primers as the target sequences diverge (Primmer *et al* . 1996b). The chance of cross species amplification of a locus is therefore inversely related to the evolutionary distance between the species.

Primmer *et al.* (1996b), in their survey of cross-species microsatellite amplification in birds, also found that the proportion of polymorphic loci amongst the successfully amplified loci decreased with increasing genetic distance between species. Kondo *et al.* (1993) tried to amplify mouse primers in rats and vice versa with only 25 out of 153 (16%) and 20 out of 166 (12%) working respectively and only 8% and 4% of these were polymorphic. The selection process for loci during isolation means that primers are often only developed for the longer loci found in the subject species. These loci are not necessarily the best selection for another species and can only be expected to be most useful in the species from which they have been developed (Primmer *et al.* 1996b). So, although some primers may prove useful for microsatellite analysis in other species, it may be preferable to develop new sets for the species to be studied if the best results are to be obtained.

4.1.1 The Isolation of microsatellite loci

Microsatellite primer sequences are stored on sequence databases, such as GenBank and EMBL, and are published in scientific literature. Notably, the journal *Molecular Ecology* (Blackwell) has a section devoted to the publication of new primer sequences. Thus it may be possible to find previously developed primers for the same or closely related species to use in a study. The amplification of loci using these primers can be attempted and, if successful and the loci prove to be variable, the reactions can be optimised as described in chapter 5. All possible microsatellite loci should be tried in the new species, not just those found to be most polymorphic in the original species as they may vary in length and polymorphism between species (Strassmann *et al.* 1996). If new loci need to be isolated this can be done using standard protocols as described by Schlötterer (1998b) and Strassmann *et al.* (1996), and discussed below. Figure 4.1 summarises the methods used in this study.

4.1.1.1 Constructing a genomic library

The construction of genomic libraries is traditionally seen as a technically demanding process but commercially prepared cloning kits and high performance enzymes are now available which make many of the steps quicker to perform and more reliable. Even so, it is a multi-stage process and as such suffers from the unreliability intrinsic to such procedures; if any of the steps does not work perfectly the result can be an inferior library with the need to repeat the whole process. Strassmann *et al.* (1996) described the isolation of microsatellites as "tedious and time consuming". This is particularly true when the construction of a library needs to be repeated several times in order to develop a useful suite of microsatellite loci. In addition to this, researchers isolating microsatellite loci are often not very experienced with molecular techniques, either coming from an ecological background or being students learning as they go. This increases the probability of mistakes being made during the many steps of the process.

A genomic library is a collection of DNA fragments from the genome of a particular species, inserted into a microbial vector (McDonald and Potts 1997). The construction of a partial library of size-selected DNA fragments is necessary for the isolation of microsatellites. This use of a genomic library was first described by Rassmann *et al.* (1991). A library constructed for microsatellite isolation does not need to be fully representative of the whole genome as only a few short sections are required. It is constructed with segments of DNA short enough to be fully sequenced using only two primers flanking the inserted DNA segments.

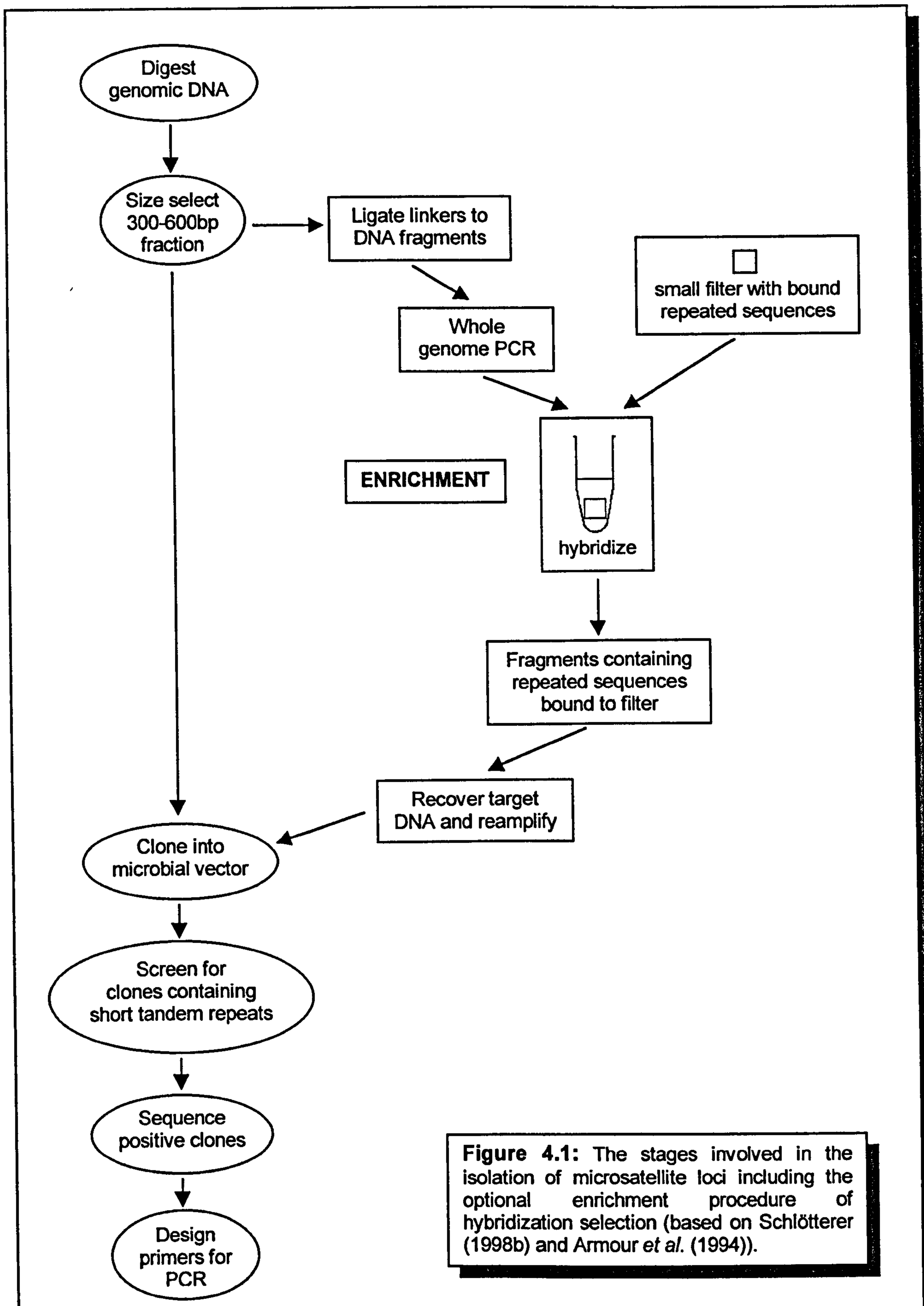


Figure 4.1: The stages involved in the isolation of microsatellite loci including the optional enrichment procedure of hybridization selection (based on Schlötterer (1998b) and Armour *et al.* (1994)).

To construct a library, high molecular weight DNA is fully digested using a restriction enzyme that leaves ends compatible with the insert site of the chosen plasmid vector. A commonly used combination is to digest the genomic DNA with SauIIIA and the plasmid with BamH1 (McDonald and Potts 1997). Sections of genomic DNA within the required length range are selected by running the digest through an agarose gel next to a size marker and excising the appropriate portion. These segments of DNA are then ligated into the vector, reconstituting the circle of plasmid DNA. A culture of competent bacterial cells is then infected with or "transformed" with the vectors containing genomic DNA. This culture is then grown for no longer than one hour, this simply revives the culture as too much growth would lead to replication of particular genomic fragments in the library (Strassmann *et al.* 1996).

The cultures can then be plated out onto agar and blue-white screening identifies colonies containing a vector (see section 2.2.5.4). These colonies are picked, grown and replicated onto a membrane. The membranes are then probed for the presence of short repeated motifs using artificially constructed and radioactively labelled sections of tandemly repeated DNA. Colonies indicated by the probing are sequenced and if a microsatellite is found, primers are designed to flank the repeat.

4.1.1.2 Enrichment techniques

The standard method for isolating microsatellite loci by simply screening a partial library can work well if only a few dinucleotide repeats are required, but unenriched libraries can be very inefficient. Dinucleotide repeats are the most common in eukaryotic genomes, even so the expected number of $[CA]_n$ repeats in a small insert library is only 1 per 100-400 colonies (Ostrander *et al.* 1992). Tri- and tetra- repeats are more rare and extremely large libraries would be required to isolate several of these loci.

Several different techniques have been developed to enrich libraries with repeat loci. Most of these depend on the process of whole genome PCR (Kinzler and Vogelstein 1989). Short double-stranded lengths of sequence called "linkers" are attached to each end of the DNA fragments and used as target sequences to prime a polymerase chain reaction (PCR) which amplifies the linkered DNA molecules. In theory, all the molecules of the genomic digestion could be amplified, but in practice it is likely that there will be some bias and a few molecules may be amplified to much higher frequency than others. This is due to the exponential way in which PCR works: any slight bias introduced in one of the early cycles will be exaggerated with each subsequent cycle.

Once whole genome PCR has been carried out, sufficient DNA in short fragment lengths is then present to allow manipulation and selection of a few preferred segments. In this case, sections containing microsatellite loci are required. Several methods exist to select for such fragments, some of which only allow for the selection of one type of repeat (for example see Nishikawa *et al.* 1995 and Ostrander *et al.* 1992). When collecting loci to use in a population genetics study a range of different loci is preferred. It is also necessary to be able to select many different repeat types, when isolating tri- and tetra- repeats which are relatively the rare, in order to find sufficient variable loci. Therefore methods that can enrich for many different repeat types are more appropriate, such as those described by Armour *et al.* (1994) and Prochazka (1996).

Prochazka's hybrid capture technique involves hybridizing biotinylated repeat probes (artificially constructed sections of repeated DNA with biotinylated ends) to the amplified fragments of the genome. These probes bind to matching repeat sequences and, when streptavidin-coated paramagnetic beads are added, the biotinylated probes stick to the beads thereby "catching" the DNA fragments containing the required sequences. The captured sequences are washed off the beads and amplified again before being cloned into an enriched library.

The method used in this study is that described by Armour *et al.* (1994). This also involves selecting DNA fragments containing short repeat sequences prior to constructing the library by hybridization using short target probes. In this case the target sequences are bound to a small filter and the digested DNA hybridized to it, sections of DNA containing repeats will bind to the target DNA on the filter. These fragments can then be washed off, reamplified and used to construct a library.

4.1.1.3 Primer design

If many sequences are obtained for microsatellite loci it may be preferable to prioritise the design of primers so those most likely to be variable are tested first. Dinucleotide repeats are generally the most commonly occurring type, but they are often more difficult to score due to the small differences in allele size and problems with "stutter" bands (see chapter 5). Tri- and tetranucleotide repeats are becoming more popular for use by population geneticists as they are not only easier to score, but some researchers have also found them to have a higher mutation rate than dinucleotide repeats (Ellegren 1995; Weber and Wong 1993). This pattern is counterintuitive as dinucleotide repeats would be expected to experience more slip-strand mispairing and it has been contradicted by more recent studies. Chakraborty *et al.* (1997) felt

that in some studies the trend towards more mutable tetranucleotide repeats was due to one or two highly variable loci and when they are removed the remaining trend is for dinucleotide repeats to show more mutation than the tetranucleotides. Overall, the consensus does seem to be that tetranucleotide repeats are more variable than dinucleotide repeat loci, but there have been few comparative studies (Jame and Lagoda 1996).

Early studies suggested that loci containing a higher number of repeats are more polymorphic than shorter ones, but this has not always been confirmed by more recent work (Schlötterer 1998b). However, such a trend would be expected if slippage is the usual cause of mutations as long lengths of repeats may be more vulnerable to mispairing events. Strassmann *et al.* (1996) recommended that regions containing 8 or more repeats may be of use and are worth testing for variability; this is also the threshold size for microsatellite expansion identified by Rose and Falush (1998; see section 1.5.2.1). Perfect repeats are consistently more polymorphic than compound and interrupted repeats so are therefore more preferable targets for PCR.

Computer programs, such as *Oligo*, which was used in this study, can be used to aid in the design of primers. The full sequence is entered and the program carries out a search for the best primer pair. The two primers should be of similar length (preferably at least 20bp long to reduce the probability of spurious binding), have 40-60% GC content and have very similar melting temperatures (T_m) (Hoelzel and Green 1998). Care should be taken that they do not bind to each other or form hairpin loops, especially at the 3' end which actually primes the DNA replication; in practice it is usually possible to avoid overlaps of more than 3 bp.

A few "G" or "C" nucleotide residues at the 3' end is often recommended as the G-C bond is the strongest, allowing the primer to anneal to the template DNA more tightly and reliably. However some researchers feel that too many G or C bases at the 3' end of the primer can lead to mispriming as the G/C residues could bind to complementary sequence at the wrong place in the genome well enough to prime the reaction regardless of the mismatch in the remainder of the primer (pers. com., "micro-sat" internet newsgroup). Therefore, ideally, it seems that some G or Cs at the 3' end are useful but not too many, although the final 3' residue should be G or C. An A-T rich 3' end should definitely be avoided due to the possibilities of mispriming. In practice, there is often not a lot of room to manoeuvre when designing the primers, it is simply a case of identifiying the best possible with the available sequence.

It has been suggested that primers close to the repeat sequence are more likely to encounter amplification problems due to mutations in the primer site, but this has been denied by other workers (Callen *et al.* 1993). The aim is to have a product of between 100 and 200 bp in size; short products run faster through electrophoresis gels and so are easier to score (Strassmann *et al.* 1996). Once manufactured the primers can be tested by amplifying the required section from the vector using a standard PCR protocol to ensure they bind to the correct points in the genome.

4.1.2 Testing the loci

Once isolated, before being used as markers in population studies, the loci must be tested for the presence of null alleles and for linkage. Both of which, if they are not recognised, can lead to false results or the incorrect interpretation of the results.

4.1.2.1 Null alleles

Null alleles are alleles that are present in a sample but are not identified in the analysis of a sample or population using the chosen technique. They have been a recognised problem since the early days of protein polymorphism studies and in more recent minisatellite work (Callen *et al.* 1993). In microsatellites they are alleles that have not amplified for some reason and are assumed to be the result of mutations in the primer site preventing the priming of the PCR reaction. Paetkau and Strobeck (1995) identified a null microsatellite allele in bears and, by redesigning the offending primer and sequencing the amplified product, found that the null allele was due to a G→C transversion at the exact 3' end position of the original primer. Callen *et al.* (1993) found a more dramatic example in humans of a null allele due to an 8bp deletion located at the 3' end of the primer target sequence. Other factors such as template DNA quality can also affect the success of amplification reactions and may be the cause of more occasional non-amplified alleles.

Null alleles can be identified in parentage studies by the apparent non-inheritance of alleles. For example, if the parental genotypes are AB and CD but the offspring appears homozygous for C then either allele A or B from the first parent is not being identified in the offspring and is a null allele. Null alleles have often been identified in mother/offspring comparisons where the offspring does not appear to carry either of the maternal alleles (Pemberton *et al.* 1995).

Where known relatives are not available, it is possible to test for null alleles by testing for heterozygote deficiency in a population. If the observed number of heterozygotes is less than that expected in a population conforming to the Hardy-Weinberg law then null alleles may be suspected. Care must be taken however because such a deviation from Hardy-Weinberg expectations could also be due to the Wahlund effect (apparent heterozygote deficiency that is in fact the result of population subdivision) or to inbreeding so these possible causes must be excluded.

Under Hardy-Weinberg expectation the number of heterozygotes (H_e) in a population is:

$H_e = 1 - \Sigma x_i^2$ where x_i is the frequency of each allele.

Heterozygote deficiency can be tested for significance by carrying out a goodness of fit test, such as Fisher's exact test, on the observed and expected numbers of heterozygotes and homozygotes. If significant, the frequency of null alleles (*r*) can be estimated as (Brookfield 1996):

$r = (H_e - H_o) / (H_e + H_o)$ when all the samples showed an amplified product

or

$r = (H_e - H_o) / (1 + H_e)$ when some samples have apparently not amplified

where H_e is the expected number of heterozygotes
and H_o is the observed number.

The data can then be corrected to allow for the null alleles (van Treuren 1998).

4.1.2.2 Linkage disequilibrium

Loci are linked when they are located physically close to each other in the genome and so do not segregate independently during meiosis. Recombination or crossing over may occur between the two loci to separate them, the further they are apart the more recombination will occur between them, therefore the degree of linkage is approximately proportional to the distance between them.

Linkage can be measured by examining many parent and offspring groups. If loci are linked then the same allelic combinations found in the parents will usually be found in the offspring. If they are not linked, then the alleles found in the offspring will be random combinations of the parental alleles. The degree of linkage can be estimated from the frequency with which the parental combinations are maintained in the next generation.

The non-random distribution of alleles is termed linkage disequilibrium and can be caused by chance in small populations, by non-random mating or the recent mixing of previously separated populations, as well as by linkage in the genome. When family groups are not available for analysis then linkage disequilibrium can be estimated instead but it is important to remember that several other factors could account for any disequilibrium seen as well as linkage.

In a haploid genome disequilibrium can be directly measured from observable genotypes but in a diploid organism this is not possible. In an individual with the genotype $A_1A_2B_1B_2$ it is not possible to determine whether the alleles have segregated as A_1B_1 and A_2B_2 or as A_1B_2 and A_2B_1. It is therefore not possible to measure directly whether certain allelic combinations are more common in a population than would be expected at random. Statistical estimates can be made of the effect of linkage on a group of loci using contingency tables. A table can be constructed for each pair of loci as illustrated in figure 4.2 and an exact test carried out using a Markov chain, as described in section 3.1.2.1. This calculates the probability of the genotypes seen occuring by chance if the null hypothesis of no association between the alleles is correct and so can be used to indicate linkage between loci.

		RSμ5			
		A_1A_1	A_1A_2	A_2A_2	total
RSμ6	B_1B_1	1	1	0	2
	B_1B_2	4	2	1	7
	B_2B_2	6	4	0	10
	total	11	7	1	19

Figure 4.2: An example of a contingency table for a pair of loci and the incidences of each combination of genotypes in a population. The data used is for loci RSμ5 and RSμ6 amplified in individuals from the Belgian Brede Zijpe population, each of which showed two alleles in this population (A_1 = allele 139 and A_2 = allele 141 in RSμ5; B_1 = allele 122 and B_2 = allele 128 in RSμ6, the result of the exact test on this table is *p=0.81*).

4.2 METHODS

Three libraries were created using the enrichment procedure of hybridization selection described by Armour *et al.* (1994). Each library varied slightly in methodological detail; the following is the general protocol used in this study.

4.2.1 DNA preparation

Two or three high quality DNA extractions from Belgian red squirrel samples were digested using the restriction enzyme *Mbo1* (Gibco BRL). This is a four base cutter recognising and cleaving the double stranded DNA oneside of the motif -GATC- leaving a four base overhang at each end of the fragments.

The digestions were set up as follows:

freshly extracted DNA	15µl	
React 2 buffer (GibcoBRL)	2µl	
Spermidine	2µl	
Mbo1	1µl	(10 units)

The reactions were incubated at 37°C overnight.

The concentration of the digested DNA was measured using a GeneQuant RNA/DNA calculator (Pharmacia Biotech) and ~3mg of each digestion was run through a 1.5% agarose gel (see section 2.2.3) with ϕX174 DNA Hae III digest size marker (Kramel Biotech). The gel running time was kept to a minimum to facilitate the separation of the genomic fragments over the smallest possible area of agarose (the smaller the piece of agarose the more efficient the gel extraction process). The fragments of the digestion between 400bp and 1000bp in length were excised and extracted from the gel matrix using the QIAquick Gel Extraction Kit (Qiagen). For library 3, fragments between 400bp and 1300bp were excised. The concentration of the excised DNA was measured using the DNA DipStick Kit (Invitrogen), which can be used to measure very low concentrations of DNA.

4.2.2 Linker construction

The oligos SAULA (5'- GCGGTACCCGGGAAGCTTGG -3')
with SAULB (5'- GATCCCAAGCTTCCCGGGTACCGC -3') (ROYLE *et al.* 1992)
were used in libraries 1 and 2

and LinkerA (5'- GGGTAGGATGGGGGATGGG -3')
with LinkerB (5'- GATCCCCATCCCCCATCCTACCC -3') (U.H. Refseth, pers. com.)
in constructing library 3.

The linkers were created by mixing equal quantities of each complementary oligo with 14% annealing buffer (Pharmacia, from the T7 sequencing mixes) and raising the temperature of the mixture to 60°C then slowly reducing it to 0°C over a 2 hour period on a thermocycler. During this process, denatured oligos should pass through their optimum annealing temperature and anneal to each other to form short double stranded fragments with a four base pair overhang of sequence complementary to that left by the digestion with *Mbo1*.

4.2.3 Ligation of linkers to the digested DNA

The ligation of the linkers to each end of the DNA fragments was carried out using the ligase enzyme and buffer from the pGEM-T vector system (Promega). 2μg of linkers were ligated to 200ng of DNA; ten times more linkers than DNA were used to saturate the system and to try to ensure that the ligation reaction created molecules consisting of linker-fragment-linker rather than fragment-linker-fragment. To the mixture of linkers and size selected DNA, 1μl T4 DNA ligase and 1μl ligase buffer (10x) was added per 10μl of reaction volume and the reaction was incubated at 14°C overnight.

For libraries 2 and 3, the whole reaction was then run out on a 1.5% agarose gel with a size marker (see section 2.2.3). The same size fraction was again excised and cleaned with the QIAquick gel extraction kit (Qiagen). In constructing library 1, the excess of linkers was removed by cleaning the reaction with the QIAquick PCR purification kit (Qiagen) but the shorter fragments that may have been incidental products of the ligation reactions would not have been removed.

4.2.4 Whole genome PCR

The purpose of ligating linkers to the DNA fragments is to provide flanking sequences that can be used to prime a PCR reaction. All the appropriately sized fragments of genomic DNA can then be amplified in a single reaction. As the linkers are the same on either end of the fragments, only one primer is needed in the reaction: one of the original oligos used in the linker construction sequences is used. The reaction mixture was set up as in table 4.1.

Reagent	Conc^{n}.	Quantity	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	2.5 µl	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	4 µl	0.2 mM
Magnesium chloride (Advanced Biotechnologies)	25mM	3 µl	3 mM
primer SAULA	10µM	2 µl	0.8 µM
Red Hot *Taq* (Advance Biotechnologies)	5 U/ µl	0.2 µl	1 U
template		4 µl	
sterile distilled water		8.5 µl	

Table 4.1: The PCR reaction mix used to perform whole genome PCR.

The reaction was run on a thermocycler (PTC-100, MJ Research, Inc.) using the following cycling program:

Step	Temperature (ºC)	Time (minutes)
1	95	1
2	95	1
3	67	1
4	70	2
5	Go to step 2	31 more times
6	72	5

The whole reaction was run out on a 1.5% agarose gel (section 2.2.3), the appropriate size fraction was excised and extracted using the QIAquick gel extraction kit (Qiagen). The concentration of eluted DNA was measured using the DNA DipStick Kit (Invitrogen). The size selection at this stage was omitted in the creation of libraries 1 and 2; the concentration of DNA in the PCR product was estimated directly from the agarose gel by comparison with the size marker and then used direct from the PCR in the hybridisation selection process.

4.2.5 Hybridization selection

4.2.5.1 Preparation of target DNA sequences

Short stretches of synthesised DNA repeats were manufactured by the Oligonucleotide Synthesising Unit at the University of Nottingham; the oligo repeats are shown in table 4.2. These were created so that matching pairs with complementary seqence could be annealed and extended using linear PCR. This is a PCR reaction set-up with each oligo acting as both primer and template. During the cycles of the reaction the complementary oligos anneal to each other at various places and prime the extension reaction creating longer and longer molecules with each cycle. The reaction mix was set up as shown in table 4.3.

Synthesised oligo repeat pairs:		Used in the contruction of library:		
		1	2	3
$(GATA)_n$	$(TATC)_n$	✓	✓	
$(GACA)_n$	$(TGTC)_n$	✓	✓	✓
$(CCAT)_n$	$(GTAG)_n$	✓	✓	
$(ACCT)_n$	$(GGAT)_n$		✓	✓
$(TTGG)_n$	$(CCAA)_n$			✓
$(GGAA)_n$	$(TTCC)_n$			✓
$*(TTTG)_n$	$(CCAA)_n$			✓
$*(TTTC)_n$	$(GGAA)_n$			✓
$(GTA)_n$	$(CAT)_n$		✓	
$(GAT)_n$	$(CTA)_n$		✓	
$(GCT)_n$	$(CGA)_n$		✓	
$(CGT)_n$	$(GCA)_n$		✓	
$(TCC)_n$	$(AAG)_n$		✓	
$(CAC)_n$	$(GTG)_n$		✓	
$(GTT)_n$	$(AAC)_n$			✓
$(AAG)_n$	$(TCT)_n$			✓
$(GT)_n$	$(AC)_n$			✓

Table 4.2: The synthesised repeat sequences used as target sequences for hybridization selection. The pairs marked * are not perfect matches but were extended successfully. In addition, the amplified human microsatellite wg1c4 (Armour *et al.* 1994) and the repeat $(GGAGGGAA)_n$ were used as a target sequences for the construction of libraries 1 and 2 respectively.

Reagent	Conc^{n}.	Quantity	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	5 µl	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	4 µl	0.1 mM
Magnesium chloride (Advanced Biotechnologies)	25mM	3 µl	1.5 mM
oligo 1	~35µM	2 µl	1.4 µM
oligo 2	~35µM	2 µl	1.4 µM
Red Hot *Taq* (Advance Biotechnologies)	5 U/ µl	0.2 µl	1 U
sterile distilled water		33.8 µl	

Table 4.3: The reaction mix used in the linear PCR reactions to extend the oligonucleotide repeat sequences.

The extension reactions were run on a thermocycler (either PTC-100 or PTC-200, MJ Research, Inc.) with the following program:

Step	Temperature (ºC)	Time (minutes)
1	93	3
2	93	1
3	39	1
4	72	1
5	Go to step 2	29 more times
6	72	2

The whole reactions were run out on 1.5% agarose gels (section 2.2.3) and fractions of 200bp and above were extracted using the QIAquick Gel Extraction Kit (Qiagen). Further rounds of extension PCR (three in total) were carried out as above with 10µl of cleaned extended DNA as template (with the amount of water in the reaction mix reduced appropriately) and 1µl (0.7 µM) of each original oligo added as primers until a large concentration of long target DNA sequences was visible on an agarose gel.

After the final round of extension, the reactions were not size selected and gel extracted as this results in a large reduction in DNA concentration; a high concentration was required for the hydridization selection. After several rounds of size selection and extension, there were very few short fragments remaining so size selection was not really necessary. A 5µl aliquot of the reaction was run on an agarose gel to check that it had been successful. In the

construction of libraries 1 and 2 the extended target DNA sequences were used neat from the extension reactions and the concentrations were estimated from the agarose gels. In construction of library 3 the target DNAs were ethanol precipitated to remove the other PCR components using the following protocol and the concentrations of each target sequence were measured using a GeneQuant RNA/DNA calculator (Pharmacia Biotech).

1. Add the remaining 45µl of PCR product to 200µl 100% ethanol and 5µl 3M Sodium acetate.
2. Vortex the solution briefly and incubate at -80°C for 15 minutes.
3. Carefully remove all the solution above the pellet.
4. Rinse pellet with 400µl of 70% ethanol.
5. Carefully remove all the ethanol and dry the pellet in an oven (55°C).
6. Resuspend the pelleted DNA in 30µl sterile distilled water by incubating overnight at 55°C.

4.2.5.2 Preparation of filters and genomic DNA

The target DNA is bound to small pieces (approximately 3mm x 3mm) of Zeta-Probe GT Genomic Tested Blotting Membrane (Bio-Rad). For library 1, four filters were prepared each with ~1µg of mixed target DNA; for libraries 2 and 3, the target DNA sequences were divided into two and three groups respectively and ~1µg of DNA was bound to each.

The target DNA was denatured and neutralised by the addition of Potassium hydroxide to a final concentration of 150mM and 1/4 volume 1M tris-Hydrochloric acid. The small filters were placed on a layer of filter paper and dotted with drops of the prepared target DNA solution. The filter paper helped to draw the solution through, leaving the DNA on the surface of the filters. The filters were left to dry and then exposed to UV to bind the DNA to the filter using the dry membrane crosslinking program on the GS Gene Linker (Bio-Rad).

The filters were incubated in 1ml pre-hybridisation solution (0.5M "phosphate buffer" (Na_2HPO_4, pH 7.2), 7% SDS) rotating at 65°C for 45 minutes

4.2.5.3 Hybridization

1μg of the mixed amplified genomic DNA was also denatured and neutralised with KOH and tris- HCl as for the target DNA. The filters were transferred to 100μl fresh pre-hybridization solution and the input DNA was added. The reaction was incubated, rotating at 65°C, overnight.

The filters were washed twice in 1ml of pre-warmed 0.2xSSC/0.01% SDS by rotating at 65°C for 5 minutes; this removed the non-specifically bound genomic DNA. The remaining DNA bound to the target repeats was recovered from the filters by submerging them in 50μl of 50mM KOH for a few minutes followed by the addition of 50μl 50mM Tris-HCl. This solution was removed and the DNA recoved by precipitation with the addition of 1/10 volume 3M Sodium acetate and 2 volumes 100% ethanol as described at the end of section 4.2.5.1. 1μg of one of the linker oligos was also added as a carrier to assist in the precipitation. The pelleted DNA was resuspended in 20μl of sterile distilled water by incubation at 55°C overnight.

4.2.6 Whole genome PCR

The selected DNA fragments were used as template in another whole genome PCR reaction exactly as described in section 4.2.4. During library 3 construction, 50μl sterile distilled water was put over the filters and heated to 95°C for 5 minutes. This denatured any DNA that may have remained bound to the filters, releasing it into the solution. The water was then used as a template in another whole genome PCR reaction.

The whole PCR reactions were run through a 1.5% agarose gel, size selected as before (section 4.2.4), and cleaned with the QIAquick kit. In the case of library 3, the amplified DNA from the second PCR seeded by the denatured water was then combined with the other amplified selected DNA. In constructing library 1, the post-selection PCR reactions were not size selected but cleaned using the QIAquick PCR Cleaning Kit (Qiagen). The concentrations of eluted amplified DNA were measured using either the DNA DipStick Kit (Invitrogen) or the GeneQuant RNA/DNA calculator (Pharmacia Biotech).

4.2.7 Assessment of the success of the enrichment procedure

During the construction of library 2, the pre- and post- selection PCR reactions were compared for target repeat sequence content by probing each with the target DNA oligos. The same quantity of DNA from each whole genome PCR reaction was run through a 2% agarose gel until the DNA molecules were well spaced down the gel. Southern blotting was carried out to transfer the DNA to a piece of Zeta-Probe GT Genomic Tested Blotting Membrane (Bio-Rad) using the following standard protocol adapted for a small gel.

1. Soak the gel in 0.2M Hydrochloric acid for 7 minutes followed by 1.5M Sodium chloride, 0.5M Sodium hydroxide for 15 minutes to denature the DNA and finally neutralise by soaking in 1M Tris, 1.5M Sodium chloride for 7 minutes twice.

2. Trim the gel to remove the loading wells and set up the Southern Blot as shown in figure 4.3. Lay a glass or plastic plate over a tub containing 20xSSC (3M Sodium chloride, 0.3M Sodium citrate). On this plate lay a strip of 3mm chromatography paper (Whatman) pre-wetted in 20xSSC bent so each end is submerged in the reservoir of SSC. This acts as a wick keeping the gel moist. Lay the agarose gel on top of the wick surrounded by a layer of Saran wrap (Dow) (the wrap is laid over and cut around the gel). Lay a piece of membrane the same size as the gel on top of it and on top of that layer two pieces of 3mm chromatography paper and a stack of dry absorbent paper towels about 10cm in depth. This is all held down by a 200g weight placed on top.

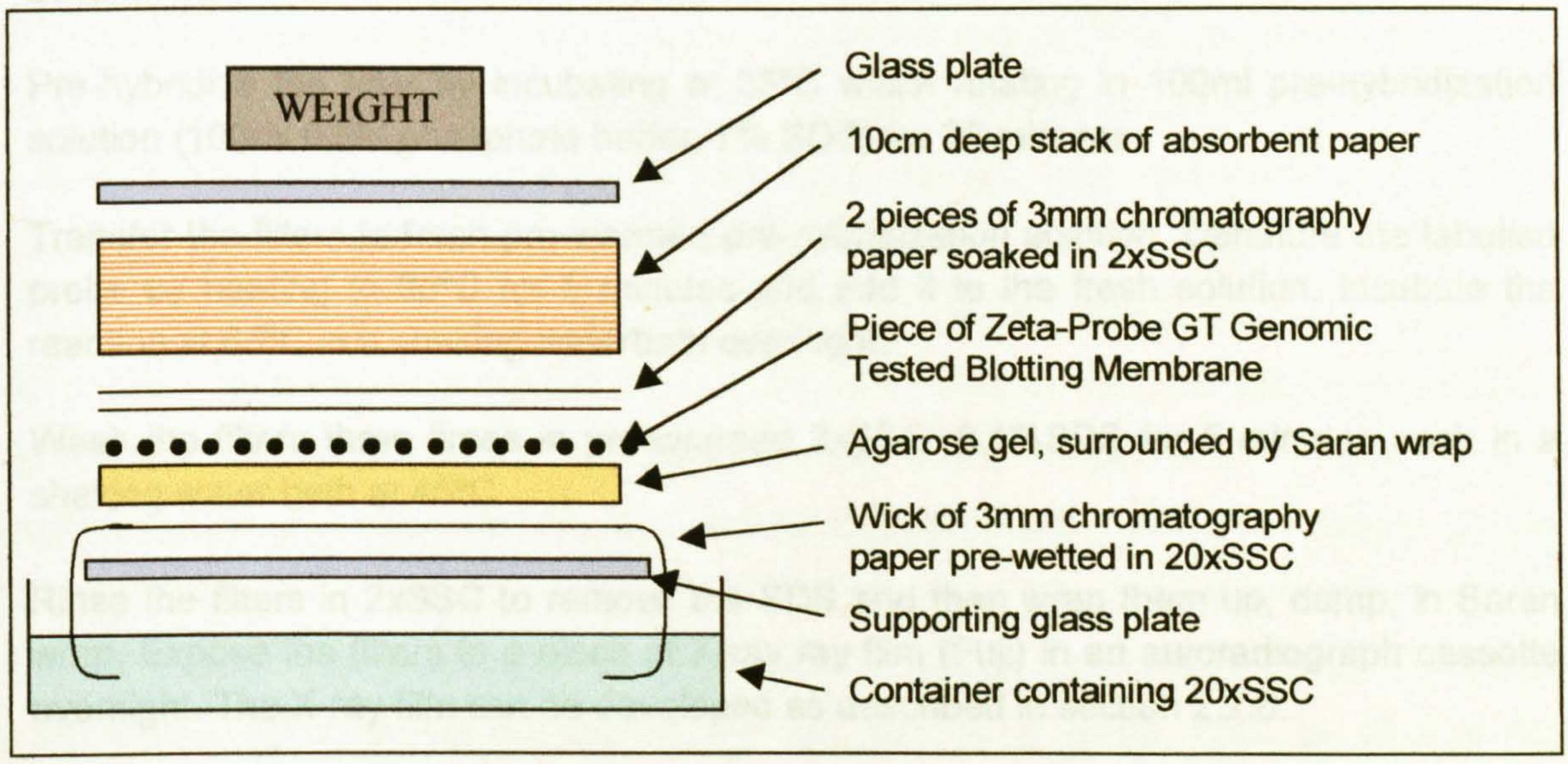


Figure 4.3: A diagram illustrating the set-up required for a Southern blot.

3. Leave the blot for 18 hours during which time the DNA is transferred to the membrane by capillary action. Once dismantled, rinse the membrane in 2xSSC for 1 minute and leave to dry at room temperature. Fix the DNA to the membrane by exposing it to UV, using the dry membrane crosslinking program on a GS Gene Linker (Bio-Rad).

4. Prepare the probe using the oligolabelling kit from Pharmacia Biotech. This kit works by random priming denatured probe material with random sequence hexanucleotides contained in the kit. Radioactively labelled nucleotides are then incorporated into the probe DNA as it is replicated. The oligos used to enrich the library are used as probe material and prepared according to the kit protocol:

    1. Mix equal quantities of each oligo making a total volume of not more than 34μl containing up 50ng DNA and heat denature by boiling in a waterbath for 5 minutes.

    2. Add 10μl reagent mix and 1μl Klenow fragment (5-10 units/μl) from the kit and 5μl ^{32}P α-dCTP (3000 Ci/mmol).

    3. Mix thoroughly and centrifuge very briefly to bring to contents of the tube to the bottom of the eppendorf. Incubate at 38°C for 4 hours.

    4. Stop the reaction by adding 20μl Nick-stop mix (0.9% blue dextran, 0.03% bromocresol purple, 20mM EDTA Na_2).

    Remove unincorporated labelled nucleotides and unlabelled probe DNA by spin column chromatography. Fill the barrel of a 1ml syringe, with a glass fibre plug inserted, with sephadex G50 and put into a plastic test tube. Equilibriate the column by spinning it in a centrifuge at 3000 rpm for 5 seconds and again with 50μl TE added. Add the probe reaction to the top of the column and spin it again for 5 seconds. Add 50μl of TE and spin the column for another 5 seconds, repeat this 5 times. The labelled probe accumulates in the tube below the column.

    5. Pre-hybridize the filter by incubating at 65°C whilst rotating in 100ml pre-hybridization solution (100ml 0.5M phosphate buffer, 7% SDS) for 30 minutes.

    6. Transfer the filters to fresh pre-warmed pre-hybridization solution. Denature the labelled probe by heating to 95°C for 5 minutes and add it to the fresh solution. Incubate the reaction at 65°C in a shaking waterbath overnight.

    7. Wash the filters three times in pre-warmed 2xSSC, 0.1%SDS for 5 minutes each in a shaking water bath at 45°C.

    8. Rinse the filters in 2xSSC to remove the SDS and then wrap them up, damp, in Saran wrap. Expose the filters to a piece of X-ray ray film (Fuji) in an autoradiograph cassette overnight. The X-ray film can be developed as described in section 2.2.8.

4.2.8 Ligation into a plasmid vector

The enriched amplified DNA was ligated into the pGEM-T vector (Promega) to construct libraries 1 and 2, and into the pNoTA/T7 shuttle vector of the Prime PCR Cloner Cloning System (5 Prime → 3 Prime, Inc.) for library 3, following the protocols provided with the ligation kits. In both cases, control ligations were set-up with components provided with the kits as recommended.

Ligations into pGEM-T vector depend on the activities of many thermostable DNA polymerases including Taq. These enzymes usually add a deoxyadenosine to the 3' ends of the amplified products. The pGEM-T vectors are prepared by the manufacturers by cutting the circular DNA molecule with EcoRV and adding a terminal thymidine to the 3' ends. This leaves the vector and PCR products with compatible ends for ligation. The ligation into the shuttle vector is blunt-ended and the insert fragments are exposed to a modification reaction to blunt and phosphorylate the PCR products. They can then be ligated into the vector.

The ligations into pGEM-T vector were carried out with an insert:vector ratio of 3:1, more insert than vector should encourage the incorporation of the insert. This required 35-40 ng of insert DNA. The ligation reactions were set-up with 1μl ligase buffer, 1μl (50 ng) pGEM-T vector, 1μl T4 ligase, the insert DNA solution and sterile distilled water to make the volume up to 10μl. The reactions were incubated at 15°C for three hours and stored at 4°C overnight. The ligation into pNoTA/T7 shuttle vector was done with an insert:vector ratio of 4:1 which required 222ng of cleaned PCR product. The modification reaction was set-up by mixing 2μl 10x prime PCR cloning reagent, 1μl prime PCR cloner nucleotide mix, 1μl 0.1M DTT solution with the appropriate volume of cleaned PCR product and enough sterile distilled water to make up the volume to 18.5μl. After mixing briefly, 1.5μl Prime PCR modification reagent was added. The reaction was vortexed gently and centrifuged briefly before incubating at 16°C for 15 minutes and then at 75°C for 15 minutes to terminate the reactions. The ligation reaction was set-up by adding 5μl sterile distilled water, 1μl 0.1M DTT solution, 2μl 10x prime efficiency ligation buffer and 1μl pNoTA/T7 vector DNA to 10μl of the modification reaction. After mixing this briefly 1μl of T4 DNA ligase enzyme was added. This was mixed by gentle vortexing and consolidated at the bottom of the tube by brief centrifugation. The reaction was incubated at 25°C for 30 minutes then terminated by heating to 65°C for 2 minutes.

4.2.9 Transformation of competent cells

The libraries were constructed by transforming *Epicurian coli* XL2-Blue MRF ultracompetent cells (Stratagene); cultures of these cells come in a kit ready prepared to be transformed. These cultures are very expensive so before using these cells the success of the ligation reaction was checked by transforming competent cultures of *E. coli* DH5α. The cells were prepared as for *E. coli* JM101 following the protocol given in section 2.2.5.2. Approximately 50ng of the ligation reactions were used to transform 200μl of competent cells following the protocol in section 2.2.5.3. 100μl of the transformed cells were plated onto an agar plate for blue-white screening as described in section 2.2.5.4. If white colonies were seen to grow on this plate, the ligation reactions were deemed successful and used to transform the ultracompetent cells. This was done following the protocol provided in the kit and 100μl of the transformed culture was plated on an agar plate for blue-white screening. If the agar plate was covered with many colonies, most of which were white ones, the transformation was successful and the remaining transformation reaction was plated out onto agar plates ready for colony picking.

Where positive control ligation reactions had been carried out using the kit protocols, competent *E. coli* DH5α cells were also transformed with these ligations as an indicator of the quality of the ligation reactions. For each transformation, control reactions were also done using approximately 50ng of the cleaned pGEM-T vector containing insert DNA produced when cloning the red squirrel mitochondrial control region (see section 2.2.6). This was a useful positive control to show that the competent cells were prepared properly and the transformation reactions carried out successfully.

4.2.10 Colony picking and storage

Glycerol stocks of white colonies were made in 96-well microtitre plates with lids (Nalge Nunc International). A glycerol mix of 50ml LB medium, 11.5ml 80% glycerol and 0.75ml 4mg/ml ampicillin was made up and 125μl was pipetted into each well of the microtitre plate. White colonies were picked using a p200 Gilson pipette by stabbing the pipette tip into the colony and inoculating the medium in a well by gently pipetting. The colonies in the plates were allowed to grow for a couple of hours at 37°C before storing in a freezer at -20°C. Long term storage was at -80°C. Library 1 consisted of 960 colonies, library 2 of 1152 and library 3 of 672.

4.2.11 Replication of colonies onto nylon filters

The stock colonies were replicated onto membranes for probing using a technique known as "hedgehogging", as illustrated in figure 4.4. This term refers to the tool used in replication which, when held upside down, could be said to bear some resemblance to a hedgehog. The DNA within the colonies is then fixed to the filters to prepare them for probing. The following method was used:

1. Pour 300ml melted agar (LB agar plus 15g/L bacto-agar) plus 15mg amplicillin into three large (9″ by 9″) agar plates. Allow to set and dry thoroughly in an oven.

2. Cut pieces of Zeta-Probe GT Genomic Tested Blotting Membrane (Bio-Rad) to the same size as the microtitre plates, cut the same number of pieces as microtitre plates to be replicated and number them to match the plates.

3. Lay the pieces of membrane on the large agar plates.

4. Fire the hedgehog by dipping it in ethanol and igniting it in a bunsen flame. Leave it for two minutes to cool after the flames have gone out.

5. Press the prongs of the hedgehog into the wells of the microtitre plate and then press them firmly straight onto the corresponding filter so the lay out of the colonies in the plates is exactly replicated onto the filter.

6. Repeat the process of firing the hedgehog, allowing it to cool and transferring colony medium onto the membranes until all the colonies in the microtitre plates are replicated onto the membranes.

7. Place the membranes on the surface of the agar and incubate the plates upside down at 37°C overnight to allow colony growth to develop on the membranes.

8. Record any unusual colony growth such as extra large colonies, faint colonies or colonies that have not grown at all, for reference after probing.

9. Soak some pieces of 3mm chromatography paper (Whatman) in 2xSSC, 5% SDS. Lift the filters off the agar and onto the Whatman paper, leave for 2 minutes.

10. Transfer the Whatman paper with the filters on to a microwave oven and heat on full power for 2 1/2 minutes (with a 240V bulb), just enough to dry the membranes thoroughly without burning. This fixes the DNA to the membranes.

11. Wet the membranes again in a tray containing 5xSSC, 0.1% SDS and rub off the remaining bacterial material on the membranes with a gloved finger.

12. Leave the membranes to dry at room temperature then wrap in Saran wrap and expose them face down to UV for 50 seconds in the GS Gene Linker (Bio-Rad) using the dry zeta-probe crosslinking program.

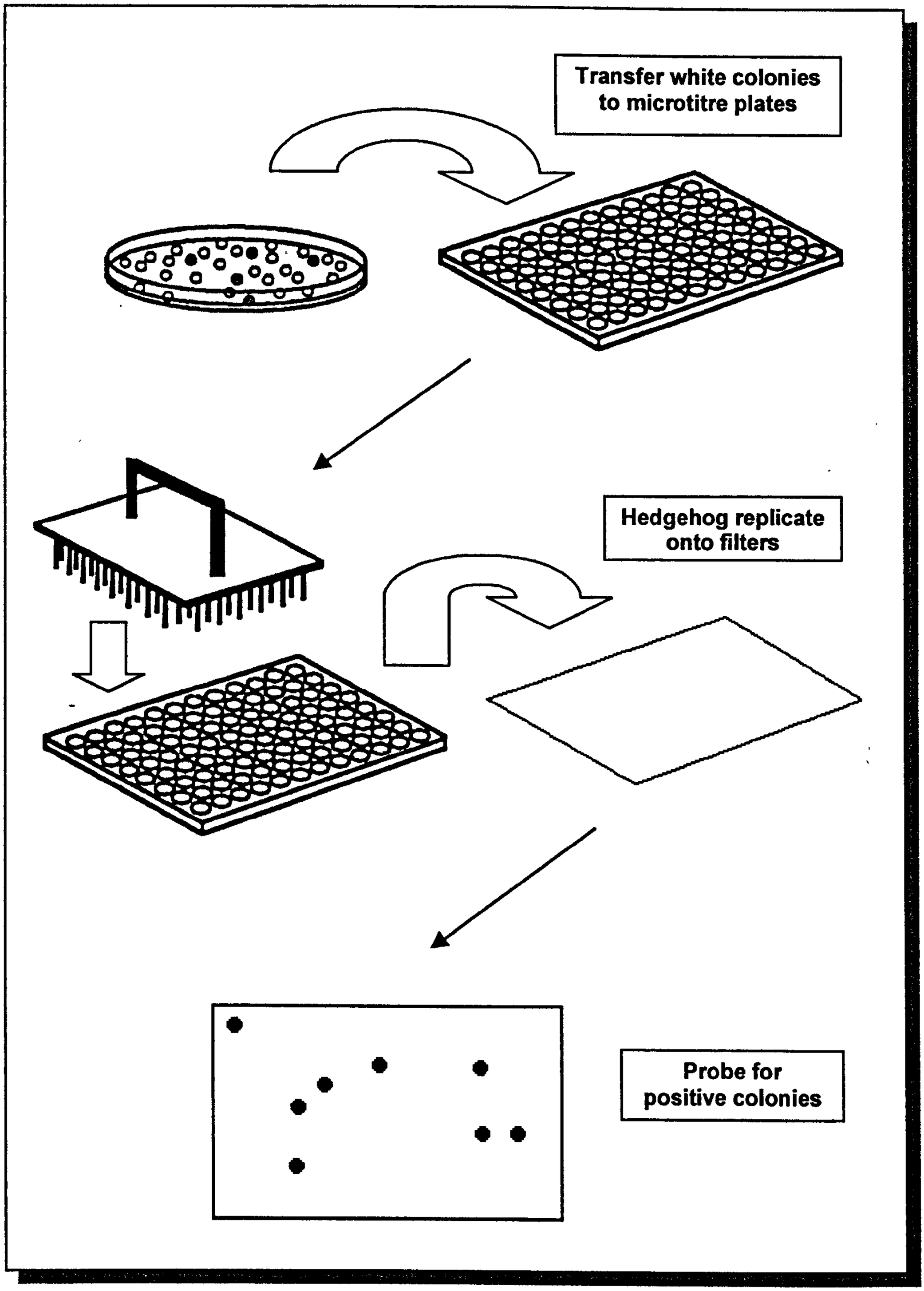


Figure 4.4: The stages of colony selection and storage after the construction of a library by cloning genomic fragments into a bacterial culture. See text sections 4.2.10 – 4.2.12.

4.2.12 Identification of inserts containing repeats

Inserts containing tandem repeat sequences are identified by probing the library filters with the oligo sequences used as target DNA in the enrichment steps. All the libraries were probed with a mixture of the oligos used in their construction; in addition, library 3 was probed with the $[GT/AC]_n$ tandem repeat sequences individually. The same methods were used as described in steps 4 - 8 of section 4.2.7 with slight variations:

50ng of oligo DNA was labelled using the oligolabelling kit (Pharmacia Biotech) as previously described, although, as the probe was used immediately, it was not necessary to add Nick-stop mix. The filters were pre-hybridized in a plastic box covered in pre-hybridization solution in a shaking water bath and the denatured probe was added directly to the solution surrounding the filters. The hybridization reaction was again allowed to proceed overnight at 65°C with constant agitation. The filters were washed 3 times, constantly agitating in the shaking water bath, using 150-160ml of the prewarmed solutions listed in table 4.4.

	Library:			
	1	**2 (1)**	**2 (2)**	**3**
wash 1	0.25M phosphate buffer, 1% SDS	2xSSC, 0.1% SDS	1xSSC, 0.1% SDS	1xSSC, 0.1% SDS
wash 2	2xSSC, 0.1% SDS	2xSSC, 0.1% SDS	1xSSC, 0.1% SDS	1xSSC, 0.1% SDS
wash 3	1xSSC, 0.1% SDS	1xSSC, 0.1% SDS	1xSSC, 0.1% SDS	1xSSC, 0.1% SDS
conditions	15 minutes at 65°C	5 minutes at 65°C	10 minutes at 65°C	10 minutes at 65°C

Table 4.4: The solutions used to wash the probed library filters from each of the three libraries. Library 2 was probed twice and the details of both sets of washes are given. All the filters were washed in enough of each solution to cover the membranes in a shaking waterbath.

The filters were then wrapped in Saran wrap and in a darkroom, placed face down on an X-ray film in an autoradiograph cassette. Before shutting the cassette, the film with the filters on top, was pre-flashed by setting off a camera flash held directly above. This exposes the X-ray film around the filters allowing the outlines of the filters to be seen when the film is developed. The locations of any positive colonies can then be accurately identified. Positive colonies containing tandem repeated DNA of the sequences probed for show up as dark circles on the autoradiograph in a position corresponding to the location of that colony growth on the Zeta-probe membranes.

After probing, the filters were stripped to remove the radioactive probe. The filters were laid out in a large plastic box with the DNA side upwards and completely visible. 0.8g of SDS was sprinkled evenly over the filters and 800ml of boiling distilled water was poured over them. The lid was then put on the box and the box was left, insulated in polystyrene blocks, until cold. The filters were rinsed in 2xSSC to remove the excess SDS and allowed to air dry. This process only removes the probes and the DNA remains bound to the membranes so they could be reprobed if necessary.

4.2.13 Storage of positive cultures

Larger glycerol stocks were made of positive colonies. Overnight cultures were grown up by innoculating 5ml LB media containing 250 μg ampicillin with 5μl of the glycerol stock in the microtitre plates. The cultures were grown up overnight at 37°C with constant agitation. Glycerol stocks were made up in eppendorfs with 1ml of overnight culture and 0.5ml 80% glycerol, they were stored frozen at -20°C.

4.2.14 PCR of the inserts

Stabs from the frozen glycerol stocks were streaked out onto sections of agar plates (containing ampicillin, poured as described in section 2.2.5.4 for blue-white screening) and grown upside down overnight at 37°C. This allowed individual colonies to be picked and used as template in a PCR reaction to amplify the genomic DNA inserted into the vector. The reactions were primed with

M13 for	(5' -GTAAAACGACGGCCAGT- 3')	(GibcoBRL)
M13 rev	(5' -CAGGAAACAGCTATGAC- 3')	(Promega)

designed to bind to the vector DNA either side of the insert site. These primers were applicable to both the vectors used in library construction. The reactions were set-up as described in table 4.5.

Reagent	Conc^{n}.	Quantity	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	2.5 µl	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	4 µl	0.2 mM
Magnesium chloride (Advanced Biotechnologies)	25mM	3 µl	3 mM
primer M13 for	10µM	2 µl	0.8 µM
primer M13 rev	10µM	2 µl	0.8 µM
Red Hot *Taq* (Advance Biotechnologies)	5 U/ µl	0.2 µl	1 U
template		single colony	
sterile distilled water		11.3 µl	

Table 4.5: The PCR reaction mix used to amplify the genomic inserts in the vectors of positive library bacterial cultures.

The reactions were run using the following program on either a PTC-100 or PTC-200 thermocycler as before.

Step	Temperature (ºC)	Time (minutes)
1	94	5
2	94	1
3	55	1
4	72	1 1/2
5	Go to step 2	29 more times
6	72	3

The reactions were run out on a 1.5% agarose gel with an appropriate size marker. The approximate size of each insert could then be estimated by comparison to the size marker. If the insert DNA was to be sequenced directly the clean, bright bands of amplified product were cut out of the gel and the DNA extracted using the QIAquick Gel Extraction Kit (Qiagen).

4.2.15 Sequencing of the insert DNA

In library 3, where many positives were indicated in the initial probing, the sequencing of inserts was prioritised. Tri- and tetra- repeats were preferred so all the positives indicated by probing with the (GT/AC) repeat oligo were not sequenced initially. Longer fragments were sequenced first and any inserts of 100bp or less in length were deemed too short to be of use. For the earlier libraries attempts were made to sequence all of the positive inserts.

4.2.15.1 Sequencing from cleaned plasmid template

Most of the inserts from libraries 1 and 2 were sequenced from the plasmid vector. The plasmid DNA containing the insert was harvested from each positive culture using a shortened version of the method described in section 2.2.6

1. **Harvest the bacteria**
   Inoculate 12ml of LB media containing 0.6mg ampicillin with the bacterial culture carrying the vector from the glycerol stock. Grow the culture overnight at 37°C agitating constantly.

2. Transfer the culture to a large Falcon tube and spin at 4000rpm in a benchtop centrifuge for 10 minutes at 4°C. Pour off the supernatant broth to leave a large pellet of bacterial culture.

3. **Alkaline lysis**
   Resuspend the bacterial pellet in 300µl solution 1 (50mM glucose, 25mM TrisHCl pH8, 10mM EDTA pH8), transfer the solution to a 1.5ml eppendorf and leave at room temperature for 5 minutes.

4. Add 300µl solution 2 (0.2M NaOH, 1% SDS), shake gently and incubate on ice for 5 minutes.

5. Add 300µl solution 3 (3M potassium, 5M acetate solution made by mixing 60ml 5M potassium acetate, 11.5ml glacial acetic acid and 28.5ml H_2O) shake well and leave on ice for 15 minutes.

6. Spin solution in a microfuge at 13000rpm for 15 minutes and transfer the supernatant to a fresh eppendorf.

7. **Purification of plasmid DNA**
   Perform two extractions, one with phenol:chloroform and one with just chloroform (see steps 2 & 3 of the DNA extraction protocol, section 2.2.2).

8. Precipitate the clean plasmid DNA by adding 1/10th volume 3M Sodium acetate and 2x volume cold 100% ethanol, incubate at -80°C for 30 minutes and spin at 13000rpm for 15 minutes. Remove the supernatant, rinse the pellet with 70% ethanol, spin again briefly, remove supernatant and dry the pellet thoroughly. Resuspend the DNA in 50µl TE.

The insert DNA was then sequenced manually using the primers SP6 and T7 or M13 for and M13 rev following the protocol given in section 2.2.7 and visualised on 6% polyacrylamide gels as described in section 2.2.8.

4.2.15.2 Direct manual sequencing of PCR products

All of the positives sequenced from library 3 and a few of those from library 2 were sequenced directly from the insert PCR products. The reactions were carried out using T7 sequencing mixes (Pharmacia Biotech).

The following protocol was used to sequence all the amplified PCR products.

1. **Primer annealing**
   The primer is annealed to the template DNA by a sudden drop in temperature, the primer "snap" anneals to the template sequence.

   For each template set-up the following reaction:

primer (10mM)	2μl
annealing buffer	2μl
5% Nonidet P-40	1.4μl
template	8.6μl

   Mix well by gentle pipetting, boil for 3 minutes in a boiling waterbath and immediately lower the tube into liquid nitrogen.

2. Set-up the other reactions; pipette 2.5μl each of the terminating reaction mixes "G", "A", "T" and "C" into four wells of a sequencing plate (Pharmacia). Prepare the labelling mix:

labelling mix A	3μl	
5% Nonidet P-40	0.6μl	
enzyme dilution buffer	1.1μl	
T7 polymerase enzyme	0.3μl	
[α-^{35}S]dATPαS	1μl	(0.37 Mbq)

3. **Sequencing reactions**
   Carry out the sequencing reactions as follows:

   1. Pre-warm the sequencing plate on a hot plate at 37°C.
   2. Pulse the template reaction in a microfuge.
   3. Labelling reaction:
      Add the 6μl of labelling mix to the template reaction and leave at room temperature for 30 seconds or more.
   4. Terminating reactions:
      Add 4.5μl of the labelling reaction to each termination mix in the sequencing plate and incubate at 37°C on the hot plate for 5 minutes.
   5. Add 6μl of stop solution to each reaction in the sequencing plate.

Once completed the reactions can be stored in a fridge at 4°C. The reactions were visualised by running through 6% polyacrylamide as described in section 2.2.8.

4.2.16 Primer design

Primers were designed with the aid of the program Oligo, ver. 4.0 (National Biosciences Inc.). This program searches a sequence for the best primer pairs. The primer sequences it suggested were usually adjusted by eye to improve them according to the requirements discussed in section 4.1.1.3. The oligos were manufactured by MWG Biotech and the Oligonucleotide Synthesising Unit at the University of Nottingham. The primers were tested by carrying out an amplification reaction using the plasmid DNA from which they had been designed as template, either the isolated plasmid DNA was used or the colony from which it was derived. The reaction conditions described in section 4.2.14 were appropriate for this reaction, only varying the annealing temperature to match that required by each primer set. If DNA in solution or liquid culture was used as template, the water content was reduced correspondingly.

4.2.17 Microsatellite amplification

The microsatellite loci isolated were amplified using protocols described in chapter 5. Initially the amplification reaction for each locus was tested with unlabelled primers and visualised on 4% metaphor agarose (Flowgen) with Ethidium bromide staining. Gels of metaphor agarose have a finer matrix than LE agarose so have a greater resolving power and can be used more effectively to visualise smaller PCR products. Metaphor agarose can be used as with LE agarose (see section 2.2.3) except that it requires more careful disolving in the TAE buffer. The metaphor agarose powder must be added extremely slowly to the buffer with constant mixing (preferably using a magnetic stirrer) to create an even suspension and then dissolved slowly with gentle heating in a microwave. The gel can then be poured and run as with LE agarose.

The primer pairs were then used to amplify the loci in a panel of 10 German samples to test for variability. Primer pairs amplifying loci that showed variation in these samples were applied to all the populations, as described in chapter 5.

4.2.18 Testing the loci

The presence of null alleles was tested for by carrying out exact tests on the 2x2 contingency tables of observed and expected numbers of heterozygotes and homozygotes. Each locus was tested on the data from the German Waldhäuser population sample of 27 individuals, the German population being the least likely to have reduced heterozygosity due to other factors. The program GENEPOP (v. 3.1) (Raymond and Rousset 1995a) was used to calculate the expected frequencies of heterozygotes and homozygotes and Fisher's exact tests were carried out using BIOMSTAT (v.3.2) (Applied Biostatistics Inc.).

Linkage was tested for in all the populations, German and Belgian. An apparent association may occur by chance in a small sample, therefore only testing one population may be misleading. If the same disequilibrium is consistently seen in all the populations then it is more likely to be due to real linkage. The probability of linkage disequilibrium between each pair of loci was also calculated using GENEPOP (v. 3.1), probabilities were obtained for each locus pairwise comparison in each population and a global exact test was performed on each locus pair across all populations.

4.3 RESULTS

4.3.1 Microsatellite isolation

Three libraries were constructed using the hybridisation selection method of enrichment as described by Armour *et al.* (1994). Library 1 was constructed by selecting for three tetra repeats. Five positive colonies were identified and on sequencing no repeat was found in two of the insert sequences. $[GGAT]_n$ repeats were found in the other three but only one was long enough to be used as a locus for amplification; this was the variable locus RSμ1.

The presence of only 5 positive colonies out of 960 suggests that the enrichment procedure had not worked as these are numbers that may be expected from an unenriched library (Armour *et al.* 1994). As has already been discussed, the isolation of microsatellite loci is a complicated multistage process and a mistake or failure of a reaction during any step can jeopardise the whole library. The methods used were examined carefully and consideration was given to how the process could be improved for future libraries. A few changes were made which included carrying out more careful size selection at each stage by extraction from an agarose gel and the use of a larger panel of oligo repeats as target DNA.

The most important size selection step is probably the one immediately prior to cloning after the second round of whole genome PCR. In constructing library 1, size selection was not carried out at this point, but instead the PCR reactions were cleaned using a PCR cleaning process to remove the remaining reagents. This would have allowed short molecules that were not long enough to contain a microsatellite locus to enter the ligation reaction. Shorter molecules may have been preferentially ligated, finally resulting in a library of short inserts containing very few repeat loci. Gel extraction was carried out in place of PCR cleaning during construction of subsequent libraries.

By increasing the panel of synthetic oligo repeat sequences, more loci may be selected and more loci should be successfully cloned. Four tetra- and six tri- repeats were used in the construction of library 2 and out of 1152 colonies, 36 probed positive in the first instance. No repeat was found in 10 of these positives on sequencing. This was either due to the presence of a short imperfect repeat, to which the probe bound, that was not identified in the sequence or to the hybridisation of the probe to random sequence. 12 of the positives were either not successfully amplified for sequencing or not sequenced cleanly. It can be difficult to sequence through tandem arrays of short sequence as the polymerase enzyme seems to find such motifs difficult to replicate. In some cases this meant it was difficult to obtain sequence on both sides of a repeat.

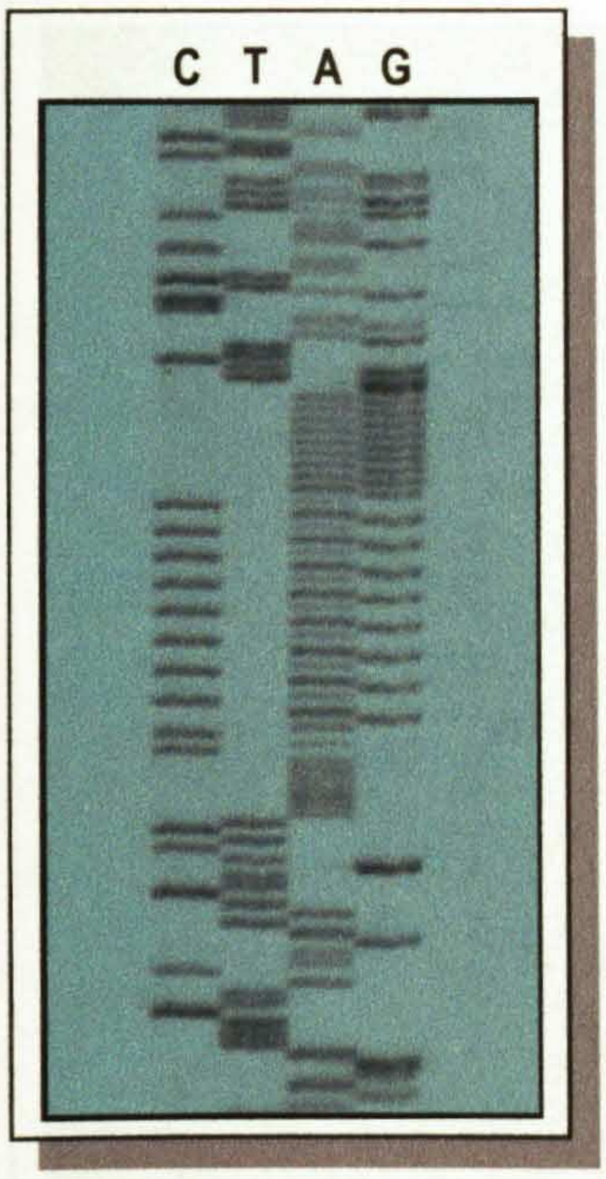


Figure 4.5: The sequence of the repeat motif contained within locus RSμ3.

Of the 14 positives containing repeats, 10 did not have enough flanking sequence on one side to design a primer. This is a common problem in microsatellite isolation by partial genomic library construction as the genome fragments are only a few hundred base pairs long so there is a reasonable probability that a repeat array will fall to one side or the other of the insert. Primers were designed flanking repeats found in four of the inserts but one locus repeatedly failed to amplify cleanly, despite the design of two different primer pairs. This region may have been replicated in the genome resulting in the messy amplification of many different sized products. The loci RSμ2, RSμ3 and RSμ4 were successfully amplified. Figure 4.5 shows the autoradiograph of a repeat sequence obtained by sequencing library clone N of library 2, this is the repeat is contained in locus RSμ3.

This library was more successful than the previous one with 3.1% of clones probing positive and 1.2% found to contain repeat sequences, although most of those did not have enough flanking sequence. This is more efficient than might be expected from an unenriched library. For example, May *et al.* (1997) isolated microsatellite loci from the Northern Idaho ground squirrel using an unenriched library and found microsatellite loci in only 0.04% of clones screened with 3 tetra- and 5 tri- repeats. However, Makova *et al.* (1998) found microsatellites in 1.8% of clones from an unenriched library of striped field mouse DNA but only 1.1% of these were for four tetra- and tri- repeats, the remaining were for the more common $[CA]_n$ repeat not probed for in this library. Even if the success of the library created by Makova *et. al.* is unusual, a much greater efficiency of microsatellite identification is expected from the enrichment process. Armour *et al.* (1994) found approximately 30% of their clones gave a positive signal on probing with a mixture of tetra- and tri- repeats, whereas only 3.1% of clones in library 2 were positive.

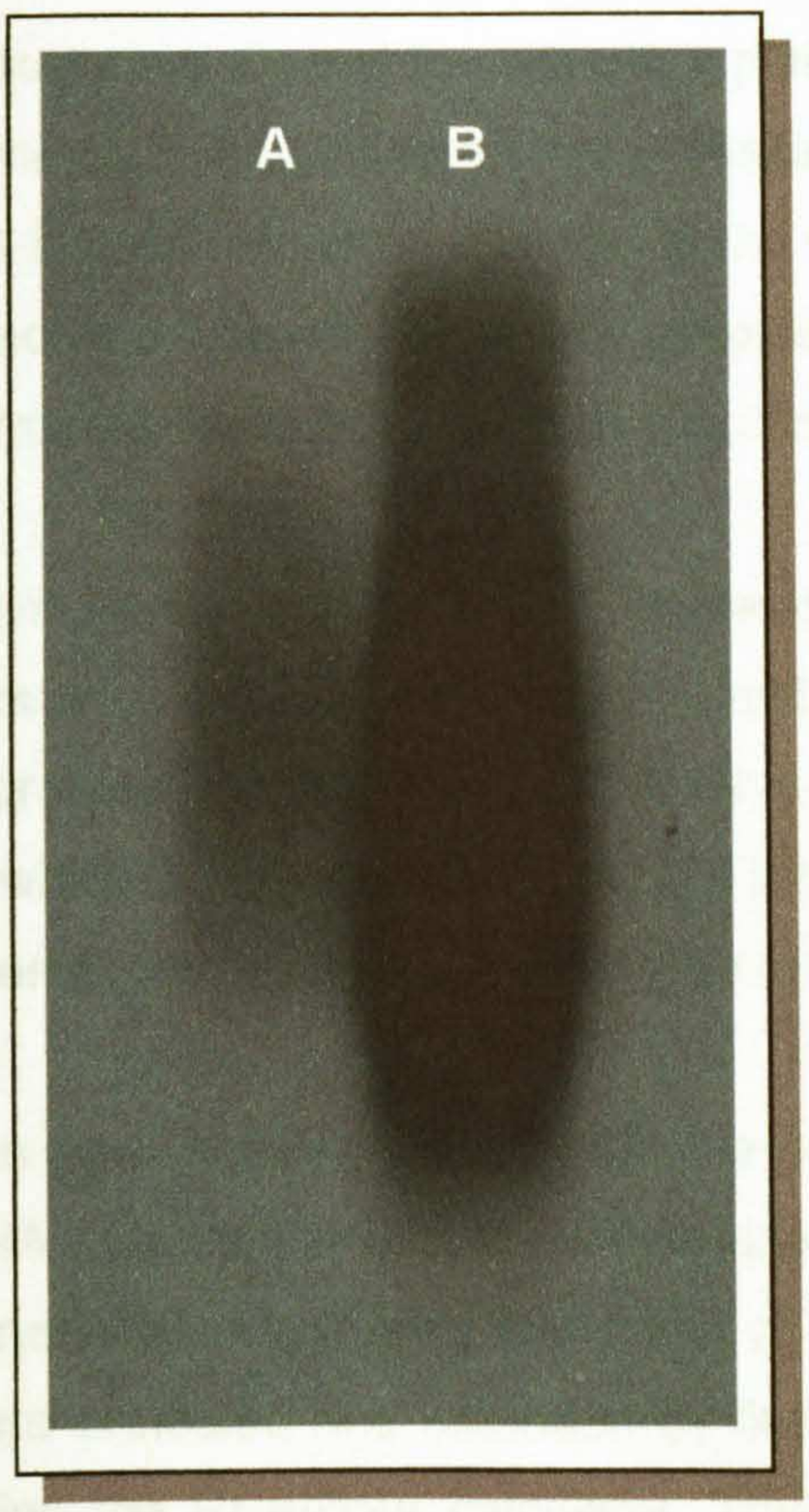


Figure 4.6: The autoradio-graph resulting from probing the pre- and post-hybridization selection whole genome PCR reactions. Lane A contains the pre-hybridization reaction and lane B is the post-hybridization reaction.

During the contruction of this library an attempt was made to assess the enrichment process by probing the products from the two whole genome PCR reactions with the mixture of oligos used as target sequences. The autoradiograph that resulted from the Southern blot is shown in figure 4.6. The smear in lane A is the pre-enrichment product and the smear in lane B is the post-enrichment product. The post-enrichment product is much more dense and darker indicating the presence of a greater concentration of repeated DNA. This suggested that the enrichment procedure had been successful. The subsequent library was indeed slightly enriched with these repeat sequences but not as much as would have been expected for the method. The Southern blot of the product DNAs was not sensitive enough to give any indication of the level of enrichment, it simply showed that one product contained more repeated DNA than the other. Had the enrichment not worked at all it probably would have shown that quite clearly, but it cannot be used as an indicator of the degree of enrichment.

Some new oligo repeat sequences were obtained by the laboratory and these were used to reprobe library 2. The repeats used were $[GTTC]_n$, $[TTTA]_n$, $[TTTG]_n$, $[TTTC]_n$, $[TATC]_n$, $[TTCC]_n$ and $[CCAT]_n$, most of which had not been used in the previous enrichment process but it was thought that they may identify other loci that may be present in the library by chance. 38 new clones were indicated along with 17 of the previous ones. Despite the higher stringency of the washes used in the second probing (1xSSC used instead of 2xSSC in the first two washes, the lower salt content results in higher stringency) no repeat was actually found in 13 of the positives when sequenced.

No sequence was obtained for 17 of the new positives as by now the glycerol stocks were quite old and this may have caused problems when trying to isolate clean template for sequencing. One of the microtitre plates had been spilt in the freezer resulting in the loss of some colonies and some of the positive colonies failed to grow when streaked out onto agar, this is probably due to the age of the stocks. Old stocks of bacteria containing a plasmid can reject the plasmid when the ampicillin in the media runs out. These stocks had been defrosted several times which would hasten their degeneration.

Eight of the new positives contained repeats but without enough surrounding flanking sequence so no new primers could be designed. Overall, library 2 probing revealed 22 microsatellite repeats in 1152 colonies (1.9%) and three loci were successfully amplified, two of which were variable. This left a total of 3 polymorphic loci; more were required so a third attempt was made at producing an enriched library.

The most notable difference in the construction of library 3 was the use of new oligo target DNA sequences. Some of those used in the previous library had been quite old and of low concentration resulting in a panel of target sequences varying greatly in quality. New oligos were annealed and extended by linear PCR and their concentrations accurately measured using the RNA/DNA calculator. This meant that equal quantities of each repeat could be measured accurately rather than estimated and used as targets in both the enrichment procedure and for probing for positive colonies. In producing an enriched library the process of selection by binding genomic fragments containing repeats to target DNA is the most crucial stage, so it is important to have good quality target sequences. Nine pairs of oligos were used in total (see table 4.2), including a $[GT/AC]_n$ repeat which is known to be very common in the genomes of mammals (Schlötterer 1998b).

A few other changes were made to the methods. A larger size range was included in the size selection in an attempt raise the average size of the inserts in the library. The upper limit was increased from 1000bp to 1300bp but the size of inserts in library 3 remained similar to that found in library 2 with most positives containing between 250 and 350bp of genomic DNA and the occasional insert of 500 or 600 bp. Shorter fragments may have been ligated into the vector with greater efficiency and vectors containing shorter inserts may have been preferentially taken up by the competent bacteria during cloning. Different linkers were used for the 3rd library, these were designed by U.H. Refseth at the University of Oslo (pers. com.) and were recommended as being more reliable than the previously used Saul linkers.

The ligation kit "Prime PCR Cloner" was used instead of the "pGEM-T vector" kit as this had given improved cloning results when used by another research group in the department. Included in the ligation reaction was DNA amplified both from the initial washes from the hybridisation selection filters and from denaturing the filters by boiling (section 4.2.6). A separate whole genome PCR reaction using the selected DNA removed from the filters by heat denaturing the water was successful, producing a smear of amplified DNA indicating that some selected molecules were left bound to the filters after the initial treatment with KOH. These molecules may include longer and more useful repeat loci that were bound to the target DNA more tightly. This extra step may have been very important in producing a more useful library.

A library of 672 clones was produced of which 84 probed positive when probed with a mixture of the target DNA sequences. Subsequent probing with the $[GT/AC]_n$ repeat showed that 58 of the positives probably contained this dinucleotide repeat motif. The remaining 26 were prioritised for sequencing as, although dinucleotide repeats are numerous, they are usually less variable and are harder to score. Time constraints meant that not all the inserts were sequenced, but of those that were, 7 contained no useful repeat, 5 included repeats but did not include enough flanking sequence to design primers and 6 contained loci for which primers were designed. Two of these inserts were replicates of another two, so only 4 different loci were amplified, loci RSμ5, RSμ6, RSμ7 and RSμ8. RSμ5 and RSμ7 are $[GT/AC]_n$ repeats, even though the clones they were in did not show as positives when probing was carried out with the dinucleotide repeat.

Library 3 contained 12.5% positives in total, yet without the $[GT/AC]_n$ repeat only 3.8% were positive, only a small improvement on library 2. Again, there has been some enrichment but not as much as could be hoped. This library was mostly enriched for $[GT/AC]_n$ repeats, several of the positives thought to contain tetra- or tri- repeats were actually found to contain this dinucleotide. The differences in success of the 2nd and 3rd libraries is probably due to the differences in frequencies of the dinucleotide repeat $[GT/AC]_n$ and the other tetra- and trinucleotide repeats selected for in the genome of red squirrels.

None of the libraries contained many incidences of replication where several clones contain the same fragment of DNA. This was both pleasing and surprising as both sets of whole genome PCR reactions could lead to some fragments being replicated many more times than the others and so being cloned several times. This potential weakness in the system did not show itself in this study.

4.3.2 The microsatellite loci

The primer sequences and details of all eight isolated loci are given in table 4.6. One locus, RSμ1, was isolated from library 1and was found to be variable when amplified in a panel of ten German red squirrels. Of the others, RSμ3, RSμ4, RSμ5 and RSμ6 were also found to be polymorphic. Out of the three compound and interrupted repeats only the compound repeat was polymorphic, this is consistent with previously described patterns of polymorphism and the observation that interruptions seem to stabilize arrays of repeats (Jarne and Lagoda 1996). However, it also generally thought that loci containing larger numbers of repeats are more variable, in this study a repeat of 10 [GT]s was variable whereas an array of 16 of the same dinucleotide motif was not.

4.3.3 Testing the loci

The five variable loci were tested for the presence of null alleles using Fisher's exact tests carried out on the observed and expected numbers of heterozygotes and homozygotes for each locus in the German sample of 27 individuals. The results are given in table 4.7. Three of the loci, RSμ1, RSμ3 and RSμ6, show lower heterozygosities than expected, RSμ3 only marginally, but none of the loci show a significant deviation from the heterozygosities expected under Hardy-Weinberg equilibrium. When null alleles are present they often have quite a dramatic effect on the heterozygosity resulting in a large difference between the observed and expected numbers, for example Neumann and Wetton (1996) found the observed heterozygosity was only 0.4 in contrast to an expected heterozygosity of 0.75 in one of the loci isolated from the house sparrow. The differences between the expected and observed heterozygosities at all the red squirrel loci are quite minimal in comparison. Therefore there is no evidence of null alleles at any of the loci.

Locus	no. of alleles	H_o (%)	H_e (%)	exact probability
RSμ1	7	70.37	77.92	*0.55*
RSμ3	7	48.15	48.43	*1.0*
RSμ4	8	96.3	82.67	*0.19*
RSμ5	7	48.15	43.26	*1.0*
RSμ6	4	55.56	58.91	*1.0*

Table 4.7: A table showing the results of Fishers exact tests to detect null alleles showing the observed and expected heterozygosities (H_o and H_e respectively) for each locus in the German sample, the number of alleles found at each locus and the exact probability of each test.

Locus	Repeat	PCR Primers	Size (bp)	Sequence
RSμ1	$[GGAT]_{13}$	f 5'- CTGGGTTCACTGACTTCTCC -3'	172	**CACTCTCAGAGGCCAAAGTC**TGCTGTCTTTTCACCAGACCATTACTAATACT CAGAGGATGAACTAGTAGNNNNNN$[GGAT]_{13}$GGAGTGGGTAGATAAATAAGAT AGAT**GGAGAAGTCAGTGAACCCAG**
		r 5'- CACTCTCAGAGGCCAAAGTC -3'		
RSμ2	$[CA]_{10}CGT[AC]_7$ $[ATCT]_{18}$	f 5'- CTTGGGATCTCTCTTCTC –3'	186	**CTTGGGATCTCTCTTCTC**CACCCTTTC$[TCCC]_4[TC]_7$TTT$[CA]_{10}$CGT$[AC]_7$[AT CT]$_{18}$**AATCTATCCAGCCTTGT**
		r 5'- ACAAGGCTGGATAGATT-3'		
RSμ3	$[GA]_9[GACA]_9$	f 5'- GCCAAAATCTAGCCCAAGAAG -3'	168	**GCCAAAATCTAGCCCAAGAAG**TTTCTGTGG$[GA]_9[GACA]_9$CANAAAATCTCTG TTCTATAGAACATTCTTTAGGAGATAAAGAAAGCAAGACAATGTCCTATT**GC TTCTTTCCCACACCTGAG**
		r 5'- CTCAGGTGTGGGAAAGAAGC -3'		
RSμ4	$[ATCC]_{12}$	f 5'- CAATCCTCCCATCCTGCTGC -3'	274	**CAATCCTCCCATCCTGCTGC**CCTCCATTCACCTATCCACCCATCCACCCATC CATCAGTCCTTCCATCCACCCATTCATCCTTCCATCTATCCCTCCATCCAAC CATTTATTCATTCACCCATCCATCCACCTGCCTATCCTCCCATCCCTCCATC CAACCATTTATTCATTCACCCATCCATCCACCTGCCTATCCTCCC$[ATCC]_{12}$ ATCTT**CCCACCTATCTGACTGCCT**
		r 5'- TAGGCAGTCAGATAGGTGGG -3'		
RSμ5	$[GT]_{10}$	f 5'- CCCAGTCTACATTAAAGGGC -3'	119	**CCCAGTCTACATTAAAGGGC**AGGGCATGGCAGGGGAGCATAAACTCCACGA $[GT]_{10}$GCATGCACTATAGTCATACACAATTA**CAGTCAATTATAGTGTATAGGC**
		r 5'- GCCTATACACTATAATTGACTG -3'		
RSμ6	$[GTT]_{10}$	f 5'- GGCATAGGGCACGTGAAG -3'	125	**GGCATAGGGCACGTGAAG**TATTTTACTAAAAGAACTTCTTTGTTTTTGTTT $[GTT]_{10}$GAAAGTCTCTTCTTAGTTCTCCTAG**CTTGGTATGAGATTGGCCC**
		r 5'- GGGCCAATCTCATACCAAG -3'		
RSμ7	$[GT]_{16}$	f 5'- CAGCAGGCCTCAAGGCAATG -3'	108	**CAGCAGGCCTCAAGGCAATG**TC$[GT]_{16}$GCATGTTACCAGGGATCGAACCCAGG GCCTTGCA**CATGCTGGGTAAACCCTCCA**
		r 5'- TGGAGGGTTTACCCAGCATG -3'		
RSμ8	$[GT]_9AT[GT]_4$ $[GA]_{14}$	f 5'- GCACAGTGAGGAGGCATATG -3'	120	**GCACAGTGAGGAGGCATATG**$[GT]_9AT[GT]_4[GA]_{14}$GCGAGGAGAAGACAGCAGA CTGACA**CTTTGGGACTGTTCTTGGTG**
		r 5'- CACCAAGAACAGTCCCAAAG -3'		

Table 4.6: The eight microsatellite loci isolated from the genome of *Sciurus vulgaris*. The table shows the repeat sequence, the PCR products designed to amplify the loci, the size of the PCR product and the full sequence of the PCR product. RSμ1, RSμ3, RSμ4, RSμ5 and RSμ6 were found to be variable in a panel of German red squirrels.

The results of the global exact tests for linkage disequilibrium over all populations are given in table 4.8. None of these comparisons showed a significant difference to the genotypic distributions that may be expected under a null hypothesis of no linkage. However, the Belgian population of Peerdsbos and the German Waldhäuser population showed significant linkage disequilibria when the loci RSμ3 and RSμ4 were tested (*p=0.0019* and *p=0.0211* respectively). After a Bonferroni correction, only the Peerdsbos result remains significant (the corrected type 1 probability being *0.005*). The global test for loci RSμ3 and RSμ4 does have a low probability but it is not significant (*p=0.1*). It is unlikely that these loci are closely linked as a significant result is only seen in one population. Some comparisons would be expected to be significant by chance when multiple tests are carried out and this is the likely explanation of this result.

Locus pair	χ^2	d.f.	probability
RSμ1 & RSμ3	10.236	20	*0.964*
RSμ1 & RSμ4	12.198	20	*0.909*
RSμ3 & RSμ4	28.420	20	*0.1*
RSμ1 & RSμ5	10.241	20	*0.964*
RSμ3 & RSμ5	25.038	20	*0.2*
RSμ4 & RSμ5	20.633	20	*0.419*
RSμ1 & RSμ6	4.010	14	*0.995*
RSμ3 & RSμ6	5.565	14	*0.976*
RSμ4 & RSμ6	6.482	14	*0.953*
RSμ5 & RSμ6	4.806	14	*0.988*

Table 4.8: The results of the global exact test on pairwise comparisons of all the loci across all populations, showing the resulting χ^2 value and the probability that the null hypothesis of no linkage is correct.

4.4 CONCLUSIONS

Eight microsatellite loci, including five variable ones, were isolated using the hybridisation selection method of library enrichment. This enrichment process did not prove very efficient in this study for isolating tetra- and trinucleotide repeat microsatellite loci, several libraries were required to isolate five different loci (the other three loci isolated contain dinucleotide repeats). Library 3 was more successful, being reasonably enriched with dinucleotide repeats but as these are far more common in the genome this was not surprising. Lower success rates were observed than could have been hoped for when using an enrichment method but the loci isolated should be sufficient for population studies.

The variable loci fit the previously described patterns of polymorphism in different types of loci with the dinucleotide and compound repeats being the least variable and the tetranucleotide repeats the most. All have heterozygosities in the more variable German population of around 50% or above. None of the loci showed any evidence of null alleles when the expected and observed heterozygosities were compared and there is no evidence for any linkage between loci. It is impossible to completely exclude the possibility of either null alleles or linkage at these loci without the examination of many family groups, unfortunately, they were not available for this study.

CHAPTER FIVE:

THE MICROSATELLITE ANALYSIS OF THE STUDY POPULATIONS

5.1 INTRODUCTION

Microsatellites have been enthusiastically adopted by researchers as a useful molecular marker. The levels of variability found at the loci led to a great deal of initial optimism as to their potential usefulness for answering a wide-range of phylogenetic questions, from parentage assignment to species comparisons (McDonald and Potts 1997). Microsatellite loci have indeed been proved useful in population genetic analyses and in parentage studies (for example, in the elk (Talbot *et al.* 1996) and in humans (Alford *et al.* 1994)). However, cross-species amplification is not always reliable, limiting their use for species comparisons, and the nature of the step-wise mutation processes makes their use for evolutionary tree construction dubious, as alleles of the same size do not necessarily share the same ancestral history (Schlötterer and Pemberton 1994).

Microsatellites have been used with great success in population genetic investigations into the structure, relatedness and variability of populations. Conservation geneticists have found them useful because the use of PCR technology means amplification products can be obtained from very poor tissue samples, such as hair, feather and faeces, allowing non-invasive sampling techniques to be used (see Kohn and Wayne (1997) for a review of the development of "molecular scatology"). The high levels of polymorphism found at microsatellite loci also make them a useful tool for conservation genetics. Small populations found to be lacking in variability at other markers often show sufficient diversity at microsatellite loci to facilitate population level analysis. Paetkau and Strobeck (1994) used microsatellite loci isolated from the black bear (*Ursus americanus*) to compare the levels of diversity found in populations occupying three Canadian national parks. Bears tend to have inherently low levels of genetic variation due to factors such as low population densities and low effective population sizes and previous mitochondrial studies had found that variability of the mitochondrial genome was too low to distinguish between even widely distributed populations of North American bears (Paetkau and Strobeck 1994). Microsatellites showed enough variation for differences in the populations to be identified and it was found that one of the three populations was genetically depauperate, although there was no evidence that the health of the population was being adversely affected.

This chapter describes the application of the red squirrel microsatellite loci to the study populations in northern Belgium and Germany. The levels of variation found in the populations were compared and the effects of habitat fragmentation on the Belgian red squirrel populations ascertained.

5.1.1 Using microsatellites

Microsatellites are used in population studies by employing the polymerase chain reaction (PCR) to amplify each locus in each sample. Enough DNA is quickly accumulated to be visualised on a polyacrylamide gel and data as to allelic diversity and heterozygosity is easily obtained.

5.1.1.1 The polymerase chain reaction

The polymerase chain reaction is a remarkable technique both in its simplicity and its effectiveness. It employs a DNA polymerase enzyme to replicate a short stretch of selected DNA sequence exponentially until there are millions of copies. In theory millions of copies can be obtained of a single template DNA strand enabling the DNA to be visualised, sequenced and manipulated. PCR is used in microsatellite analysis to quickly amplify the locus under consideration and quickly obtain enough DNA to be visualised and sized on a gel. The principles and applications of PCR are reviewed by Newton and Graham (1997) and Hoelzel and Green (1998).

The amplification process itself involves cycling all the reagents, mixed together in one tube, through three temperatures. First the template is denatured by heating to 94°C. This temperature is usually sufficient, but non-denatured template is a common cause of PCR failure, so a temperature of 98°C may be required. Care must be taken not to reduce the activity of the polymerase enzyme by overheating. The annealing step follows, where the primer anneals to the template DNA, and finally the extension step where the new DNA strand is synthesised. The polymerase enzyme works best between 70-80°C but as it synthesises at a rate of at least 100 nucleotides per second and is active at temperatures outside its optimum range, the extension step can often be omitted when amplifying short products such as microsatellites. This also can reduce the amplification of unintended products.

Although the reaction itself is simple, it can entail complex biochemical interactions and is not always guaranteed to be successful. It is usually necessary to optimise the conditions in which the reaction takes place to obtain the maximum amount of clean product DNA. The quantity of reagents in the reaction can be adjusted. For example, magnesium chloride is an important cofactor for heat stable polymerases, so reducing its concentration also reduces the quantity of product obtained, but increasing it too much leads to non-specific amplification. Therefore, the quantity of magnesium chloride represents a trade off which depends on the stability of the particular reaction. The nucleotides in the reaction mix bind

magnesium ions so any change in the concentration of one may need to be compensated by an alteration in the concentration of the other. It is also important to consider the concentration of template DNA as too much can inhibit the reaction.

Cobb and Clarkson (1994) described a simple procedure for optimising the reagent concentrations in a PCR reaction. They considered there to be four reagents whose concentrations should be varied to optimise the reaction. If three different concentrations were tried for each reagent in individual reactions that would require 81 separate reactions. They applied Taguchi theories, often used in industrial processing, to suggest that a set of nine different PCR reactions would cover the essential combinations and help establish the optimal conditions for each reaction. This methodology can be useful in optimising PCR reactions.

Raising the annealing temperature can reduce the number of spurious products but too high a temperature will result in the primer binding with lower efficiency. The actual temperature at which the primer anneals to the template DNA with maximum efficiency (T_m) is dictated by many factors including primer length, GC content of the region, and the order of the nucleotides, but a temperature of T_m - 5°C is usually taken as a starting point. In general the higher the annealing temperature, the greater the binding specificity resulting in fewer undesired products (Hoelzel and Green 1998). The closer the annealing temperature to the actual one for the particular primer the greater the efficiency and accuracy of amplification.

Other methods to improve the accuracy of the PCR reactions include the addition of certain additives such as non-ionic detergents, formamide and DMSO, and the use of methods such as the hot-start. In this procedure the DNA polymerase enzyme is not added until the initial prolonged denaturation step. This prevents the reaction starting until a high temperature is reached thereby reducing the production of artefacts caused by early mispriming. This can be done artificially by covering the reaction with wax and putting the enzyme on top of it. When the reaction heats up the wax melts and the enzyme drops in.

5.1.1.2 Visualisation techniques

The amplified microsatellites are best separated on a thin denaturing polyacrylamide gel as used for sequencing, this matrix provides the required resolving power to distinguish alleles that may only differ in size by 2 bp. Some laboratories then silver stain the gels to visualise the DNA; this does provide a cheap and sensitive alternative to radioactive labelling but it is not sensitive enough to detect a sequencing ladder, which is used to size the alleles (Schlötterer 1998b), and the delicacy of the gel can make the handling required in the process difficult.

The most commonly used methods involve radioactive labelling, either by the incorporation of radioactively labelled nucleotides into the PCR product during amplification, or by the end-labelling of one of the PCR primers. The use of radiation is hazardous, but primer end-labelling gives clean, straightforward results. If only one primer is labelled only one strand of the double-stranded PCR product will be visualised. This reduces the number of bands to be scored and the number of stutter bands which may cause confusion. A homozygote is represented with one band and a heterozygote with two.

As automated DNA sequencers become more accessible, the application of fluorescence-based technologies to the detection of microsatellites is becoming more popular. The amplified PCR products are labelled with fluorescent dyes by incorporation into the product or by end-labelling a primer. The DNA is electrophoresed through a matrix passing a laser which activates the label. The size of the product is calculated automatically by an attached computer.

This method has many advantages over radioactive labelling, aside from the obvious ones of safety and time saving. Radioactive labelling often requires multiple gel exposures to obtain autoradiographs clearly showing all the reactions, the greater linear range of signal intensity from fluorescent labels means that signals of greatly varying intensity can be scored at the same time with improved accuracy (Ziegle *et al.* 1992). Omitting the steps of X-ray film development and scoring greatly reduces the time spent manipulating the data and making database entries (David and Menotti-Raymond 1998). The greatest disadvantage of this method is expense, the automated detection machines are extremely expensive and the cost of the kits and custom dye-labelled oligonucleotides is also very high (David and Menotti-Raymond 1998). The use of automated fluorescent-based detection systems in microsatellite analysis is reviewed by David and Menotti-Raymond (1998).

5.1.1.3 Scoring the data and avoiding errors

Ideally, all samples in a study would have each locus amplified and visualised repeatedly to ensure that the genotypic assignment was correct. In a large population study this is not feasible, but a subset of samples can be rerun to check that there is consistency and that samples have not been muddled at any stage. Another source of error is in the allelic assignment when examining the images of the microsatellite amplification products. If the PCR reactions have been satisfactorily optimised, most loci will produce an image of one or two clearly defined bands per individual.

Some loci, usually dinucleotide repeats, produce a ladder of bands when amplified, the extra bands are "stutter bands" (Schlötterer 1998b). These are thought to result from slip-strand mispairing (see section 1.5.2.1) occurring during the PCR reactions (Hauge and Litt 1993). A classic stutter pattern, produced when a heterozygote is amplified, consists of at least three bands, the darkest being the second from top and the next darkest being the top band. The top band is the longest of the pair of alleles, the next band (the darkest one) is both the other real allele and a stutter product of the first allele. The other bands below these are shorter versions of both alleles, resulting from slippage events. A homozygote will produce a pattern with the darkest band at the top, which is the true allele, and other bands below being shorter stutter products. Unfortunately, not all stutter band products are easy to interpret according to these rules, but it is usually possible to develop an "eye" for how each locus amplifies; consistency across samples for each allele is the most important consideration (microsatellite internet newsgroup, pers. com.).

5.1.2 Population genetic analysis of microsatellite data

Microsatellites are found in the nuclear genome, therefore the data produced are diploid in nature with two alleles being present at each locus. This means that data regarding both allelic diversity and heterozygosity can be gathered and analysed. Most population studies involve the analysis of a sample of individuals found in the population so the measure of allelic diversity (a straight count of the number of alleles in the sample) cannot be considered to be a reliable indicator of the variation present in the population as it is heavily dependent on the sample size. In the case of the Belgian fragment populations of Brede Zijpe (BZ), Gasthuisbos (GH), Kegelslei (KE), Luisbos (L), Tallaarthof (T) and Antwerp water works 1, 2 and 3 (AWW1-3) investigated here, all the individuals in the populations (except one individual from each of Brede Zijpe and Antwerp water works areas 1 and 2) have been sampled so the count of alleles can be used as an accurate measure of the variation.

Another measure of variation in populations is heterozygosity. This refers to the proportion of individuals in the population that are heterozygous at the locus in question; the average heterozygosity is the average proportion of individuals heterozygous across several loci. The observed heterozygosity is given by a count of the number of heterozygotes in a population at each locus.

In a population with allelic frequencies as expected under Hardy-Weinberg equilibrium (see below) the expected proportion of heterozygotes (H_e) in the population is equal to:

$$He = 1 - \sum_{i=1}^{m} x_i^{\,2} \qquad \text{where } x_i \text{ is the frequency of the } i\text{th allele and } m \text{ is the number of alleles.} \qquad (1)$$

Nei defines this quantity as "gene diversity" (Nei 1987, p177) but it is also referred to as the "expected heterozygosity". It can be compared with the observed heterozygosity at each locus to identify deviations from Hardy-Weinberg expectation or it can be used as a measure of genetic diversity in its own right. As a measure of genetic diversity it is an indicator of allelic diversity that takes the allele frequencies and sample size into account. It reflects both the number of alleles and evenness of their distributions, low frequency alleles make very little contribution to the measure of gene diversity.

Gene diversity cannot be calculated accurately without data for all the individuals in the population. When this information is not available, an unbiased estimation of the population gene diversity is calculated from sample data as: (Nei 1987, p178)

$$\hat{H}e = 2n(1 - \sum \hat{x}_i^{\,2})/(2n-1) \qquad \text{where } n \text{ is the number of individuals sampled} \qquad (2)$$

The data analysis packages used in this study, such as GENEPOP and BOTTLENECK, automatically calculate this estimation of gene diversity (equation 2), as would be appropriate for most population studies. In this study the gene diversity for the Belgian fragment populations can be calculated accurately using equation 1 rather than estimated.

5.1.2.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg law is central to population genetics. The law or principle describes a relationship between allele and genotype frequencies that was independently demonstrated by Hardy and Weinberg in 1908 (Falconer 1981; Nei 1987). The law states that in a random mating population with no selection, mutation or migration, the allele frequencies and the genotype frequencies are constant from generation to generation and that there is a simple relationship between them (Guo and Thompson 1992).

In its simplest form, where there are only two alleles at a locus, the relationship between the allele and genotype frequencies is (Falconer 1981):

	alleles:		genotypes:		
	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
frequencies:	p	q	p^2	$2pq$	q^2

This relationship can be expanded for multiple alleles. When the frequency of the genotype A_iA_j is given by X_{ij}, then the frequency of the ith allele (xi) is given by:

$$x_i = X_{ii} + \frac{1}{2}\sum_{j \neq i} X_{ij} \qquad (3)$$

where $\Sigma j \neq i$ indicates the summation of all the genotypes including allele i except the genotype A_iA_i (where $j=i$), which is already included in the calculation.

The frequencies of the genotypes A_iA_i (homozygotes) and A_iA_j (heterozygotes) are given by:

$$X_{ii} = x_i^2 \quad \text{and} \quad X_{ij} = 2x_i x_j \qquad (4a\&b)$$

These frequencies of alleles and genotypes are called Hardy-Weinberg proportions. Populations which display the described relationship between the frequencies are in "Hardy-Weinberg equilibrium" and it can be shown that, in the absence of mutation, migration and selection, these frequencies are maintained from generation to generation (for a complete demonstration of the proof of this rule see Falconer 1981, p8-9). For an autosomal locus, a single generation of random mating in a population is enough to establish Hardy-Weinberg equilibrium (Nei 1987).

Deviations from Hardy-Weinberg equilibrium can be caused by a number of processes that result in changes in allele frequencies; a full study of each process is given in Falconer (1981). Non-random mating where individuals preferentially mate with individuals expressing the same genotype as themselves (assortative mating) or with individuals of a different genotype (disassortative mating) will lead to increase in the frequencies of homozygotes and

heterozygotes respectively, ultimately leading to changes in allele frequencies. The process of selection, which results in increased frequencies of alleles if they confer increased fitness to individuals carrying them, will also alter allele frequencies. As microsatellites are generally accepted as being neutral, selection is unlikely to affect the frequencies of microsatellite markers. It is only possible for the frequencies of neutral alleles to be affected by selection or non-random mating if they are linked to loci being influenced by these processes, this is referred to as "hitch-hiking".

Migration and mutation both change allele frequencies in populations by introducing new alleles. Mutation and selection can have counteracting influences. Recurrent mutation may repeatedly introduce an allele to a population; if that allele is slightly deleterious, selection will tend to reduce its frequency. When this is the case, an equilibrium may be reached where the counteracting pressures of mutation and selection result in the maintenance of the frequency of the allele at a constant level. This is mutation-selection equilibrium.

Small populations, such as those of interest to these studies, are more likely to be influenced by the random sampling effects of drift, as described in section 1.3.2. The Hardy-Weinberg law applies to a population of infinite size or, in practice, a large population in which drift has no noticeable effect. Allele frequencies in small populations will fluctuate at random due to the random sampling of gametes in each generation. Unlike the other processes described, genetic drift is unpredictable and can result in an increase or decrease of allele frequencies. However, on average, it will lead to the loss of rare alleles and the fixation of common ones.

Testing for deviations from Hardy-Weinberg equilibrium can be carried out by comparing the observed genotype frequencies with those expected if the population was in equilibrium. The expected genotype frequencies are calculated from the observed allele frequencies using equations 3 and 4a&b. The expected and observed frequencies were traditionally compared using goodness-of-fit tests such as the χ^2 test but these tests can be unreliable when the sample sizes are small and/or some of the cell frequencies are small (Guo and Thompson 1992). In these cases an exact test is more reliable but until recently a lack of computing power limited the use of exact tests. Now it is possible to calculate exact probabilities for large contingency tables, such as those for allele frequencies at microsatellite loci that typically show large numbers of alleles. If the allele number is large the probability can be estimated using a Markov chain algorithm (Guo and Thompson 1992). Such an exact test of Hardy-Weinberg equilibrium can be carried out by the computer package GENEPOP (v. 3.1) (Raymond and Rousset 1995a) and the test was applied to all the populations in this study.

5.1.2.2 The expected allelic diversity

An unusual feature of this study is the total sampling of the fragment populations under investigation. This means that both the sizes of these populations and the actual allelic diversities are known. Ewens (1972) derived a formula for calculating the expected number of alleles ($E(k)$)in a population at mutation-drift equilibrium with loci mutating under the infinite alleles model. It is dependent on the effective population size (Ne) and the mutation rate (μ) and is calculated as follows:

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots\ldots\ldots + \frac{\theta}{\theta+2Ne-1} \qquad \text{where } \theta = 4Ne\mu \qquad (5)$$

As the population size for the fragment populations is known, this can be used to calculate the expected number of alleles for given mutation rates, although the effective population size is likely to be smaller. Microsatellites have been reported to mutate at rates ranging from 10^{-2} to 10^{-6} (Ellegren *et al.* 1995; McDonald and Potts 1997; Schlötterer 1998b) so the expected number of alleles can be calculated for each population at each of these two extremes of mutation rate. It should then be possible to see whether there is a different number of alleles in each of these populations than is expected for its size.

Kimura and Ohta (1978) developed a similar but much more complex formula for the expected number of alleles under the stepwise mutation model. Unfortunately this is too complicated to apply in this study as it involves several double and triple integrals in the generation of an equilibrium distribution of allele frequencies (Estoup *et al.* 1995). However, the expected number of alleles under the SMM is much lower than under the IAM so this can be considered when looking at the allelic diversity of the microsatellite loci (Kimura and Ohta 1978).

5.1.2.3 The effects of a bottleneck

When a population experiences a dramatic reduction in size (a bottleneck) it will lose much of its genetic variation with the individuals lost from the population. Initially, alleles are suddenly removed from the gene pool so there is a sudden and dramatic reduction in allelic diversity. The amount of diversity lost will depend on the severity of the bottleneck. Gene diversity and observed heterozygosity are reduced more gradually over the following generations at a rate dependent on the size of the resultant population. This means that, for a period immediately after a bottleneck, the population will have a transient excess of gene diversity. Conversely, a recent population expansion will result in a transient deficiency of gene diversity (Maruyama and Fuerst 1984; Maruyama and Fuerst 1985; Luikart *et al.* 1998). As the heterozygosity and

gene diversity fall in the generations following a bottleneck, the excess slowly shrinks until both allelic and gene diversities stabilise at equilibrium levels. The allelic diversity of a closed population is maintained by a balance between genetic drift, which tends to reduce allelic diversity, and mutation, tending to increase it; when this situation is stable the population is in mutation-drift equilibrium.

The most direct evidence for the historical occurrence of a bottleneck is a reduction in allelic diversity, proved by the direct comparison of the allelic diversity in the pre- and post-bottleneck populations, but in practice it is rarely possible to make such direct comparisons (but see Bouzat *et al.* 1998, reviewed in section 1.3.1). Most studies rely on comparisons with other populations thought not to have been affected by a bottleneck. For example, such comparisons have led to the conclusion that bottlenecks have occurred in populations of cheetahs (O'Brien *et al.* 1985) and humans (Sajantila *et al.* 1996). The same kind of comparison is made in this study which compares fragmented Belgian populations with a large German population thought not to have been reduced in size. However, comparisons of allele number are often not possible as the measure of allelic diversity is dependent on sample size, rendering cross population comparisons unreliable. Instead, the heterozygosities and gene diversities of the populations are used to show evidence of a reduction in population size but, if the bottleneck was recent, these measures may not yet show an effect and if the population rapidly recovered in size, they may never be affected by a bottleneck. Conclusions that a bottleneck occurred that are based solely on a comparison with a different population, sometimes species, may not be valid when nothing is known about the history of the populations. Indeed, it has been suggested that the role of bottlenecks in influencing the variability of populations has been overemphasised as other processes may lead to reductions in genetic variation (Cornuet and Luikart 1996).

Rather than using evidence from comparative studies, it may be possible to detect a bottleneck by looking at the patterns of allelic and gene diversity within the populations. The transient excess of gene diversity should be detectable for "a given window of time" (Cornuet and Luikart 1996) after a bottleneck has occurred. Cornuet and Luikart (1996) proposed a statistical test that compares the gene diversity that would be expected from the number of alleles found in the population with that observed.

At least five loci must be examined as the test of significance relies on there being a gene diversity excess at the majority of a number of loci. If a population is at equilibrium, the gene diversity at individual loci will still fluctuate around that expected due to the effects of drift and mutation, but at equilibrium around 50% of the loci would be expected to be in excess and

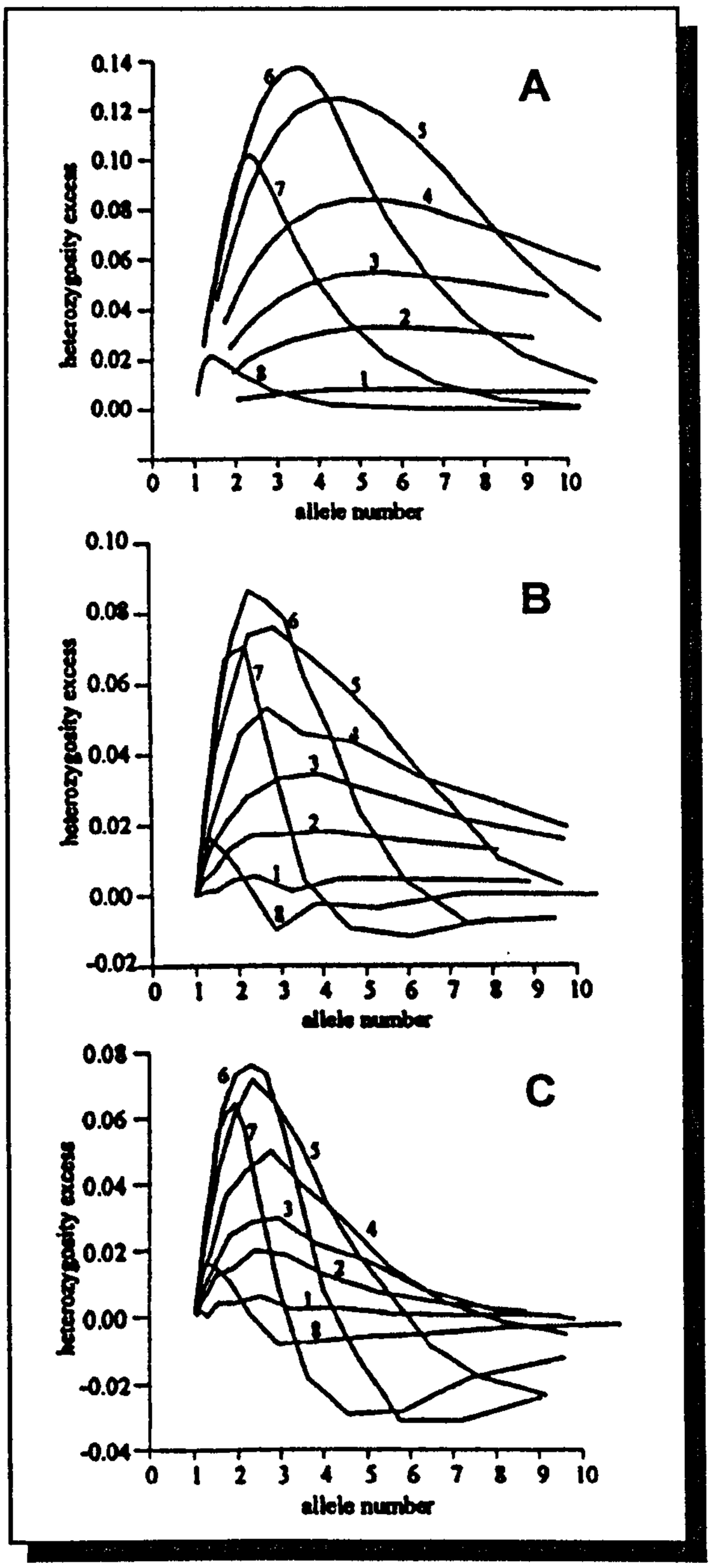


Figure 5.1: The theoretical relationship between gene diversity excess and the number of alleles in a population at different times after a bottleneck, when the loci evolve under **(A)** the IAM, **(B)** the TPM (a mixed model with 90% one-step mutations and 10% multi-step mutations), and **(C)** a strict SMM. Each line is the situation after a different time, in units of $2Ne$ generations, after the bottleneck:

1:0.005 (x$2Ne$) **2**:0.025(x$2Ne$) **3**:0.05(x$2Ne$)
4:0.1 (x$2Ne$) **5**:0.25 (x$2Ne$) **6**:0.5 (x$2Ne$)
7:1.0 (x$2Ne$) **8**:2.5 (x$2Ne$)

(reproduced from Cornuet and Luikart, 1996)

50% to show a deficiency of diversity. If most of the loci show a gene diversity excess this is taken as evidence of a recent bottleneck; the more loci considered, the greater the power of the test.

The expected gene diversity in a population depends on the mutation model of the loci considered. Figure 5.1 shows the theoretical response of gene diversity in a population reduced in size by a factor of 100, when the loci affected evolve under (A) the infinite alleles model (IAM), (B) the two-phase model (TPM) and (C) a strict stepwise mutation model (SMM) (from Cornuet and Luikart 1996). The shape of the curves is the same in all cases, as gene diversity can be expected to show a transient excess after a bottleneck under all the models. The infinite alleles model and the stepwise mutation model represent two extremes and the main difference between them is that the gene diversity excess is lower under the SMM (as the expected gene diversity is higher). The two-phase model is probably the most realistic representation of the mutation process experienced by microsatellite loci.

The graph of the effects of a bottleneck on loci mutating by the SMM shows that, after some time, a gene diversity deficiency can be seen, rather than an excess. This means that the "window of time" in which a bottleneck can be detected, due to a diversity excess, is much smaller for loci

mutating according to the SMM. However, it must be remembered that a strict stepwise mutation process is very unrealistic and only a small deviation from this model removes this deficiency (Cornuet and Luikart 1996).

To explain this deficiency, the nature of the allelic distributions of loci under the SMM and the types of mutations experienced must be considered. Microsatellite allele lengths are usually adjoining, forming a contiguous distribution. A bottleneck may cause gaps to form in the distribution when particular alleles are lost. The higher the number of alleles the greater the likelihood of gaps. After the bottleneck, these gaps will be progressively filled by mutations leading to a transient excess of alleles and a concomitant gene diversity deficiency (Cornuet and Luikart 1996). In a stable population, with no gaps in the distribution, most mutations at microsatellite loci transform one allele into another that is already found in the population. When there are gaps in the distribution, more of the mutations will result in the formation of new alleles therefore the allelic diversity at loci under the SMM can be expected to increase rapidly, temporarily, after a bottleneck.

Cornuet and Luikart (1996) applied their bottleneck test to the data gathered by Taylor *et al.* (1994) in a study of the northern hairy-nosed wombat (*Lasiorhinus krefftii*) in Australia. This is an extremely threatened species with a single population that was reduced to only 20-30 individuals in 1981. Taylor *et al.* (1994) detected an apparent reduction in variation in this species at nine polymorphic microsatellite loci, when compared with two populations of the closely related southern hairy-nosed wombat (*L. latifrons*). Cornuet and Luikart (1996) found that eight of the nine loci showed a gene diversity excess and that the results were significant, supporting the theory of a recent bottleneck. Interestingly, they also found evidence of recent reductions in population size in the more variable southern hairy-nosed wombat populations.

Their proposed test for a bottleneck can be carried out using the computer package BOTTLENECK (v. 1.2.02) (Cornuet and Luikart 1996) which calculates the expected gene diversity for each locus in each population under three microsatellite mutation models. The expected gene diversity under the IAM is calculated following the theory and formulae derived by Watterson (see Cornuet and Luikart 1996). No such formulae exist for the SMM, so the expected gene diversity is determined using a computer simulation of the mutational process (Cornuet and Luikart 1996). The observed gene diversities at each locus and each population are then compared with expectation. This test was applied to all the fragment populations to see whether there was any genetic evidence that these populations had experienced bottlenecks as a result of fragmentation in their recent history.

5.1.2.4 Population differentiation

In the 1920s, Wright introduced the concept of the fixation index (F) which he later developed into the F-statistics (Wright 1950; Wright 1973) with which population structure could be analysed in terms of the genetic differentiation between groups and populations. The F-statistics concern three parameters, F_{IS}, F_{ST} and F_{IT}, and are equivalent to the parameters f, θ and F (Weir and Cockerham 1984) that measure the correlations of genes:

f	within individuals within populations;
θ	of different individuals in the same population (coancestry);
F	within individuals (inbreeding).

These parameters can be estimated from the observed components of variance in allele frequencies:

$$1 - \hat{F} = \frac{c}{a + b + c}$$

$$\hat{\theta} = \frac{a}{a + b + c}$$

$$1 - \hat{f} = \frac{c}{b + c}$$

where
a = variance between populations
b = variance between individuals within populations
c = variance between gametes within individuals

This is the basis of the method developed by Weir and Cockerham (1984) for estimating F-statistics that has become the standard approach.

When investigating population differentiation it is the parameter F_{ST}, estimated by θ, that is of interest as it represents the proportion of variance that is due to differences between populations. When alleles are drawn at random from a group of populations, it represents the relative probabilities that two alleles drawn from the same population are the same and two alleles drawn from different populations are the same. If these probabilities do not differ then there is no discernible differentiation between the populations. Values of θ range between –1 and +1; the greater its value, the greater the amount of variation attributable to differences between populations and the greater the differentiation. Negative values indicate that there is more variation within populations than between them, therefore they are not differentiated. In this study, estimates of F_{ST} were calculated for all possible pairwise combinations of fragment populations and across all populations (using the computer package ARLEQUIN (v. 1.1)) to establish whether there was any differentiation between them and, if so, which were least related. Conclusions about any population structuring could then be drawn.

Traditional approaches to the estimation of the F-statistics are based on the assumptions of the infinite alleles model of mutation. As discussed in section 1.5.2.2, this model of mutation is unlikely to be valid for microsatellites and they generally evolve at a much faster rate than the allozymes for which the methods were intended. Slatkin (1995) introduced a new parameter, R_{ST}, analogous to Wright's F_{ST} that is appropriate for microsatellites mutating at a fast rate and under the stepwise mutation model, in which the new allelic state depends on the size of the previous allele. The method of calculating R_{ST} proposed by Slatkin is similar to the method for estimating θ proposed by Weir and Cockerham (1984) in that both estimate the between population proportion of variance. However, the R_{ST} calculation takes allele sizes into account whereas, in the calculation of θ, only the identity or non-identity of alleles features (Slatkin 1995). This method is therefore more informative as it not only considers whether alleles are the same but also how similar they are, assuming a step-wise mutational process.

Slatkin (1995) found that estimates of differentiation based on F_{ST}, using microsatellite data, showed too much genetic similarity, especially when populations were very differentiated, whereas R_{ST} did not show this tendency for inaccuracy. However, when differentiation is low, F_{ST} is reliable. This is presumably because as two populations differentiate, microsatellites mutate so that they show increasing difference in size as well as being different in state. The estimate of R_{ST} takes this into account whereas F_{ST} estimates continue just to see them as different and not take account of this increasing difference with time. Therefore, with increasing time F_{ST} will tend to underestimate the level of differentiation at microsatellite loci but over short timescales, F_{ST} may be expected to perform as well as R_{ST}.

Estimates of R_{ST} were obtained by calculating ρ, analogous to θ, as described by Goodman (1997) using his computer program RSTCALC (v.2.2), for all pairwise comparisons of the fragment populations and over all populations. To determine whether the estimates of θ and ρ were significantly different from zero, permuation tests were carried out by both computer programs used. This involves permuting haplotypes between populations and recalculating the statistics. This is done repeatedly and then the probability of obtaining the real result is determined, it represents the probability of obtaining the result by chance if the null hypothesis of no differentiation is true.

Comparisons of θ and ρ cannot be made directly as they are different statistics. However, both can be used to estimate the number of migrants per generation (Nm) between the populations and these estimates can be compared. The estimates of Nm from θ and ρ are calculated as:

$$\hat{N}m = \frac{1}{4}\left(\frac{1}{\theta} - 1\right) \quad \text{and} \quad \hat{N}m = \frac{d_s - 1}{4\,d_s}\left(\frac{1}{\rho} - 1\right)$$

where d_s is the number of populations. This factor does not need to be included when estimating Nm from θ as the sampling regime is taken into account when θ is estimated using Weir and Cockerham's (1984) method (Slatkin 1995). The reliability of these measures as actual estimates of migration is extremely doubtful; they are only intended to be used when the levels of migration are low (Wright 1943). Even then they are estimates based on estimates, both of which involve many assumptions that are unlikely to be true in real biological populations (Bossart and Pashley Prowell 1998). They can, however, be used as an alternative measure of population differentiation. Estimates of Nm were generated along with θ and ρ, by the computer packages ARLEQUIN (v. 1.1) and RSTCALC (v.2.2).

5.1.2.5 Isolation by distance

Isolation by distance is a concept first developed by Wright (1943). He mathematically described a situation where a population has complete continuity of distribution but interbreeding is restricted due to limited dispersal. Under these conditions remote groups of individuals may become genetically differentiated merely because of isolation by distance. When this is the situation there is a correlation within the samples between the pairwise genetic differentiation and geographic distance between samples. This correlation will also be seen if the level of differentiation between the samples in separate populations is only influenced by the distance between them.

Slatkin (1993) took this further and showed that, as F_{ST} is correlated with distance, the number of migrants (Nm) shows a negative relationship with distance and that this correlation can be detected by plotting log(Nm) against log(distance). This method has been used successfully to show isolation by distance, for example, Goodman (1998) showed that isolation by distance influenced the degree of differentiation between European harbor seal populations.

Pairwise population comparisons generate half matrices and the association between the elements of two such matrices can be tested using the Mantel test (Manly 1997). First the association between the corresponding elements of the matrices is calculated (e.g. by regression analysis) then the order of the elements in one of the matrices is randomly reallocated and the association recalculated. This is repeated for a specified number of permutations allowing the probability of obtaining the observed result by chance (the significance level) to be estimated; 1000 permutations is the recommended minimum (Manly 1997).

Nm estimates can be made from F_{ST} and R_{ST} estimates when the levels of migration are low but obviously cannot be made when F_{ST} or R_{ST} estimates are negative. This can lead of incomplete matrices when some datasets are analysed and in these cases Slatkin's method of testing for isolation by distance is inapplicable. Rousset (1997) showed that the correlation between genetic differentiation and geographic distance can be tested directly by testing the regression of $F_{ST}/(1-F_{ST})$ estimates for pairwise population comparisons on the natural logarithm of pairwise distances. This is a more appropriate analysis method for the data in this study and Mantel tests were carried out to compare the matrices generated using GENEPOP (v. 3.1).

5.2 METHODS

Most of the samples of DNA from red squirrels used for the microsatellite analysis were also used in the mitochondrial study described in Chapter 3. 136 samples of red squirrels from the Belgian populations of Brede Zijpe (BZ), Gasthuisbos (GH), Kegelslei (KE), Luisbos (L), Tallaarthof (T) and Anwerp water works areas 1, 2 and 3 (AWW1-3) Merodese Bossen (MB) and Peerdsbos (P) in Belgium and 27 from Waldhäuser (WH) in Germany were included in this study. The sample collection and extraction procedures are described in sections 3.2.1 to 3.2.2 and section 2.2.2 respectively.

5.2.1 DNA concentration

It was preferable to use DNA samples of approximately equal concentrations in the PCR amplification reactions. Each DNA sample was run out on a 0.8% agarose gel next to a known quantity of marker DNA and stained with Ethidium bromide (as described in section 2.2.3). The concentration of each sample was estimated, by comparison with the marker, as containing closest to 500ng/μl, or 200ng/μl, or 100ng/μl, or 50ng/μl and less. They were all then diluted in TE buffer, 1μl in 50μl, 1μl in 20μl, 1μl in 10μl and 1μl in 5μl respectively, to produce DNA samples all of approximately 10ng/μl.

5.2.2 PCR amplification

5.2.2.1 Reaction optimisation

The reaction conditions for variable microsatellite primer sets were optimised by carrying out "cold" reactions (without labelling with ^{32}P γ-dATP) which were visualised by electrophoresis through metaphor agarose gels and Ethidium bromide staining (see section 4.2.17). Initial optimisation reactions were set-up according to the approach described by Cobb and Clarkson (1994) which involves setting-up nine separate reactions (plus a negative control), each varying in the concentrations of nucleotides, magnesium chloride, primers and template. Generally, if the primers are going to amplify the region intended successfully, one of these reactions should produce the required product. The ten reaction mixes for 25μl reactions are given in table 5.1.

		Reaction mixes: volume in µl and final conc^{n}. *(units as stock)*									
Reagent	**Stock conc^{n}.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**
reaction buffer IV (Advanced Biotechnologies)	10x	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*	2.5 *(1x)*
nucleotide mix (Pharmacia Biotech)	1.25mM	1 *(0.05)*	2 *(0.1)*	4 *(0.2)*	4 *(0.2)*	1 *(0.05)*	2 *(0.1)*	2 *(0.1)*	1 *(0.05)*	4 *(0.2)*	2 *(0.1)*
$MgCl_2$ (Advanced Biotechnologies)	25mM	1 *(1)*	3 *(3)*	5 *(5)*	3 *(3)*	5 *(5)*	1 *(1)*	5 *(5)*	1 *(1)*	3 *(3)*	3 *(3)*
primer t^{pro}	10µM	1 *(0.4)*	1 *(0.4)*	1 *(0.4)*	2 *(0.8)*	2 *(0.8)*	2 *(0.8)*	4 *(1.6)*	4 *(1.6)*	4 *(1.6)*	2 *(0.8)*
primer t^{phe}	10µM	1 *(0.4)*	1 *(0.4)*	1 *(0.4)*	2 *(0.8)*	2 *(0.8)*	2 *(0.8)*	4 *(1.6)*	4 *(1.6)*	4 *(1.6)*	2 *(0.8)*
Taq DNA poly. (Promega)	1 U/µl	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*	0.2 *(1)*
template	~10 ng/µl	1 *(~0.4)*	2 *(~0.8)*	4 *(~1.6)*	1 *(~0.4)*	2 *(~0.8)*	4 *(~1.6)*	1 *(~0.4)*	2 *(~0.8)*	4 *(~1.6)*	0 *(0)*
Sterile water		16.5	12.5	6.5	9.5	9.5	10.5	5.5	9.5	2.5	12.5

Table 5.1: The ten PCR reaction mixes used to carry out the initial optimisation of each reaction. The volume of reagent in each mix is given and its final concentration in a 25µl reaction is shown in brackets.

The reactions were amplified on a PTC-200 thermocycler (MJ Research, Inc.) according to the following program. The annealing temperature of step 3 was varied for each primer set according to the optimum annealing temperature calculated by the primer design package, OLIGO (v.4.0; National Biosciences, Inc.). To begin with a slightly lower than optimal annealing temperature was tried.

Step	Temperature (°C)	Time (minutes)
1	94	5
2	94	1
3	54	1
4	72	1 1/2
5	Go to step 2	29 more times
6	72	5
7	28	1

Further optimisation was then undertaken by varying the concentration of the reagents and the steps of the cycling program. Generally, it was most productive to vary the concentrations of nucleotides and magnesium chloride and to increase the annealing temperature. If a large quantity of unincorporated primers ("primer dimers") could be seen on the gel then the primer concentration was decreased, but this did not have a great effect on the efficiency of the reactions. When a single clean product of the correct size was achieved by the amplification reaction then the reaction was tested on a few samples with labelled primers and visualised on polyacrylamide gels as described below. Further optimisation of the "hot" reactions was carried out where required, especially to optimise the amount of labelled primer used.

5.2.2.2 Primer end-labelling

In all cases the forward reaction primer (the primer with "f" as the final letter of its name) was end-labelled with ^{32}P γ-dATP using the following protocol:

1. For 48 pmol of labelled primer, set up the following reaction:

4.8µl	primer (stock concentration 10pmol/ml)
2.4µl	5x forward reaction buffer (GibcoBRL)
1.2µl	T4 polynucleotide kinase (GibcoBRL)
3.6µl	^{2}P γ-dATP (1.3 Mbq)

2. Mix gently and incubate at 38°C for at least 1 hour.

This produces primers labelled with ~0.03Mbq/pmol at a concentration of 4pmol/μl. The actual signal quantity from the primers depended on the efficiency of the labelling reaction and varied between primers. Some reactions were very efficient and the amount of ^{32}P γ-dATP used in the labelling reaction could be reduced. An equivalent quantity of water was added to the reaction to maintain the concentrations.

5.2.2.3 The PCR reactions

All of the samples were amplified with each of the five primer sets that had been shown to be variable in red squirrels (chapter 4). The reaction mixes used for each primer set are given in table 5.2. The PCR reaction mixes were prepared in the wells of thermo-fast 96 well plates covered with a sticky sealing sheet (Advanced Biotechnologies). The sterile distilled water and DNA were aliquoted into the wells of each plate. Care was taken to be sure of the order of the samples in the wells of the plates and that that order was maintained in loading the wells of the polyacrylamide gels. A master mix of the remaining reagents was prepared (enough for all the reactions to be performed) and the appropriate quantity was aliquoted into each well of the plates with the DNA templates. A drop of mineral oil was added on top of each reaction and the plates were sealed with the sticky cover before running the reactions. All the reactions were carried out successfully with the following program on a PTC-100 thermocycler (MJ Research, Inc.):

Step	Temperature (°C)	Time (minutes)
1	94	3
2	94	1
3	54	1
4	72	1 1/2
5	Go to step 2	29 more times
6	72	5
7	28	1

Reactions with primer sets:		RSμ1		RSμ3		RSμ4		RSμ5		RSμ6	
Reagent	**Stock conc^{n}.**	Volume (μl)	Final conc^{n}.	Volume (μl)	Final conc^{n}.	Volume (μl)	Final conc^{n}.	Volume (μl)	Final conc^{n}.	Volume (μl)	Final conc^{n}.
reaction buffer IV (Advanced Biotechnologies)	10x	1.25	1x	1.25	1x	1.25	1x	1.25	1x	1.25	1x
nucleotide mix (Pharmacia Biotech)	1.25mM	1.5	0.15mM	1	0.1mM	2	0.2mM	0.5	0.05mM	0.5	0.05mM
$MgCl_2$ (Advanced Biotechnologies)	25mM	0.75	1.5mM	0.5	1mM	0.5	1mM	0.5	1mM	0.5	1mM
primer **f** unlabelled	10μM or 10pmol/μl	0.3	0.24μM (3pmol)	0.4	0.32μM (4pmol)	0.9	0.72 (9pmol)	0.9	0.72μM (9pmol)	0.9	0.72μM (9pmol)
primer **f** labelled with ^{32}P	4μM or 4pmol/μl	0.5	0.16μM (2pmol)	0.25	0.08μM (1pmol)	0.25	0.08 (1pmol)	0.25	0.08μM (1pmol)	0.25	0.08μM (1pmol)
primer **r**	10μM or 10pmol/μl	0.5	0.4μM (5pmol)	0.5	0.4μM (5pmol)	1	0.8 (10pmol)	1	0.8μM (10pmol)	1	0.8μM (10pmol)
Taq DNA poly. (Promega)	5 U/ml	0.2	1 U	0.2	1 U	0.2	1 U	0.2	1 U	0.2	1 U
template	~10 ng/ml	1	~0.8ng	1	~0.8ng	1	~0.8ng	1	~0.8ng	1	~0.8ng
sterile water		6.3		7.1		5.4		6.6		6.6	

Table 5.2: The reaction mixes used to amplify the microsatellite loci RSμ1, RSμ3, RSμ4, RSμ5 and RSμ6 in *S. vulgaris*. The volumes of each reagent used and their final concentrations in a 12.5μl reaction are given.

5.2.3 Visualisation on polyacrylamide gels

The gels were poured, run and the samples visualised by exposing the gel to an X-ray film, as described in section 2.2.8. 5μl of loading buffer (stock of 20ml 0.5M EDTA with 200mg bromophenol blue (BDH) and 200mg xylene cyanole FF (Sigma) dissolved, diluted 1:10 in formamide) was added to each reaction before it was denatured for loading. Between 2μl and 5μl of each reaction, including loading buffer, was loaded into the wells of the gel and the gels were run at a constant temperature of 50°C for around 2 hours (slightly longer for the long products and less time for the shorter products).

A set of sequencing reactions were run alongside the microsatellite products in order to size them accurately. The control region sequence of the red squirrel was used as a marker, the sequencing reactions carried out directly from cleaned PCR products of DNA amplified with primers RSCR1 and RSCR4 as described in section 3.2.3. The products were then sequenced manually with primer RSCR4, as described in section 4.2.15.2. The four reactions were then run on the same gel as the microsatellite products and the scoring of the samples could then be carried out by direct comparison with the sequence.

Most gels required several separate exposures to X-ray film of varying times (between 12 hours and several weeks) to allow all the products to be visualised clearly. This was presumably because different samples amplified with varying success and incorporated varying amounts of labelled primer. Products emitting less signal could be visualised by long exposures which allowed even very feint products to be scored, and products of a very high concentration emitting a very strong signal were most accurately scored when only exposed to the film for a very short time.

5.2.4 Data Analysis

The allelic diversity and observed heterozygosity were calculated manually as was the gene diversity or expected heterozygosities of the Belgian fragment populations of Brede Zijpe, Gasthuisbos, Kegelslei, Luisbos, Tallaarthof and Antwerp water works, using the equation defined by Nei (1987; equation 1). The GENEPOP (v. 3.1) package (Raymond and Rousset 1995a) was used to estimate gene diversity in the large populations (according to equation 2) and to calculate the allele frequencies for each population.

BIOMSTAT v.3.2 (Applied Biostatistics Inc.) was used to carry out Mann-Whitney U and Kruskal-Wallis tests, and to calculate the correlation of variance in heterozygosity and sample size. To compare BZ and GH over two years, pairwise exact tests were carried out using GENEPOP, the θ estimates of F_{ST} were calculated by FSTAT (v.2.8) (Goudet 1999) and the ρ estimates of R_{ST} were calculated by RSTCALC (v.2.2)(Goodman 1997; http://helios.bto.ed.ac.uk/evolgen/).

Deviations from Hardy-Weinberg were tested for using GENEPOP (v. 3.1) (Raymond and Rousset 1995a) by carrying out an exact test as described by Guo and Thompson (1992). Markov chain parameters of 100 batches and 1500 iterations were used so that the standard errors of the probability values were less than 0.01. The expected number of alleles at each locus was calculated by hand according to the equation of Ewens (1972). The expected gene diversities for each Belgian population under three mutation models were calculated using the package BOTTLENECK (v. 1.2.02) (Cornuet and Luikart 1996, http://www.ensam.inra.fr/URLB) and the significance of the excess was tested using the Wilcoxon sign-rank test by the same package.

The differentiation across and between the Belgian fragment populations was investigated by estimating F_{ST} and R_{ST} for all the populations together and for each population pair. F_{ST} estimates were calculated as defined by Weir and Cockerham (1984) using ARLEQUIN (v. 1.1) (Schneider *et al.* 1997) and R_{ST}, as defined by Slatkin (1995), was estimated using RSTCALC (v.2.2). Both sets of estimates were tested for significance by carrying out 1000 permutations of genotypes between populations. The probability value of the test is the proportion of permutations leading to an estimate larger than or equal to the observed one and refers to the probability of obtaining the observed level of differentiation by chance. Isolation by distance was tested for by comparing the matrices of transformed pairwise F_{ST} and R_{ST} and geographic distance using a Mantel test (with 5000 permuations) as described by Rousset (1997); these analyses were carried out using the ISOLDE program in the GENEPOP package.

5.3 RESULTS

Five microsatellite loci were amplified from 163 samples of red squirrel DNA. The complete set of results is given in appendix B. Four samples from the small Belgian populations were missing and so could not be included in the study, and a few of the amplification reactions were not successful. This may be due to poor DNA quality, a low frequency of null alleles, or just random processes that affect PCR reactions.

The allele frequency distributions for each locus in each population are shown in appendix C. Most alleles are found in several populations, both Belgian and German, and there is a very low frequency of private alleles (alleles found in only one sample). The German sample has seven private alleles, one in each of loci RSμ1, 3 and 4, and two in each of loci RSμ5 and 6. There is only one private allele in the Belgian populations, allele 161 of RSμ3 found in the Brede Zijpe population only, although allele 176 of RSμ1 is only found in Brede Zijpe as well as the German sample.

Figure 5.2 shows the number of alleles found at each locus and the proportion of individuals in the total study that are heterozygous at each locus. These graphs illustrate the variability of each locus. RSμ4 is the most variable with more alleles detected and the highest number of heterozygous individuals, whereas RSμ6 is the least variable. As the comparisons are made between loci amplified from the same large sample of individuals then, if the loci are neutral, these trends can be taken to reflect the different mutation rates of the loci.

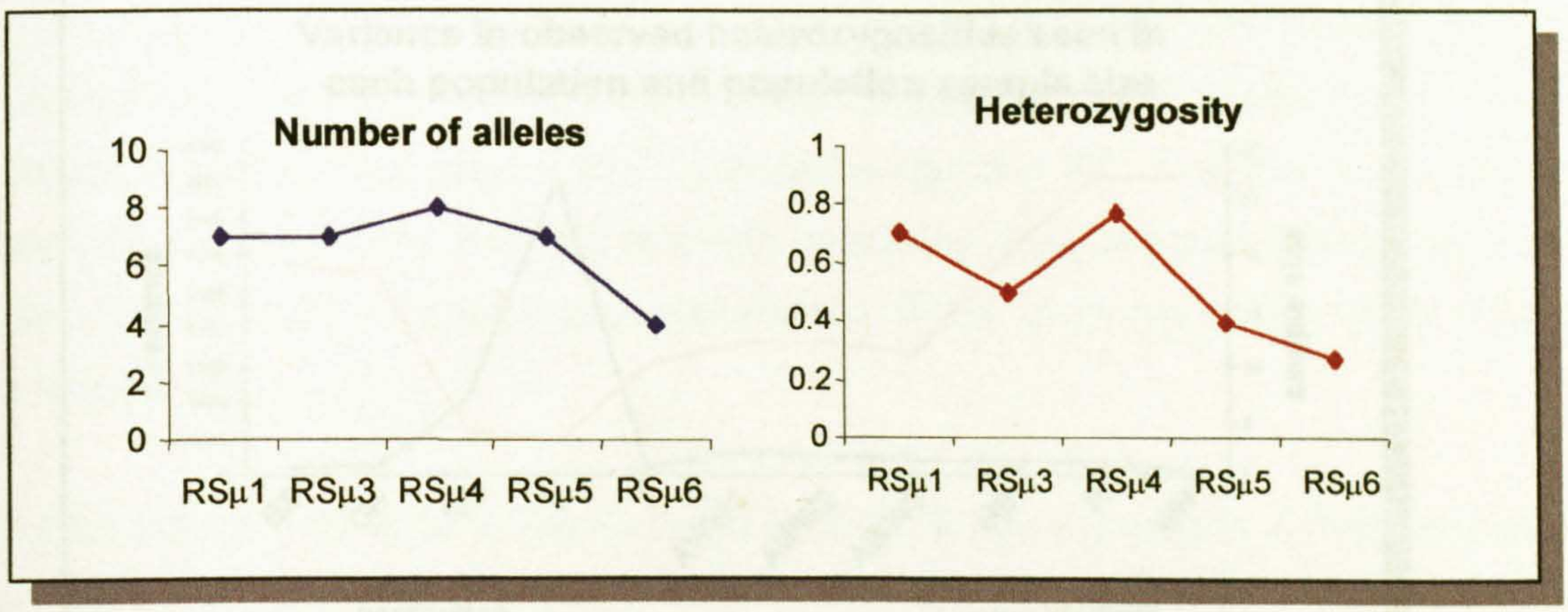


Figure 5.2: The number of alleles and the proportion of heterozygous individuals found for each locus in all the red squirrels sampled from both Belgium and Germany.

As illustrated in figure 5.3, variation in the proportion of heterozygotes for each locus in each population is negatively correlated to the sample or population size. The small populations show much more variation in the levels of observed heterozygosity at each locus than the larger populations. Closer examination of the data shows that the small populations have both very low and very high levels of heterozygosity at different loci whereas the larger populations have more similar levels of heterozygosity at each locus. This relationship has a significant negative correlation of r = -0.67 (*p=0.02*).

This can be thought of as a sampling effect where the populations studied are effectively samples of varying size drawn from a large metapopulation. The smaller the sample the less accurate will be the representation of the metapopulation variance and the larger will be the sample variance. This process can be explained in population genetic terms, as analysing each independent locus in each population is in effect the analysis of repeated independent samples of genes. Small populations are more susceptible to the random effects of genetic drift. A small sample drawn from a larger population by a process such as a bottleneck or a founder event will randomly include individuals some of which are heterozygotes and some homozygotes at different loci. If a large sample is drawn, then that sample is likely to reflect more accurately the proportions of heterozygotes and homozygotes from the original population at each of the loci analysed. However, if the sample is very small (as in the populations of Kegelslei (KE) and Luisbos (L)) then the differences seen in the numbers of heterozygotes and homozygotes at each locus, in the comparison of the same individuals, will be more marked, resulting in a greater variation in numbers of heterozygotes seen at each locus.

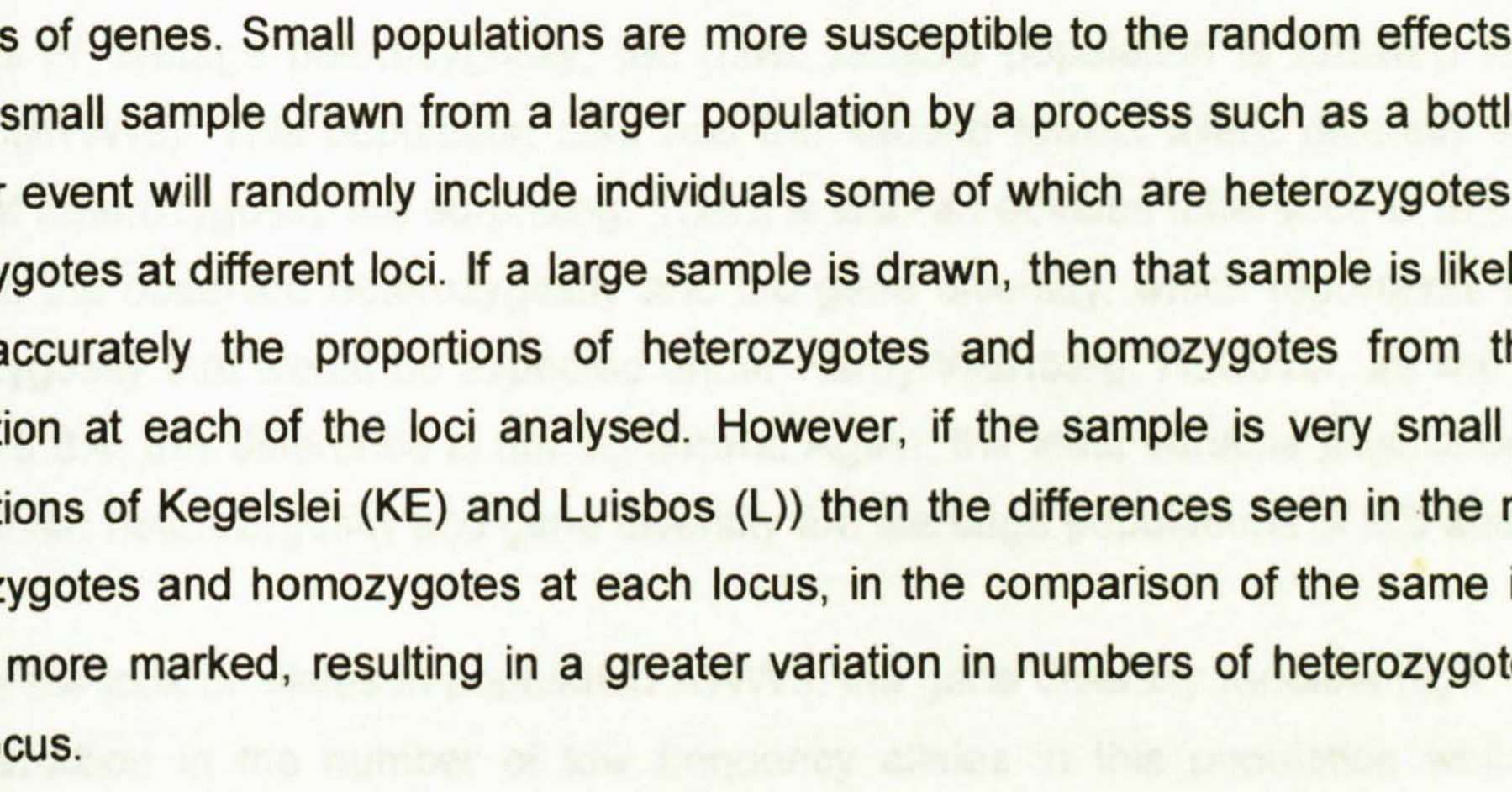


Figure 5.3: A graph to illustrate the relationship seen between the variance in the observed heterozygosities seen in each population and the sample size. The data for BZ and GH was pooled for the years 1995 and 1996.

5.3.1 Belgian intrapopulation variation levels

Table 5.3 and figure 5.4 illustrate the amount of variation found in each study population as indicated by several diversity measures. The number of alleles found in each population reflects the population/sample size, demonstrating the dependency of allelic diversity on sample size. Exceptions to this are the populations of Antwerp water works area 3 (AWW3), which has fewer alleles than the other populations of similar size (Tallaarthof (T) and Antwerp water works areas 1 and 2 (AWW1 and AWW2)), and the large populations of Merodese Bossen (MB) and Peerdsbos (P), which have fewer alleles than some of the fragment populations despite a much larger sample size. It should be remembered that the fragment populations are totally sampled (only one individual from each of a few populations was missed) so the allelic diversity measurements are valid counts of the number of alleles found in the population.

In terms of average heterozygosity, the most variable population is Antwerp water works area 3 (AWW3). This population also had the second lowest allelic diversity so the high levels of heterozygosity are surprising. There is also an obvious difference in this population between the observed heterozygosity and the gene diversity, which represents the level of heterozygosity that would be expected under Hardy-Weinberg. However, as will be seen in section 5.3.4, this difference is not significant. Again, the least variable populations in terms of observed heterozygosity and gene diversity are the large populations of MB and P.

Despite the lack of alleles in population AWW3, the gene diversity remains high. This is due to a reduction in the number of low frequency alleles in this population whilst the high frequency alleles are maintained, making the greater contribution to the gene diversity measure. This effect can be illustrated in a comparison of the allelic distributions (appendix C) of RSμ3 in the populations Gasthuisbos (GH), Kegelslei (KE) and Luisbos (L). L only has two alleles of similar frequency at this locus whilst GH and KE have five and four alleles respectively, but only one of them are at a high frequency. Despite these differences in allelic diversity, the gene diversity of all three populations at this locus are similar (0.58, 0.56 and 0.5 respectively).

To compare the overall variation levels found in all the Belgian populations they were ranked according to the independent measures of allelic diversity and observed heterozygosity. Data for the populations of BZ and GH is available for both 1995 and 1996, but as the other fragment populations were only sampled in 1996, the data for 1996 will be used in the population comparisons. The ranks were totalled for each population, the total rank indicates

the relative variability of the population. The results of this analysis are also shown in table 5.3. BZ is the most variable of the Belgian populations, with GH and T being slightly less variable. The least variable is P, due to both a low number of alleles and low level of heterozygosity. MB is also low in the rankings as it also has low heterozygosity but has more allelic diversity than P. AWW1 and 2 are equally variable in terms of allelic diversity and heterozygosity but AWW2 has slightly more gene diversity due to variations in their allele frequencies; AWW3 is more variable overall than its neighbours due to the remarkably high level of heterozygosity, although the allelic diversity is very low. The heterozygosity levels seen in AWW1 and AWW2 are the lowest of the fragment populations, but the allelic diversity in these populations remains high. These two populations along with BZ and GH are the only ones to have less heterozygosity than gene diversity (expectation under Hardy-Weinberg).

The heterozygosity found in the fragment populations (other than AWW3) remains quite consistent despite fluctuations in the number of alleles seen in these populations. Although it is interesting to note that the observed average heterozygosity seen in the populations of AWW1 and AWW2 is lower than in the other fragment populations but similar to the larger populations of MB and P.

Table 5.3 also gives a count of the number of low frequency alleles (defined here as alleles present in the population at a frequency of less than 0.1) found in each population at all five loci. Populations KE and L are too small to have alleles at this low frequency, only carrying eight and six alleles each. Of the other populations, as may be expected, the German sample has a large number of rare alleles (10) but the Belgian fragment population BZ has almost as many (9) and far more than either of the large Belgian populations. The sample size of P is the same as that for WH but despite this, it only contains one rare allele. Populations AWW1, 2 and 3 are all about the same size, yet area 3 has no rare alleles whilst areas 1 and 2 have five and six respectively.

population		population size	number of samples	total number of alleles	average number of alleles	number of rare alleles (frequency <10%)	average hetero-zygosity	average gene diversity	populations ranked by variation levels: rank by allelic diversity	rank by average hetero-zygosity	total rank	overall rank
Belgian fragment populations:												
Brede Zijpe 1995		13	12	19	3.8	8	0.53	0.54				
Gasthuisbos 1995		16	15	19	3.8	5	0.51	0.54				
Brede Zijpe 1996		17	16	20	4	9	0.51	0.55	2	6	8	2
Gasthuisbos 1996		14	14	18	3.6	5	0.5	0.51	3	7	10	3.5
Kegelslei		4	4	14	2.8	0	0.55	0.49	8.5	4	12.5	6
Luisbos		3	3	11	2.2	0	0.53	0.44	11	5	16	10
Tallaarthof		10	10	16	3.2	2	0.58	0.54	7	3	10	3.5
Antwerp Water Works	area 1	13	12	17	3.4	5	0.47	0.47	5	8.5	13.5	7.5
	area 2	13	12	17	3.4	6	0.47	0.52	5	8.5	13.5	7.5
	area 3	11	11	13	2.6	0	0.73	0.56	10	1	11	5
Other populations:												
Merodese Bossen		~300	20	17	3.4	6	0.44	0.42	5	10.5	15.5	9
Peerdsbos		~600	27	14	2.8	1	0.44	0.42	8.5	10.5	19	11
Waldhäuser			27	26	5.2	10	0.64	0.62	1	2	3	1

Table 5.3: The level of variation found in each study population indicated by various diversity measures: total number of alleles found in five loci, average number of alleles across five loci, total number of low frequency (rare, with frequencies less than 0.1) alleles at five loci, average heterozygosity across five loci and average gene diversity (as defined by Nei, 1987). The gene diversities of the fragment populations were calculated as for total populations and for the other populations, as for samples from populations. The Belgium populations were ranked by allelic diversity and average heterozygosity as a way to compare the levels of variation in each population. The total ranks were calculated and the overall rank is also given. Where populations tied the average rank was calculated and given.

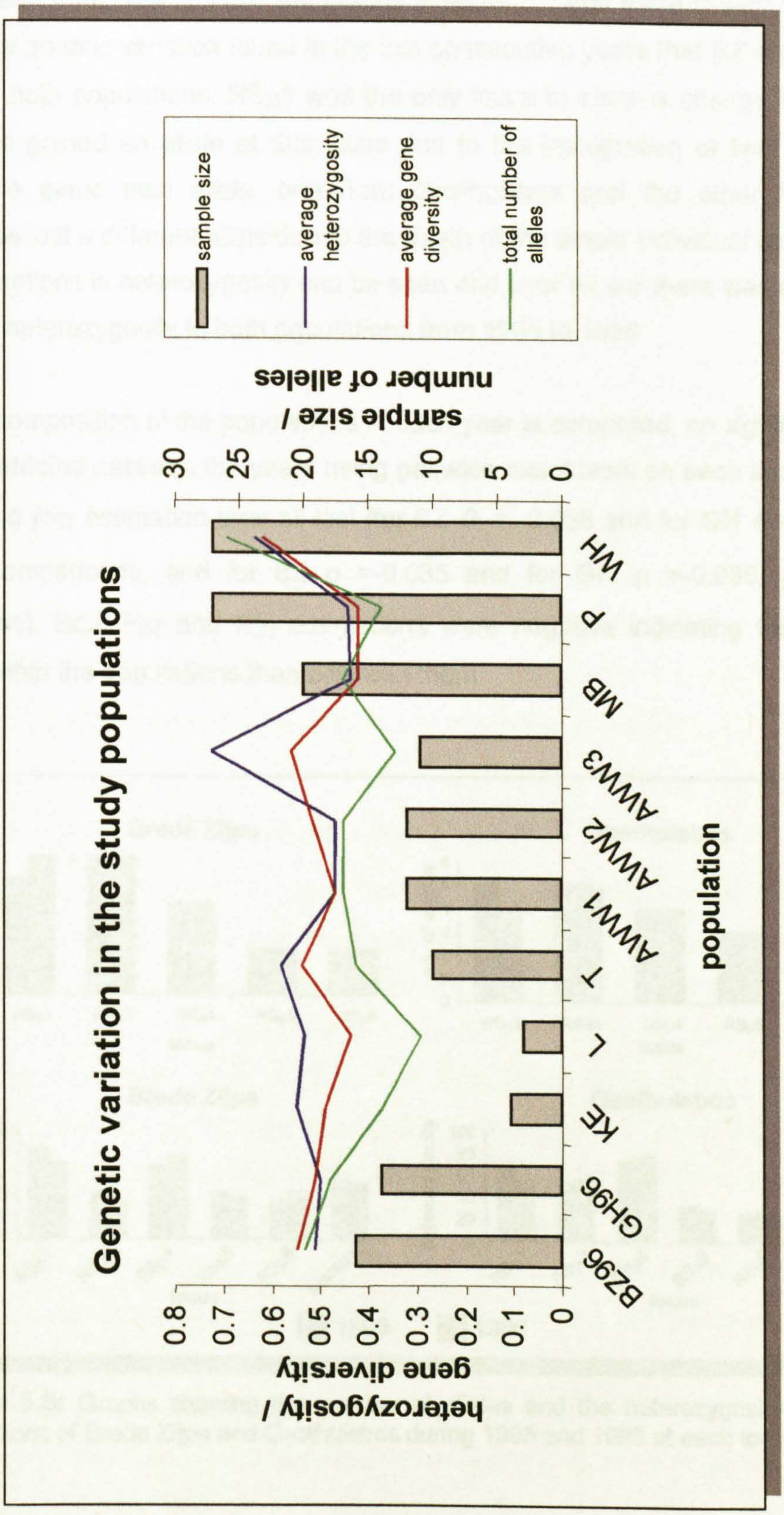


Figure 5.4: The levels of genetic variation found in the study populations at five microsatellite loci, as measured by average heterozygosity, average gene diversity and the total number of alleles found at the five loci.

5.3.2 Changes in Brede Zijpe and Gasthuisbos over two years

It can be seen from table 5.3 and the graphs in figure 5.5 that there is very little difference in the levels of genetic variation found in the two consecutive years that BZ and GH have been studied. In both populations, RSμ1 was the only locus to show a change in allele number. Brede Zijpe gained an allele at this locus due to the immigration of two individuals, both carrying the same new allele, one from Gasthuisbos and the other from Tallaarthof. Gasthuisbos lost a different allele due to the death of the single individual carrying that allele. Slight fluctuations in heterozygosity can be seen and over all loci there was a slight reduction in average heterozygosity in both populations from 1995 to 1996.

When the composition of the populations in each year is compared, no significant differences could be detected between the years using pairwise exact tests on each locus (*0.48<p<1.0*), and F_{ST} and R_{ST} estimation over all loci (for BZ θ = -0.036 and for GH θ = -0.027; *p=0.98* for both comparisons, and for BZ ρ =-0.035 and for GH ρ =-0.036; *p=0.99* for both comparisons). Both F_{ST} and R_{ST} estimations were negative indicating that there is more variation within the populations than between them.

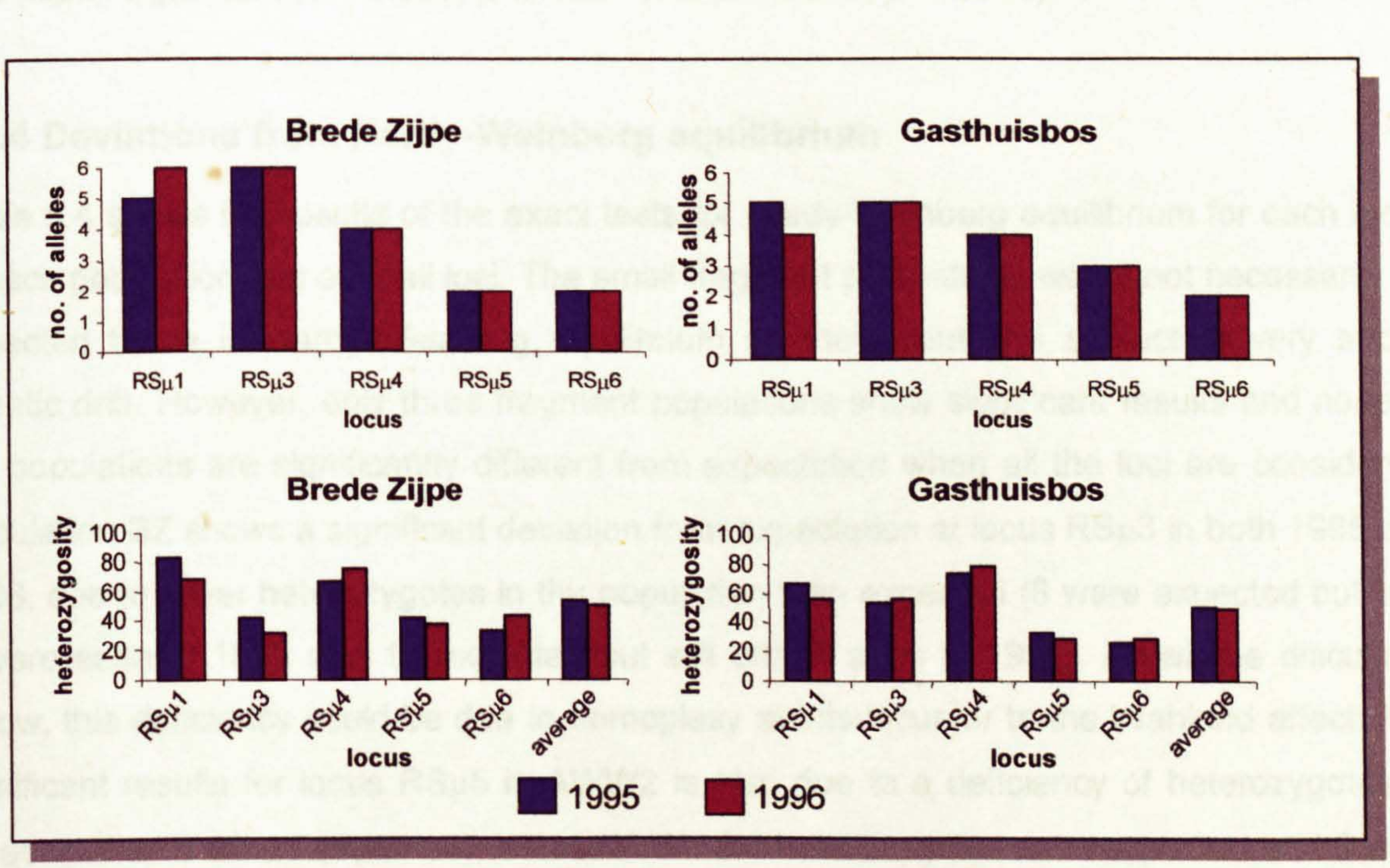


Figure 5.5: Graphs showing the number of alleles and the heterozygosity found in the populations of Brede Zijpe and Gasthuisbos during 1995 and 1996 at each locus.

5.3.3 A comparison of the German and Belgian samples

It can be seen from table 5.3 and the graph in figure 5.4 that the German (Waldhäuser) sample has more variation as both allelic diversity and gene diversity, than any of the Belgian populations. The average heterozygosity is also higher in the German sample than all the Belgian populations, except in Antwerp water works area 3 (which has a surprisingly high level of heterozygosity). Mann-Whitney U tests comparing the Waldhäuser sample with all the Belgian results for average gene diversity (U=10, *p=0.056*), average heterozygosity (U=9, *p=0.1*) and average allele number (U=10, *p=0.055*) showed that the German sample did not quite have significantly more variation than the Belgian populations, although, it is clearly the most variable of the populations studied.

From the allele frequency distributions in appendix C it can be seen that most of the alleles in the Belgian populations are also found in the German sample (only 7 alleles out of 33 are found in Belgium but not Germany). Despite the commonality in their allele content, a comparison of the pooled total Belgian sample with the German sample shows that they are in fact highly differentiated. Pairwise exact tests for each locus are highly significant (*p<0.0001*, except locus RSμ1 where *p=0.005*) and estimates of F_{ST} and R_{ST} over all loci are also highly significant ($\theta = 0.083$, *p<0.0001* and $\rho = 0.097$, *p<0.0001*).

5.3.4 Deviations from Hardy-Weinberg equilibrium

Table 5.4 shows the results of the exact tests for Hardy-Weinberg equilibrium for each locus in each population and over all loci. The small fragment populations would not necessarily be expected to be in Hardy-Weinberg equilibrium as they would be subject to very strong genetic drift. However, only three fragment populations show significant results and none of the populations are significantly different from expectation when all the loci are considered. Population BZ shows a significant deviation from expectation at locus RSμ3 in both 1995 and 1996, due to fewer heterozygotes in the population than expected (8 were expected but only 5 were seen in 1995 and 10 expected but still only 5 seen in 1996). As will be discussed below, this deficiency could be due to homoplasy at this locus or to the Wahlund effect. The significant results for locus RSμ5 in AWW2 is also due to a deficiency of heterozygotes (5 present when 8 where expected), but in AWW1 8.6 heterozygotes were expected and 8 were seen. The significant deviation from Hardy-Weinberg expectation in AWW1 must be due to differences between the expected and observed proportions of each genotype, even though, coincidentally, the number of heterozygotes is the same.

Population	RSμ1	RSμ3	RSμ4	RSμ5	RSμ6	all
BZ 1995	0.15	0.03	0.85	1	0.52	0.28
GH 1995	0.76	0.11	0.79	0.19	0.49	0.42
BZ 1996	0.18	0.002	0.54	1	1	0.07
GH 1996	0.63	0.28	0.63	0.15	1	0.61
KE	1	0.43	1	-	1	0.99
L	0.07	0.4	0.4	-	1	0.36
T	0.78	0.52	0.17	1	1	0.87
AWW1	0.49	0.36	0.5	0.04	-	0.2
AWW2	0.12	0.13	0.48	0.03	1	0.08
AWW3	0.21	0.48	0.08	0.2	1	0.24
MB	0.83	0.29	0.53	1	-	0.85
P	0.16	0.16	0.15	1	-	0.19
WH	0.29	0.03	0.1	0.51	0.19	0.05

Table 5.4: The results of the exact tests for Hardy-Weinberg equilibrium for each locus and population, and each population across all loci. The figures given are the probabilities calculated that each of the observed sets of results would have occurred by chance if the populations are in Hardy-Weinberg equilibrium.

The significant deviation from Hardy-Weinberg proportions seen for the German Waldhäuser population when all the loci are considered is very interesting as this population is large and could be expected to be at equilibrium. When locus RSμ3 is not included in the analysis the population does not show a significant result as a significant difference is only seen at this locus. Locus RSμ3 has an increased probability of showing homoplasy due to its compound structure, where alleles of the same size that are different in structure appear to be identical when visualised on polyacrylamide gels. This results in potential misdesignation of genotypes which would distort the genotype frequencies in the population, but may not necessarily reduce the apparent frequency of heterozygotes. The Waldhäuser population shows very little deviation from expectation in terms of heterozygosity (48.15% compared to 48.43% which in section 4.3.3 was shown not to be significant) therefore the significant deviation from Hardy-Weinberg seen is due to a deviation of the genotype frequencies from Hardy-Weinberg proportions rather than a deficiency in heterozygotes.

Of course, it must be considered that by the nature of significance testing one result in twenty will be significant by chance leading to the false dismissal of the null hypothesis. 55 independent tests have been carried out here so at least two significant results would be expected by chance. However, the possibility of homoplasy at locus RSμ3 cannot be ignored and the results from that locus must be treated with caution. All the following analyses were performed both with and without the inclusion of the data from RSμ3. It is preferable to retain

the data from locus RSμ3 as, when only a few loci are examined, the power and reliability of the tests improves greatly with the additional information provided by each additional locus.

When all the 1996 fragment populations were combined as if one population, it was found to differ from Hardy-Weinberg expectation in a highly significant manner (χ^2 tends to infinity). Four of the five loci showed heterozygote deficiencies. This is the "Wahlund effect"; when there is non-random mating within a population it is apparent as a deficiency in heterozygotes. The Wahlund effect confirms that there is non-random mating within the total sample due to the presence of separate populations, therefore they are not in fact one population with random mating in a patchy environment.

The lack of heterozygotes seen in Brede Zijpe at locus RSμ3 may also be due to the Wahlund effect as the area occupied by the BZ population consists of two adjoining areas of woodland. The squirrels occupying these areas were divided into two subpopulations and tested for differentiation which showed that these two subpopulations are in fact one panmictic population as estimates of F_{ST} and R_{ST} were negative (θ = -0.014, *p=0.74* and ρ = -0.027, *p=0.75*) and pairwise exact tests were not significant (*0.11<p<1.0*). The deficiency of heterozygotes in this population was only seen at the one locus so it is more likely to be due to homoplasy than to the Wahlund effect, but is it also possible that it is the result of the random effects of drift by chance reducing the number of heterozygotes at this locus. This illustrates the importance of using data from several loci so that the identification of overall trends is possible.

5.3.5 Expected number of alleles

Table 5.5 shows the number of alleles expected at microsatellite loci (mutating according to the infinite alleles model) in each of the Belgian fragment populations. This is dependent on effective population size and the mutation rate of the alleles. As the size of each fragment population is known, the expected number of alleles could been calculated for three different mutation rates (μ) within the range of rates experienced by microsatellite loci.

The effective population size is likely to be less than the actual population size as not all adults of reproductive age are expected to reproduce each year; therefore the expected number of alleles will be overestimated. However, it can be seen from the table that at mutation rates of 10^{-3} or less, variations in population size make very little difference to the expected number of alleles when population sizes are so small. The overestimate of the effective population size is only likely to have a noticeable effect on the expected allele

number at loci mutating at the rate of 10^{-2} or faster. Microsatellites are generally described as having mutation rates of between 10^{-3} and 10^{-6} and the extremely fast rate of 10^{-2} is very rarely, if ever, seen (Ellegren *et al.* 1995; McDonald and Potts 1997; Schlötterer 1998b).

Also, the infinite alleles model is unlikely to fit the mutation patterns seen at microsatellite loci. The stepwise mutation model is considered to be the other extreme of mutation model for microsatellites (Cornuet and Luikart 1996) but calculation of the expected number of alleles under this model is very complicated. However, it has been shown that the SMM predicts fewer alleles at each locus under all rates of mutation than the IAM (Kimura and Ohta 1978; Estoup *et al.* 1995), so the predictions from the IAM can be considered to be overestimates of the number of alleles expected at microsatellite loci in each population. The IAM predicts that the number of alleles in a population will increase infinitely, whereas the SMM predicts that the number of alleles rapidly reaches a plateau therefore the number of alleles predicted by the SMM can be a lot less than predicted by the IAM (Kimura and Ohta 1978).

The data in the tables clearly show that the number of alleles observed in each population exceeds expectation even for fast mutating loci. The total number of alleles in each population (found at all five loci) was compared with the total expected with the high mutation rates of 10^{-2} and 10^{-3} by carrying out a Kruskal-Wallis test of significance using the BIOMSTAT (v.3.2) package. There are more alleles than expected with a mutation rate of 10^{-2} which is just significant (H=3.99; *p=0.046*) and the difference between observed and expected is highly significant for the mutation rate of 10^{-3} (H=9.96; *p=0.002*). The number of alleles found at each locus was compared with the number expected for a mutation rate of 10^{-3} and in all cases there were significantly more than predicted ($11.9 > H > 6.4$; $p \leq 0.01$). RSμ1 and RSμ4 have significantly more alleles than expected even at the extremely high mutation rate of 10^{-2} (H=11.2; *p<0.001* and H=4.05; *p=0.044* respectively).

If the fragment populations are considered to be one large population of 85 individuals, 15 alleles would be expected at loci mutating at a rate of 10^{-2}. However, a more realistic, but still fast, mutation rate of 10^{-3} only predicts four alleles at each locus. More alleles than that are seen at four of the five loci even though the actual effective population size is likely to be much less than 85 individuals. 24 alleles in total are seen in all of these populations, with a mutation rate of 10^{-3} only 19 would be expected, therefore there are more alleles within the metapopulation as a whole than would be expected if it was an isolated unit.

A: expected number of alleles

population	size	at each locus			at five loci		
		$\mu = 10^{-2}$	$\mu=10^{-3}$	$\mu=10^{-6}$	$\mu = 10^{-2}$	$\mu=10^{-3}$	$\mu=10^{-6}$
BZ	17	3.27	1.27	1.00	16.35	6.35	5.01
GH	14	2.82	1.21	1.00	14.08	6.06	5.01
KE	4	1.38	1.04	1.00	6.90	5.21	5.00
L	3	1.25	1.03	1.00	6.27	5.14	5.00
T	10	2.22	1.14	1.00	11.10	5.70	5.01
AWW1	13	2.67	1.19	1.00	13.33	5.97	5.01
AWW2	13	2.67	1.19	1.00	13.33	5.97	5.01
AWW3	11	2.37	1.16	1.00	11.84	5.79	5.01
ALL	85	15.09	3.79	2.02	75.43	18.93	10.1

B: observed number of alleles

population	size	at each locus					at all loci
		RSμ1	RSμ3	RSμ4	RSμ5	RSμ6	total
BZ	16	6	6	4	2	2	20
GH	14	4	5	4	3	2	18
KE	4	4	4	2	2	2	14
L	3	4	2	2	1	2	11
T	10	4	2	6	2	2	16
AWW1	12	4	4	4	4	1	17
AWW2	12	4	3	4	4	2	17
AWW3	11	3	2	3	3	2	13
ALL	82	6	6	6	4	2	24

Table 5.5: (A) The expected number of alleles predicted by the infinite alleles model and **(B)** the observed number of alleles at microsatellite loci in the Belgian fragment populations in 1996.

All these comparisons show that all the loci need to be mutating at a greater rate than 10^{-3} to explain the number of alleles seen in the populations, which seems unlikely. As will be discussed in section 5.4.1, these loci show levels of polymorphism similar to other rodent microsatellites indicating that they mutate at similar rates which are typically less than 10^{-3}.

5.3.6 Gene diversity excess

If the majority of loci examined in a population show a greater gene diversity than that predicted for a population under mutation-drift equilibrium, then this can be taken as evidence that the population has suffered a bottleneck in its recent history. The computer package, BOTTLENECK, generates the expected gene diversities under different models of mutation and compares them with those seen in the populations. Table 5.6 and figure 5.6 show the results of this test for a bottleneck for each of the Belgian populations. Table 5.6 shows the proportion of loci in each population that show an excess of gene diversity and the results of the Wilcoxon sign-rank test, carried out by the BOTTLENECK package and used to test the significance of these proportions.

population		proportion of loci showing an excess of gene diversity			probability of the result (Wilcoxon sign-rank test)		
		IAM	TPM	SMM	IAM	TPM	SMM
Brede Zijpe		0.8	0.6	0.4	0.06	0.625	1
Gasthuisbos		0.6	0.6	0.4	0.16	0.81	0.81
Kegelslei		0.2	0.6	0.2	1	1	1
Luisbos		0.6	0.8	0.6	-	-	-
Tallaarthof		1	0.8	0.6	0.03	0.03	0.06
Antwerp water works	area1	0.4	0.4	0.4	0.18	0.87	0.31
	area2	0.8	0.8	0.4	0.16	0.22	1
	area3	1	1	1	0.03	0.03	0.03
Merodese Bossen		0.4	0.4	0.4	-	-	-
Peerdsbos		0.6	0.6	0.6	-	-	-

Table 5.6: The results of a comparison of the observed and expected gene diversities in each locus for each fragment population during 1996. The proportion of loci in each population where the gene diversity observed is greater than that predicted by each mutation model is shown and the results of the Wilcoxon sign-rank test for significance carried out by the BOTTLENECK (v.1.2.02) program. This is the probability that the results would occur by chance if the null hypothesis that there is no difference between the observed and expected values is correct. It should be noted that the Wilcoxon sign-rank test was carried out using observed gene diversity estimates that differ slightly from those used to generate the proportions and the graph in figure 5.6 (see text) for the fragment populations. No significance test results are given for L, MB and P as only four loci are polymorphic in these populations so significance cannot be tested reliably.

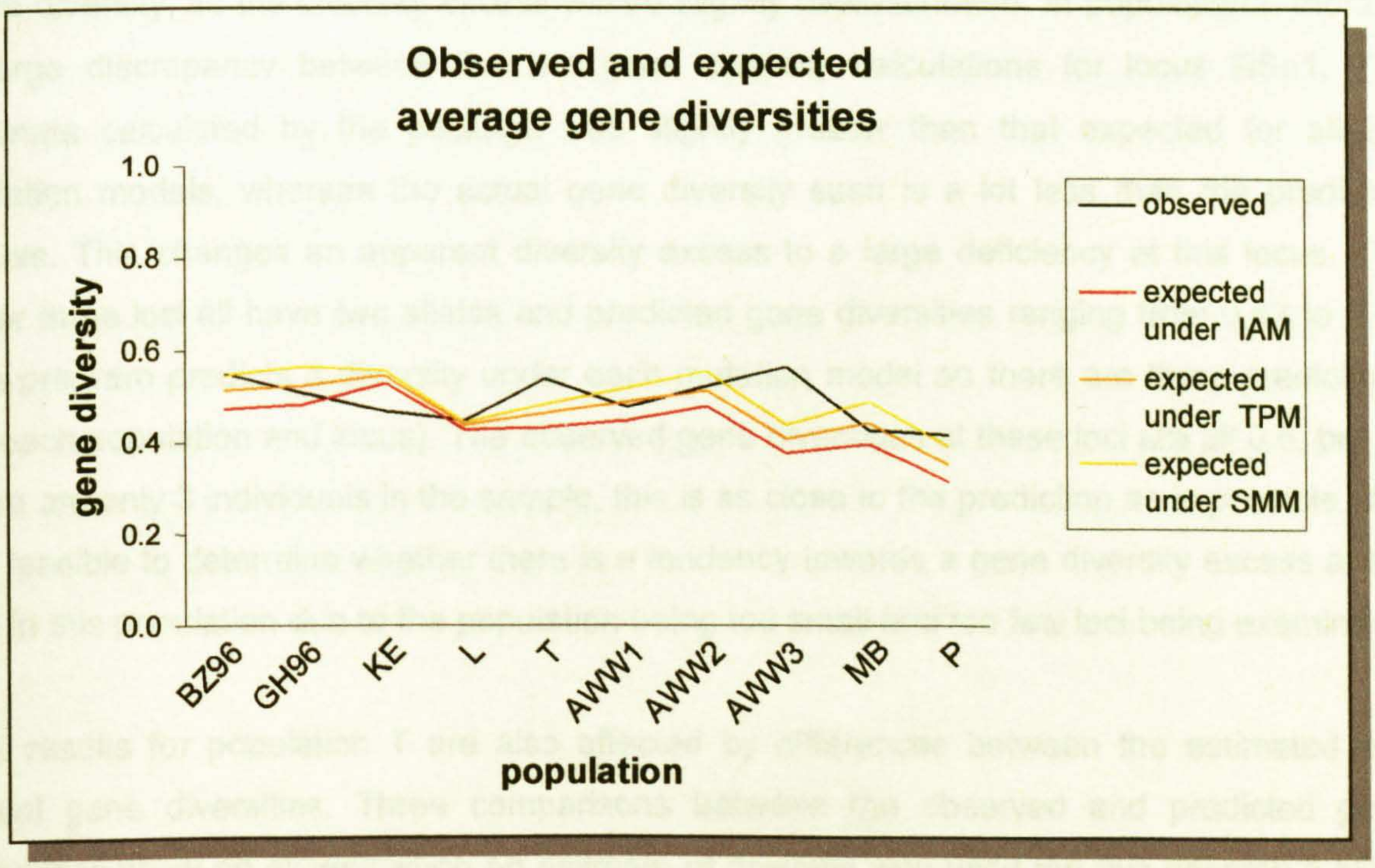


Figure 5.6: A graph showing the observed average gene diversity and expected average gene diversities under each mutation model in the Belgian populations.

The BOTTLENECK program calculates the observed gene diversity from the genotypic data for each population according to the equation appropriate for a sample taken from a population (equation 2, section 5.1.2). For the fragment populations (all except MB and P) it is possible to correctly calculate the actual gene diversity rather than an estimate, as all the individuals in the populations have been sampled (equation 1, section 5.1.2). The data shown for the proportion of loci in excess (table 5.6) and the average gene diversities (figure 5.6) were calculated by hand, for the fragment populations, by calculating the actual gene diversity. Unfortunately, it is not possible to change the gene diversity estimates calculated by the program and used in the significance test, therefore the significance of the results was tested using gene diversity estimates for these populations rather than accurate calculations.

The reliability of the significance tests are doubtful in any case as only five loci are used; more would be needed to reliably test the results. This is especially true for the populations of Luisbos, Merodese Bossen and Peerdsbos, where one locus is monomorphic. This leaves only four loci to indicate gene diversity excess and it would be impossible to statistically prove such a tendency with such a small sample, for this reason the results of the significance tests for these populations are not shown in table 5.6.

In all cases, the gene diversity estimate used by the package was greater than the actual gene diversity, so the diversity excess will be slightly overestimated. In population L there is a large discrepancy between the two gene diversity calculations for locus RSμ1. The estimate calculated by the package was slightly greater than that expected for all the mutation models, whereas the actual gene diversity seen is a lot less than the predicted values. This changes an apparent diversity excess to a large deficiency at this locus. The other three loci all have two alleles and predicted gene diversities ranging from 0.43 to 0.46 (the program predicts a diversity under each mutation model so there are three predictions for each population and locus). The observed gene diversities at these loci are all 0.5, but as there are only 3 individuals in the sample, this is as close to the prediction as is possible. It is not feasible to determine whether there is a tendency towards a gene diversity excess at the loci in this population due to the population being too small and too few loci being examined.

The results for population T are also affected by differences between the estimated and actual gene diversities. Three comparisons between the observed and predicted gene diversities show an excess when an estimate of diversity was used but this became a slight deficiency when the accurate measure was used. Despite this, three loci out of five show an excess and at two loci that excess is very large. Both RSμ3 and RSμ6 have only two alleles

in this population, which leads to gene diversity predictions of between 0.27 and 0.32 but the actual observed gene diversity at these loci is 0.49 and 0.46. Overall, 11 of the 15 comparisons show an excess. It seems that this population does have a real tendency towards an excess of gene diversity at microsatellite loci.

The results for AWW3 were not changed by the differences between the estimated and actual gene diversity calculations. In this population, all the loci under all the models show a large difference between the observed and expected value of gene diversity, with all the loci having more than expected. Therefore, this population has a significant excess of gene diversity. The significance of the excess seen in Peerdsbos could not be tested as only four loci are polymorphic in this population. However, three of the four loci show a dramatic excess of gene diversity: RSμ1 and RSμ4 have four alleles in the population with predicted gene diversities of between 0.48 and 0.6 (the large range is probably due to the larger sample size included) and locus RSμ3 has only two alleles with predicted diversities of 0.23 to 0.26, the observed gene diversity estimates were 0.72 and 0.7 for RSμ1 and RSμ4, and 0.48 for RSμ3. The size of this excess seen at three out of four loci suggests that there is also a genuine tendency towards a greater gene diversity than expected in this population.

The graph in figure 5.6 illustrates these results. This is a plot of the observed average gene diversity (calculated by hand for the fragment populations and estimated by the program for populations MB and P) against that predicted by each of the mutation models. Only populations L, T, AWW3 and P have more gene diversity than expected but this difference is very slight in L. The greatest difference is found in AWW3 but there is a real and visible excess in T and P as well.

The bottleneck test was repeated, omitting the data for RSμ3 due to the possibility of homoplasy leading to misleading results and, generally, the same patterns of gene diversity excess were seen. The significance tests were even less reliable when fewer loci were included in the tests and could not be used. The inclusion of RSμ3 did not distort the results but did increase the usefulness of the test.

5.3.7 Population structure

The population structure of the Belgian fragment populations was investigated by calculating θ and ρ, estimators of F_{ST} and R_{ST}. Calculations were performed across all the fragment populations and for all possible population pairs; the results of the pairwise analyses are given in table 5.7. The probabilities that the results differ significantly from zero are given in the table but, as multiple comparisons were carried out using the same data, a Bonferroni correction using the Dunn-Šidák method (section 3.1.2.2) was carried out to calculate the significance level for each comparison. The comparisons that were found to be significant after this correction was carried out are highlighted in yellow. The analysis was repeated without RSμ3 and no comparisons were found to be significant. No differences were seen in the overall patterns of the two sets of results, so this loss of significance is likely to be the result of a loss of power when the number of loci used is reduced, rather than reflecting distortion due to the inclusion of this locus.

The estimates across all populations are $\theta = 0.055$ ($p<0.0001$) and $\rho = 0.045$ ($p=0.004$). Both are highly significant indicating that there is structuring within this group of populations. As can be seen from the tables, only two pairwise comparisons, both estimates of θ, are significant and no estimates of ρ are significant. Overall, it can be concluded that there is some differentiation between some of the populations, resulting in significant F_{ST} and R_{ST} estimates across all the populations, but that differentiation is generally at a low level and not significant.

There are some patterns of interest within these results. The populations occupying the areas of the Antwerp water works have a linear layout along the row of reservoirs. The gene pools of AWW1 and AWW2 are likely to resemble each other as they are only divided by a track and the estimates of θ and ρ for pairwise comparisons of these populations are indeed very low (0.0031 and 0.001 respectively) confirming that they are very similar. Yet, there is no evidence that they are in fact one panmictic population as the estimates are positive, indicating that there is a small degree of differentiation between them. AWW3 is separated from AWW2 by a built up area but, despite this, there is no significant differentiation between the two populations, although they are less related than AWW2 is to AWW1. As may be expected, AWW1 and AWW3 show the greatest amount of differentiation.

A: the results of the pairwise F_{ST} analysis

		BZ	GH	KE	L	T	AWW1	AWW2
Gasthuisbos		-0.01 *(0.675)*						
Kegelslei		0.022 *(0.3)*	0.001 *(0.291)*					
Luisbos		0.065 *(0.113)*	0.061 *(0.061)*	-0.052 *(0.633)*				
Tallaarthof		0.046 *(0.034)*	0.014 *(0.209)*	0.019 *(0.377)*	0.014 *(0.467)*			
Antwerp water works	**area1**	0.103 *(<0.001)*	0.077 *(0.004)*	0.079 *(0.065)*	0.21 *(0.003)*	0.118 *(<0.001)*		
	area2	0.069 *(0.005)*	0.048 *(0.039)*	0.092 *(0.066)*	0.163 *(0.021)*	0.049 *(0.076)*	0.0031 *(0.49)*	
	area3	0.055 *(0.019)*	0.046 *(0.046)*	0.062 *(0.157)*	0.077 *(0.174)*	0.024 *(0.247)*	0.092 *(0.007)*	0.028 *(0.218)*

B: the results of the pairwise R_{ST} analysis

		BZ	GH	KE	L	T	AWW1	AWW2
Gasthuisbos		-0.01 *(0.594)*						
Kegelslei		0.047 *(0.318)*	-0.035 *(0.59)*					
Luisbos		0.016 *(0.444)*	0.026 *(0.247)*	0.078 *(0.251)*				
Tallaarthof		-0.025 *(0.816)*	-0.029 *(0.772)*	-0.017 *(0.469)*	0.012 *(0.343)*			
Antwerp water works	**area1**	0.159 *(0.002)*	0.097 *(0.013)*	0.109 *(0.087)*	0.23 *(0.033)*	0.144 *(0.002)*		
	area2	0.065 *(0.06)*	0.024 *(0.176)*	0.071 *(0.148)*	0.138 *(0.094)*	0.07 *(0.05)*	0.001 *(0.418)*	
	area3	0.023 *(0.212)*	-0.005 *(0.409)*	-0.016 *(0.421)*	-0.035 *(0.532)*	-0.016 *(0.576)*	0.109 *(0.019)*	0.067 *(0.04)*

Table 5.7: The results of population differentiation analysis on all possible pairs of fragment populations. Table **A** gives the calculated values of θ, estimates of F_{ST}, and table **B** gives values of ρ, an estimator of R_{ST}; the probabilities that these values are significantly greater than zero are given in brackets. Comparisons that are significant after a Bonferroni correction is carried out are highlighted in yellow, these are not significant when the analysis is carried out without the inclusion of locus $RS\mu3$.

The populations of BZ, GH, L and T also show a linear layout (see figure 5.10) but no relationship between distance and differentiation can be seen. This was tested using the isolation by distance analysis method of comparing F_{ST} /(1- F_{ST}) with the natural logarithm of distance, developed by Rousset (1997). The relationship of both F_{ST} and R_{ST} with distance for these four populations (involving six pairwise comparisons) is illustrated in figure 5.8 and it can be seen that there is no correlation (Mantel test on the results for F_{ST}, *p=0.67*, and R_{ST}, *p=0.7*).

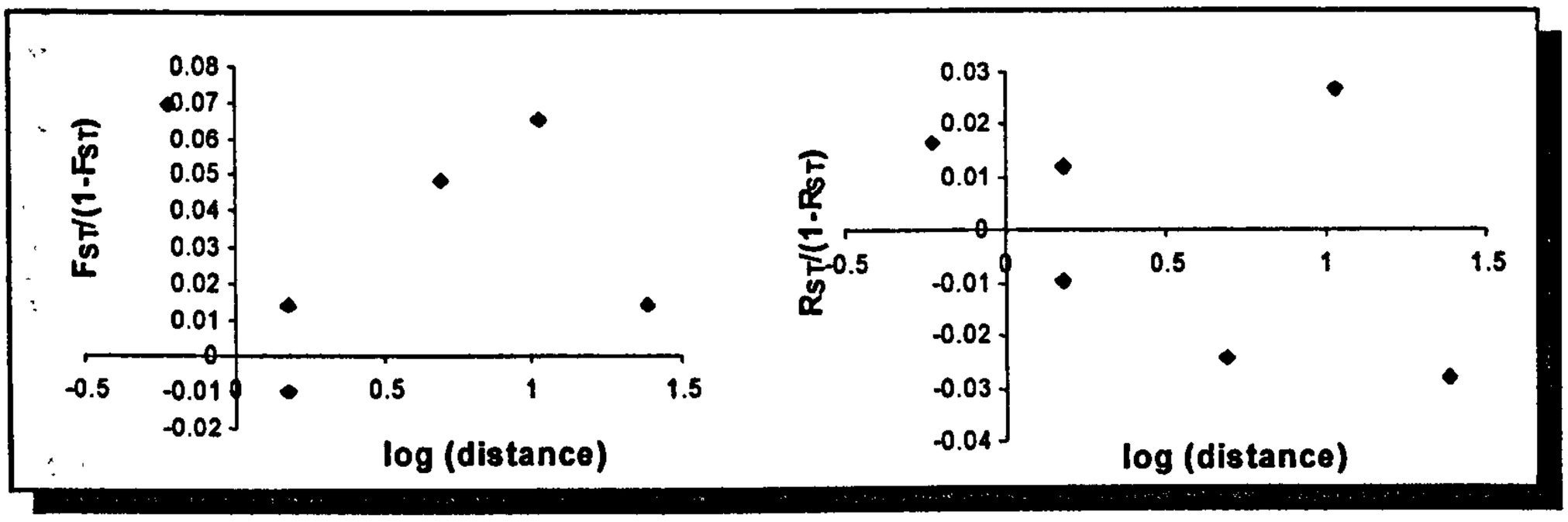


Figure 5.8: A graph showing the relationships between transformed population pairwise F_{ST} and R_{ST} estimates and the natural logarithm of the distance (km) between the populations of Brede Zijpe, Gasthuisbos, Luisbos and Tallaarthof.

Comparisons between the three AWW populations and the other fragment populations (referred to as the "core" populations) shows that, except in the one case of the comparison with KE by estimating θ, AWW1 is the most differentiated from the other populations and AWW3 the least. This pattern may be due to differences in the amount of migration between the populations, less migration may occur between AWW1 and the core fragments than between them and AWW2 and 3, and AWW3 may exchange more individuals with the core populations than either AWW1 and AWW2. These possible differences in migration rates are reflected in the estimates of Nm, shown in table 5.7. The three populations are all about equally close to the other fragment populations (see figure 1.6) suggesting that migration between AWW3 and the core populations may be easier, perhaps due to differences in the habitats between the populations such as presence or absence of corridors.

The estimates of Nm generated by both analyses are given in table 5.7. This method of estimating migration between two populations is only reliable when there is little movement between the populations. In this case, these values are not very useful as estimates of migration, some of them are quite ridiculous and in many cases they exceed the number of individuals in the populations. However, they are a comparable representation of the amount of differentiation estimated by the F_{ST} and R_{ST} analysis. In theory, F_{ST} estimates are likely to

overestimate the amount of genetic similarity and therefore may be expected to overestimate the amount of migration (Slatkin 1995). If this is true of this study, the Nm values estimated from θ would be expected to be larger than those estimated from ρ. A comparison of the Nm estimates shows that the estimate from the F_{ST} analysis is only larger than that from the R_{ST} analysis in 11 out of the 20 possible comparisons. Therefore the F_{ST} analysis does not appear to be indicating less genetic differentiation than the R_{ST} analysis. Indeed, the only significant pairwise comparisons were for estimates of F_{ST}.

		BZ	GH	KE	L	T	AWW1	AWW2	AWW3
Brede Zijpe			-	10.9	3.6	5.2	2.2	3.3	4.3
Gasthuisbos		-		208	3.85	17.4	3	4.9	5.2
Kegelslei		5.1	-		-	12.6	2.9	2.4	3.8
Luisbos		15.4	9.3	2.3		18	0.9	1.3	3
Tallaarthof		-	-	-	20.4		1.8	4.8	10.2
Antwerp	area1	1.3	2.3	2	0.8	1.5		80.4	2.4
water	area2	3.6	10.1	3.3	1.6	3.3	175.4		8.6
works	area3	10.4	-	-	-	-	2	3.5	

Table 5.7: The estimates of Nm generated from the estimates of θ (top half of the matrix) and ρ (bottom half of the matrix). There are some missing values as it is not possible to calculate Nm from negative values of θ and ρ, these population pairs can be considered to be panmictic.

The lack of bias in the analysis is perhaps an indicator of the relatedness of these populations. Slatkin (1995) showed that F_{ST} does not underestimate genetic differentiation when the time since common ancestry is small, either due to recent separation of the populations or a high degree of migration between the populations. Under these circumstances there has either not been enough time for differences in the mutation process to have a noticeable effect on the two estimates, or migration prevents the differentiation of the populations, so removing the effects of mutation.

The isolation by distance analysis is illustrated by the graph in figure 5.9. The distribution of the points on the graph appears to be quite similar but the F_{ST} results are in fact more tightly distributed. The results of the two Mantel tests on the matrices of the transformed data are quite different, the probability of the observed correlation for the F_{ST} results being *0.03* and for the R_{ST} results, *0.29*. Estoup *et al.* (1998) found a similar contrast in the results of isolation by distance analysis on F_{ST} and R_{ST} estimates when analysing geographic differentiation in the brown trout (*Salmo trutta*). They obtained probabilities of correlation from

F_{ST} estimates of between *0.01* and *0.04* whereas analysis using R_{ST} estimates produced probabilities of correlations between *0.19* and *0.33*. The differences in these probabilities are of a very similar scale to the differences in the probabilities of correlation calculated here. Estoup *et al.* (1998) attributed the lack of agreement to the higher variance in pairwise R_{ST} estimates compared to those of F_{ST} leading to a weaker correlation. In this study, the variance in ρ estimates is 0.007 and the variance in θ estimates 0.006, therefore the disparity in the results of the isolation by distance analysis may also be attributed to differences in the variance of the two genetic differentiation measures.

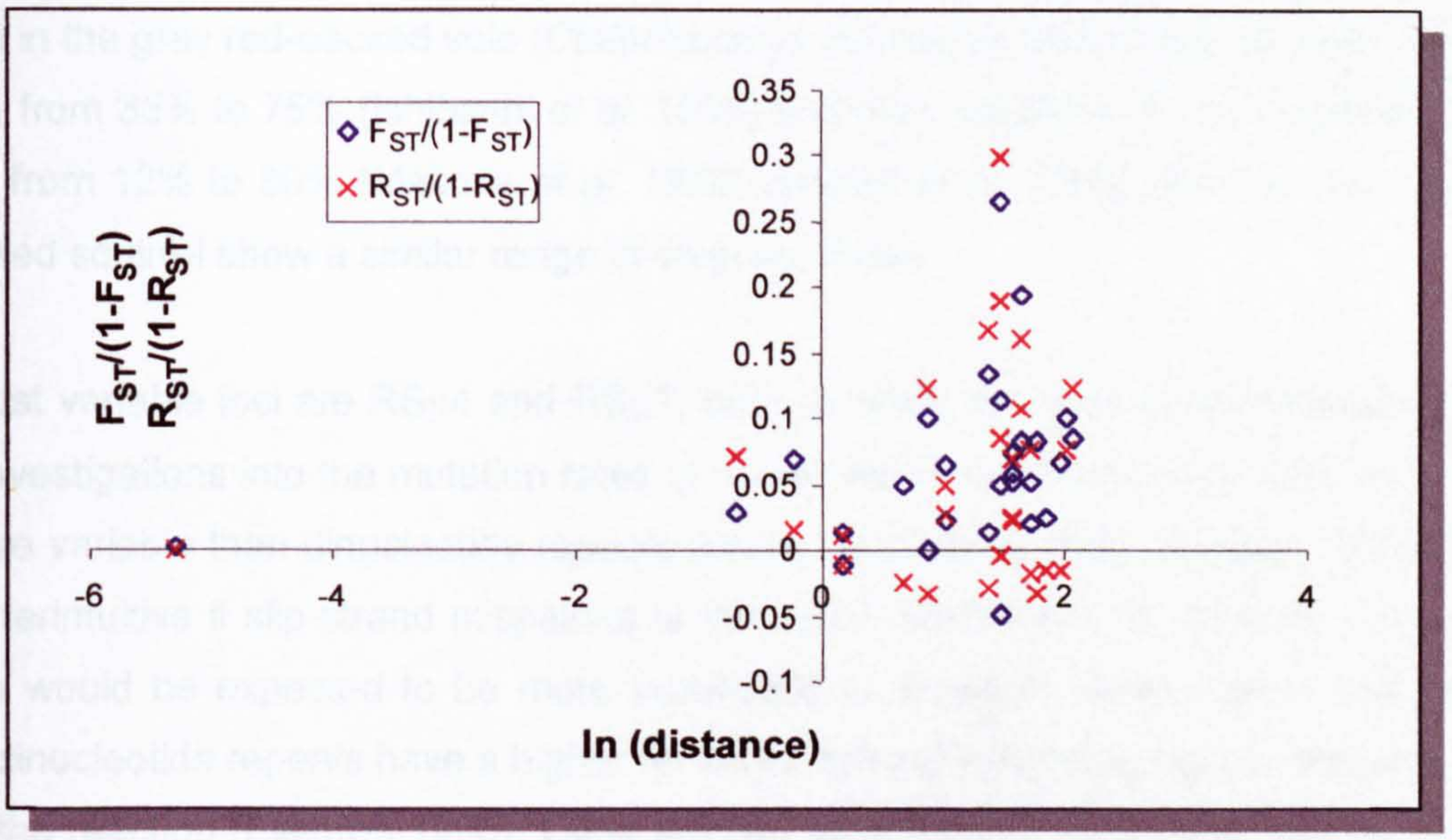


Figure 5.9: The graph produced by plotting $F_{ST}/(1- F_{ST})$ and $R_{ST}/(1- R_{ST})$ against the natural logarithm of distance (km) for all pairwise comparisons of the fragment populations.

Given the likelihood that the estimates of ρ are more accurate for assessing the differentiation between populations when microsatellites are used as markers, it would be difficult to support the argument for isolation by distance using this analysis. The differentiation of two red squirrel populations may be influenced by the distance between them but distance is not having a significant effect.

5.4 DISCUSSION

5.4.1 The microsatellite loci

The five microsatellite loci developed for this study were found to have useful levels of polymorphism comparable to other rodent microsatellites. The gene diversities of the loci in the most variable German sample ranged from 43.4% to 82.6%. Rodent microsatellites show a huge range in variability: the gene diversities of nine loci developed for the Columbian ground squirrel (*Spermophilius columbianus*) ranged from 22% to 80% (Stevens *et al.* 1997), five loci in the grey red-backed vole (*Clethrionomys rufocanus bedfordiae*) showed diversities ranging from 35% to 75% (Ishibashi *et al.* 1995) and microsatellites in three species of mice ranged from 12% to 89% (Makova *et al.* 1998; Wooten *et al.* 1999), the five loci developed for the red squirrel show a similar range of diversity levels.

The most variable loci are RSμ4 and RSμ1, both of which are pure tetranucleotide repeats. Early investigations into the mutation rates of microsatellite loci have found that tetra repeats are more variable than dinucleotide repeats (Weber and Wong 1993; Ellegren 1995) but this is counterintuitive if slip-strand mispairing is the usual mechanism of mutation. Dinucleotide repeats would be expected to be more vulnerable to slippage events, as is seen in PCR where dinucleotide repeats have a higher tendency towards stuttering due to slippage events during the reaction (Ellegren 1995). Other studies have found dinucleotide repeats to be the most mutable loci (Chakraborty *et al.* 1997) but there have been very few thorough studies into the relative rates of evolution in microsatellite loci and it may be naïve to assume that repeat length would be the only factor influencing variation levels.

Reasons to doubt this simple relationship can be illustrated by looking at the differences, for example, between the variability of the loci in the geographically distinct Belgian and German samples. In both samples RSμ1 and RSμ4 are the most variable but in the German sample the least variable are RSμ3, the compound locus (confirming the observation that compound structure tends to stabilise loci (Jarne and Lagoda 1996)), and RSμ5, the pure dinucleotide repeat. This pattern is consistent with previously observed patterns for other loci (Jarne and Lagoda 1996). However, in the Belgian sample RSμ3 is more variable and the least variable locus is RSμ6, the trinucleotide repeat. The Belgian populations would have been affected by the random process of drift, so the variation at the microsatellite loci in these populations may be more influenced by the chance loss of alleles rather than biases in the mutation processes.

The random effects of drift, which weaken the correlation between heterozygosity and mutation rate (Li 1979), explain the inconsistencies seen in the variability of different loci in different Belgian populations. For example, RSμ5 is the least variable locus in five of the eight Belgian fragment populations but is the most variable in AWW1 and 2. These inconsistencies also emphasise the need to look at several loci when investigating the genetics of small populations to get an accurate overall impression of the state of each population.

Alleles at locus RSμ3 may be susceptible to homoplasy due to its compound structure, with a dinucleotide repeat being contiguous to a tetranucleotide repeat. However, the exclusion of RSμ3 from the analysis did not lead to any noticeable changes in the results, apart from a reduced power for some of the statistical tests. Deviation from Hardy-Weinberg equilibrium was only seen in two of the populations at this locus, including one of the small fragment populations in which such a result could be attributed to chance. No evidence for homoplasy has been seen and, as the allele sizes observed all differ by units of two base pairs in size, it may be the case that most of the variation at this locus is attributable to changes in the number of dinucleotide repeats. This was shown to be the case in a locus of similar structure found in humans (locus D11S527, Hauge and Litt 1993) where sequence data from 12 individuals showed that the length variations all occurred in the region of the dinucleotide repeat. If all the variation is occuring in one repeat region then no effects of homoplasy would be seen. Overall, a negligible amount of homoplasy may be resulting in the misdesignation of a few alleles, but it has not had any notable effect on the analysis carried out in this study.

5.4.2 The genetic variation within the populations

The German sample from Waldhäuser contains the most genetic variation, but the difference between the variation levels found in the German and Belgian populations is not significant. This population is large and in an extensive area of habitat that is not known to have experienced any population size reductions, at least in its recent history. This population is useful as a comparison for the populations in the fragmented habitat of Belgium. The total Belgian sample and the German sample share many alleles, perhaps reflecting a historical common ancestry, but they are highly diverged in the present day.

If the fragment populations are considered to be one metapopulation, a comparison with the German sample shows that the metapopulation as a whole contains almost as many alleles as the German sample (24 as opposed to 26) but the sample taken from the German population is much smaller than that from the totally sampled metapopulation. All the alleles

in the metapopulation have probably been detected (only a few individuals are missing from the analysis) but there may be many undetected alleles in the Waldhäuser population as only 27 individuals were screened. However, gene diversity and observed heterozygosity are less in the metapopulation than the German sample (56.5% and 53.6% compared with 62.2% and 63.7% in the German sample), suggesting that the whole Belgian metapopulation is less variable that the German population, but that the difference is slight.

Of the Belgian populations, it is the fragment population of Brede Zijpe that is the most variable closely followed by those of Gasthuisbos and Tallaarthof. The large populations of Peerdsbos and Merodese Bossen are amongst the least variable. The lack of variation in these large populations is surprising and not easily explained: they may be more isolated from other red squirrel populations and experience less migration to offset the effects of drift, or they may have been more greatly affected by an historical bottleneck.

There is some evidence that Peerdsbos has recently experienced a bottleneck: three out of four loci have much higher levels of gene diversity than would be expected under all three mutation models investigated and the sample from this population only contained one allele at a frequency of less than 0.1. When a bottleneck occurs it is the low frequency alleles that are most easily lost so a population size reduction may be indicated by a loss of rare alleles; the other Belgian populations typically contained around 5 or 6 low frequency alleles. Merodese Bossen showed no evidence of a recent bottleneck although that does not mean that one has not occurred from which it is recovering. If Merodese Bossen exchanges more individuals with surrounding populations than Peerdsbos it would recover from a bottleneck more quickly as new rare alleles may enter the population dissipating the gene diversity excess.

Tallaarthof shows these same characteristics of gene diversity excess and a low frequency of rare alleles, but to a lesser degree. During 1995 and 1996 eleven adult squirrels arrived in this area boosting the population from what was just a couple of individuals to ten permanent residents in 1996. Therefore this is in effect a founder event caught in the act by this study. The sample examined for 1996 represents the founding population during its first generation and as is predicted by theory, an excess of gene diversity can be seen. Founder events are chance sampling events that will tend to lead to a loss of alleles without a reduction in expected heterozygosity, but the element of chance dictates that not all loci will necessarily show this trend. Only five loci were examined in this study, so it is not possible to prove statistically that this population shows a gene diversity excess, although three of the five loci showed an excess and at two it was very large.

A tendency towards an excess of gene diversity was also detected for Luisbos by the BOTTLENECK program, but with such a small population and a small sample of alleles, it is not possible to determine whether this is real. However, the demographic records for this population show that it increased in size during 1995 and 1996, although only four individuals could be reliably identified as permanent occupants. The results for the populations of T and L are very encouraging, as they show that the methods employed to identify higher levels of gene diversity than would be expected have correctly identified what is effectively two founder events, despite the low number of loci available for examination.

The Antwerp water works area 3 population also contains low levels of allelic diversity relative to heterozygosity and it is the only population that showed a testable and significant gene diversity excess. AWW3 is also lacking in low frequency alleles, so it seems likely that it too has recently experienced a bottleneck or that this population is the result of a recent colonisation event (founder event). Demographic records are not available for this population before 1996, so it is not possible to determine what has caused the observed patterns. Given the rapid increase in numbers experienced by red squirrels in this region in recent years, it seems likely that this population is a new one recently founded in this area.

The evidence that a bottleneck or founder event has occurred is transitory and populations recover at different speeds depending on their size and the amount of immigration they experience. Immigration introduces new alleles to the populations, increasing the number of alleles without affecting the level of heterozygosity. This offsets the gene diversity excess that results from a loss of alleles in the first place, so populations experiencing high rates of migration may not show the tell-tale signs of a bottleneck in the form of gene diversity excess. To be able to draw more definite conclusions from the presence of gene diversity excess at the loci in the populations, a larger number of loci would need to be examined, especially when the population sizes are so small making them extremely vulnerable to random outcomes during sampling events.

Despite any possible historical bottlenecking, all the fragment populations contain more alleles than would be expected for populations of their size at mutation-drift equilibrium. The expected number of alleles was calculated using Ewens' formula (1972) which assumes the loci are mutating under the infinite alleles model. Microsatellite loci are more likely to mutate in a way closer to the stepwise-mutation model under which the number of alleles in a population reaches a plateau at a relatively low level. The number of alleles predicted by Ewens' formula is therefore an overestimate which could potentially greatly exceed the number of alleles that would be predicted by the SMM. As well as this, the predicted number

of alleles was calculated from the actual population size when the effective population size is the correct measure. The effective population is likely to be smaller than the actual population, but this is not likely to have much of an effect on the allele number predictions when the population sizes are so small. Even with all this potential overestimation, all the fragment populations were found to contain more alleles than expected for loci mutating at a faster rate than 10^{-3} indicating that there must be an alternative to mutation as a source of alleles for these populations.

An excess of alleles relative to gene diversity would be expected in an expanding population (Maruyama and Fuerst 1984; Maruyama and Fuerst 1985) and would be visible as a gene diversity deficiency. Just as gene diversity excesses are detectable using the BOTTLENECK program, a gene diversity deficiency would be detected as a very low proportion of loci showing an excess (therefore a large proportion showing a deficiency). Examination of table 5.6 shows that no consistent pattern of gene diversity deficiency can be seen. The elevated levels of allelic diversity in the fragment populations are matched by high levels of heterozygosity and gene diversity and are not a temporary result of population expansion; even the populations showing gene diversity excess have more alleles than predicted.

5.4.3 Population structure

The F_{ST} and R_{ST} estimates across all the fragment populations are both highly significant indicating that there is differentiation among the populations. Slatkin (1995) observed that F_{ST} would tend to underestimate genetic differentiation to an increasing degree as the populations diverged. In this case the opposite pattern is seen as the estimate of F_{ST} is larger than the estimate of R_{ST} ($\theta = 0.055$ and $\rho = 0.045$). F_{ST} does not take differences in allele size into account and as populations diverge, loci mutating under the stepwise mutation model will show increasing differences in allele size which is included in the R_{ST} estimate of differentiation. Therefore, in highly diverged populations, the R_{ST} estimate will tend to be larger than the F_{ST}. Here, F_{ST} produces the greater estimate of divergence indicating that, whilst there is some differentiation in allelic composition, there is little differentiation in allele size. F_{ST} may be expected to be larger or similar to R_{ST} when there is a low average coalescence time (time to the most recent common ancestor) for the alleles as there is then little time for differences in the mutation process to take effect (Estoup *et al.* 1998). Short coalescence times may result from a high migration rate between the populations or a short time since the populations were separated. Either or both explanations may apply to these fragment populations.

Although, overall the populations show some divergence, little or no significant differentiation is detected in pairwise comparisons. This also could be due to a short time since the populations were separated, high rates of migration or a combination of both. There is no evidence for a correlation between the amount of differentiation between two populations and the distance between them, even in the four linearly arranged populations of Tallaarthof, Luisbos, Brede Zijpe and Gasthuisbos. Therefore, if the amount of differentiation between the populations is dictated by the number of migrants they exchange, the rate of migration cannot be determined by distance alone.

A lack of correlation between gene flow and geographic distance was also found between populations of another small mammal, the hedgehog (*Erinaceus europaeus*) by Becher and Griffiths (1998). Their study was on a similar small scale to this study as all the populations analysed were located within a 15km radius in a highly fragmented landscape in Oxfordshire, UK. They also concluded that other factors than geographic distance, such as geographic barriers, probably affect gene flow. The habitat between the populations and the ease with which it can be crossed will have an effect on the migration rate; the presence or absence of corridors and barriers may be crucial. The relative qualities of the habitat patches will also affect the number of immigrants as higher quality patches will attract more dispersing individuals.

The differentiation between the populations may not only be determined by migration. It will also be influenced by time since common ancestry and the demographic history of the populations. If fragmentation has simply divided one continuous population into several separated populations at the same point in time then migration and the ongoing effects of drift will be the only factors influencing their divergence. However, it is more likely that the populations became isolated at different times and that time since separation will effect their relatedness. It is also possible that some of the populations result from recent colonisation events, either due to previous extinction of the local population or to new occupation of habitat patches. The differentiation of the populations would then be determined by the time since the founder event occurred and which population or populations were the source of the immigrants.

5.4.4 Migration within the metapopulation

Migration is a frequent occurrence between these fragment populations and may have an important effect on the genetics of the populations. As was discussed in section 1.4.2, red squirrels can disperse over long distances; several kilometres may be covered so, in theory, all the populations included in this study may be reached by squirrels from any of the others. Five migration events were tracked amongst the populations of Brede Zijpe, Gasthuisbos, Kegelslei, Luisbos and Tallaarthof during 1995 and 1996 (the Antwerp water works populations were only studied during 1996 and no migration events were followed) and these are illustrated in figure 5.10. Two dispersal events were tracked between Brede Zijpe and Gasthuisbos, another from Tallaarthof to Brede Zijpe, one into Brede Zijpe from outside the study populations and one emigration from Kegelslei to elsewhere (Goedele Verbeylen, pers. com.). These are only the few migration events that could be followed by radio tracking or because a previously tagged squirrel was identified in a new location.

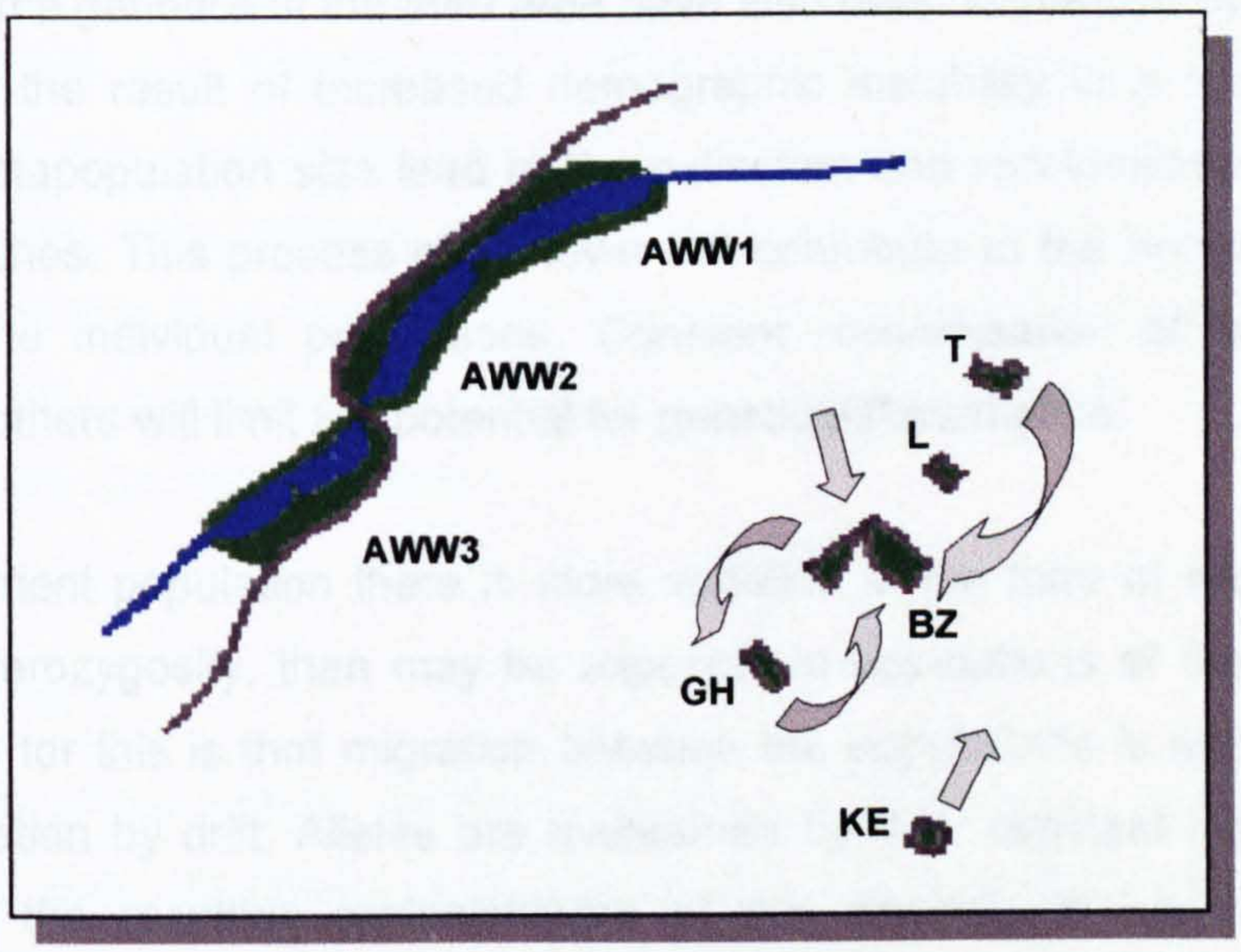


Figure 5.10: The known migration events that occurred between the Belgian fragment populations during 1995 and 1996.

Examination of the demographic records of these populations during these two years shows that 17 squirrels disappeared from these populations during this time and 24 adult squirrels arrived from unknown locations. The disappearances may be due to dispersal or to the death of the squirrel and unless the carcass of the squirrel is found or the individual appears in another population it is not possible to distinguish these fates. Of the recorded immigration events, 11 occurred in Tallaarthof and may represent the colonisation of a new area as the numbers of red squirrels in northern Belgium increase. These additional figures show that a lot of migration probably occurs between the Belgian red squirrel populations, especially as the metapopulation currently expands and new areas are colonised.

5.4.5 The effects of habitat fragmentation

All the Belgian populations individually and as one metapopulation, show less variation, both in allelic diversity and heterozygosity, than the German sample from the large Waldhäuser population, although this difference is quite small. It is not possible to tell from these results whether the slightly reduced variation in the Belgian populations is due to the current effects of reduced population sizes or historical processes such as a previous bottleneck.

There is some evidence that at least three of the fragment populations have been affected by bottlenecks or founder events recently. For two of the populations, demographic data indicate that there has been a large amount of immigration into the areas in the years immediately prior to and during this study. These populations at least are showing the reduction in allelic diversity, but elevated levels of heterozygosity associated with founder events and, given the current expansion of the red squirrel population in northern Belgium, it seems likely that the genetics of the third area have also been influenced by a founder event. Such events are the result of increased demographic instability in a fragmented habitat; fluctuations in metapopulation size lead to the extinction and recolonisation of some of the small habitat patches. This process of turnover will contribute to the homogenisation of the gene pools of the individual populations. Constant recolonisation of some patches by immigrants from others will limit the potential for genetic differentiation.

Within each fragment population there is more variation in the form of allelic diversity, and consequently heterozygosity, than may be expected in populations of that size. The most likely explanation for this is that migration between the populations is replacing alleles lost from each population by drift. Alleles are maintained by their constant movement between populations and the resulting replenishment of the diversity in each population. The populations are likely to have a high rate of turnover in alleles as they will be lost from such small populations by drift at a rapid rate and then replaced by migration from other populations. The lack of private alleles within the fragment populations further supports the hypothesis that migration is having a large role in the genetics of these populations. Migration is an homogenising factor resulting in the same alleles being found in many populations; with less migration there would be a greater chance of alleles only being found in the one population.

When the whole metapopulation is considered, it can be seen that it also contains more alleles than would be expected in a continuous population of its size (table 5.5). This is especially noticeable when it is remembered that at this scale the level of overestimation of

the expected number of alleles, due to an inappropriate mutation model being used and the effective population size being much smaller than the actual population size, is likely to have a greater effect. Despite the constant migration, alleles may still be maintained in the metapopulation by the physical separation of the populations, drift tending to remove particular alleles from some populations whilst maintaining them in others. Without the homogenising effects of migration the populations would eventually contain completely different sets of alleles. It is also likely that new alleles are entering the metapopulation from other nearby red squirrel populations. The populations included in this study do not represent all the red squirrel populations in the region, merely a sample of them.

The demographic data indicates that these populations could be experiencing a lot of migration, especially at the current time when squirrel numbers are increasing and there will be increasing pressure on young squirrels to disperse to new areas in order to find a suitable home range. It is therefore plausible that the levels of variation in the Belgian metapopulation, which are comparable to a large continuous population, are maintained by migration between the populations included in this study and other populations in the surrounding area. Migration between the populations will maintain the allelic diversity within the populations and limit their differentiation. Immigration into the study populations from the surrounding area may introduce new alleles, so maintaining and maybe even increasing the allelic diversity within the metapopulation.

5.4.6 Conclusions

The fragmentation of this habitat, which is forcing red squirrels to occupy small areas as small populations, does not appear to be leading to a loss of genetic variation. The Belgian squirrels are slightly less variable than the German squirrels, either because of the current population structure or the result of historical processes such as a bottleneck. However, these differences are only slight and all the fragment populations are in fact more polymorphic than would be expected for populations of their small size. The separation of the populations, with some migration between them, appears to be maintaining genetic diversity by retaining allelic diversity and, consequently, heterozygosity. As alleles are lost from a population by the effects of drift, they are replaced by the immigration of individuals from the surrounding areas carrying new alleles. The high levels of diversity seen in the metapopulation as a whole may be attributed to the separation of populations maintaining allelic diversity as differences between the populations and to immigration from other populations in the region.

CHAPTER SIX

DISCUSSION

6.1 The combined use of mitochondrial and nuclear markers

With the choice of molecular markers now available to the population geneticist, it is possible to make use of the features of different marker systems within one study. The mitochondrial genome has long been a popular choice when investigating relationships within and between populations because portions of the genome, notably the control region, evolve at a suitable rate to provide quantifiable variation at the population level. Other features, such as maternal inheritance and lack of recombination, mean it provides a unique picture of the history of a population that is more sensitive to demographic changes and is necessarily female biased. Microsatellites are nuclear markers that also show appropriate evolutionary rates for population studies. They are quick and easy to analyse and have, in recent years, tended to supersede mitochondrial DNA as the marker of choice for population studies. Being nuclear markers, they are less sensitive to population size changes and are equally inherited through both parents so they show no sex bias; they provide easily quantifiable data about both allelic diversity and heterozygosity.

Mitochondrial DNA is still a useful complement to microsatellites. The more different markers that can be applied to populations of interest, the more complete the picture of their history and relationships. With the recent developments in PCR and sequencing technology, both can be easily applied to a large numbers of samples. The differences in the evolution and inheritance patterns of these two markers can be used to answer specific population genetic questions.

The differences between the inheritance patterns have been most usefully applied to questions regarding sex specific dispersal patterns and kin structuring. Ishibashi *et al.* (1997) investigated kin related structuring within a population of grey-sided voles (*Clethrionomys rufocanus*) using both the maternally inherited mitochondrial genome, specifically sequences of the control region, and biparentally inherited microsatellites. Analysis of the mitochondrial marker revealed some clustering in females from the same lineage and the microsatellite data for females showed a negative correlation between relatedness and geographic distance. In males, a positive correlation was seen between relatedness and geographic distance and no clustering was seen within males from the same mitochondrial lineage. The combined sources of information from both the markers allowed the authors to conclude that there is sex-related kin structuring in this species, as females tend to cluster with relatives but males do not. They suggested several life history traits which may cause such patterns to develop.

Male-biased dispersal was assumed to be the norm for the African buffalo (*Syncerus caffer*) until a study looking at mitochondrial DNA sequences and microsatellite loci found no evidence to support this notion (Simonsen *et al.* 1998). If there was a male-bias to dispersal between populations then differences between the data from the nuclear and mitochondrial markers would be expected when they are tested against geographic distance. However, the two data sets were shown to give congruent results suggesting there is no sex bias to their dispersal patterns.

Differences in the inheritance patterns of the nuclear and mitochondrial genomes make the markers useful for detecting sex dependent processes in natural populations, but other differences between them can also be employed in population studies. In this project, differences between the nuclear and mitochondrial loci in their sensitivity to changes in effective population sizes and in their mutation rates mean that each marker shows quite different results when the Belgian populations are compared with the German population and each other. The results obtained are not contradictory, and they allow a more detailed picture of the history of the populations and the effects of habitat fragmentation to be determined. As Wilson *et al.* (1985) stated in their early review of the potential usefulness of the mitochondrial genome in evolutionary studies: "Joint comparative studies of both mitochondrial DNA and nuclear DNA variability give us valuable insights into how effective population size has varied through time."

6.2 The genetic variation in the red squirrel populations of northern Belgium

Analysis of the control region of the mitochondrial genome revealed remarkably little variation in the red squirrel populations of northern Belgium. Comparison with the German population indicates that these populations have lost most, if not all, the variation in this genome at some point in their history. In contrast, the fragment populations in Belgium are almost as variable as the German squirrels at the microsatellite loci examined. This disparity can be explained by considering the different characteristics of these two types of molecular markers.

The lack of variation in the mitochondrial genome suggests the populations have experienced either a reduction in effective population size or a selective sweep at some point in their history. The effective population size of the mitochondrial genome is one quarter that of the nuclear genome so a reduction in population size due to a bottleneck, or other demographic process, could have a dramatic effect on the mitochondrial genome whilst leaving the nuclear genome virtually unaffected (Wilson *et al.* 1985). The nuclear genome would also be unaffected by periodic selection affecting the mitochondrial genome. Even if variation in the nuclear genome was slightly reduced by a demographic contraction, microsatellite loci have a rapid rate of evolution and would therefore recover more quickly than mitochondrial loci. Even the fast evolving control region mutates at a rate which is at least one hundred times less than the slowest evolving microsatellite loci. A demographic contraction resulting in the reduced levels of variation at this locus could be caused by a simple bottleneck or by the populations taking on a metapopulation structure; these two possibilities will be discussed in full further on.

A selective sweep occurs when an advantageous mutation occurs in one copy of a linked genome and spreads through the population so that the whole chromosome becomes fixed (Ballard and Kreitman 1995). This extreme form of hitch-hiking is also called "periodic selection" and it necessarily has a dramatic effect on the genetic diversity in a population of linked genomes such as those of mitochondria (Maruyama and Birky 1991). The effects of a selective sweep on the genetics of a population are exactly the same as a severe bottleneck, all or most of the variation in the organelle DNA previously present in the population is eliminated. However, the effects of a selective sweep are locus specific, affecting only the selected locus and any loci linked to it. If a bottleneck had occurred, other loci would be expected to show evidence of it, but no evidence of a bottleneck would be expected in loci unlinked to the selected loci if a selective sweep had occurred.

In this case, the loci examined that are unlinked to the mitochondrial control region are nuclear microsatellites, which have a very fast mutation rate. The variation in the Belgian populations at these loci is only slightly reduced compared with the German population but the slight reduction in variation could be a result of the present day metapopulation structure, therefore these loci show no real evidence of an historical bottleneck. However, it is not possible to exclude the possibility of a bottleneck using these markers as the rate of evolution they experience means that they could have recovered from an historical bottleneck whilst the mitochondrial genome, which would have been more affected by a population size reduction in any case, remained genetically impoverished. Nuclear loci experiencing a slower rate of evolution, for example allozyme loci, may show the effects of an historical bottleneck if one occurred, thereby proving the role of demographic pressures.

Unfortunately, as the mitochondrial genome has a much smaller effective population size than the nuclear genome, it would never be possible to totally exclude the possibility that a bottleneck or founder event had occurred. It would always remain feasible that the bottleneck reduced the variation in the mitochondrial genome but left much of the nuclear variation. This difference could then be further exaggerated if the population remained at an intermediate size for a period after the bottleneck occurred. If the effective population size was such that mitochondrial variation continued to be lost due to the influence of drift on small populations, yet sufficient for the nuclear genome, with a larger effective size, not to feel the effects of drift to such as extent, the nuclear genome could retain its remaining diversity, perhaps even recovering some variation if sufficient time elapsed, whist the mitochondrial genome remained genetically depauperate. Therefore, if the nuclear genome shows no evidence of having experienced a bottleneck, it does not prove that a bottleneck of some size is not responsible for a loss of mitochondrial variation. For this reason, it is possible to prove a bottleneck has occurred but it is not possible to prove that one did not and that a selective sweep occurred.

Given the history of persecution and habitat destruction experienced by most mammal species in western Europe, it seems intuitively more probable that the lack of mitochondrial variation found in the red squirrels of northern Belgium is due to demographic processes. However, the possibility of a selective sweep occurring whilst the population was large enough to have been influenced by selection cannot be ruled out. In the end, the effects of both the demographic and selective processes are so similar that it is extremely difficult to distinguish between them.

6.3 Demographic processes and the cheetah controversy

The simplest explanation for a population's loss of genetic diversity is a bottleneck and reduced genetic variation in natural populations is frequently explained as being the result of an historical demographic contraction. Recently the certainty of these conclusions has been challenged as considerations of the genetics of metapopulations have revealed that populations taking on metapopulation structuring would also experience a greatly reduced effective population size. This was first pointed out by Wright in the 1930s but was more recently examined by Maruyama and Kimura (1980) and Gilpin (1991).

In a metapopulation, each separate population experiences processes that make it less than "ideal" (see section 1.3.2), thereby making the effective population size less than the actual population size. As each population has a smaller effective population size than actual size, the metapopulation overall has a much smaller effective population than it would have if it were one large population of the same total size. This reduction in effective population size is further enforced by the processes of extinction and recolonisation experienced by populations in a metapopulation, that also greatly reduce the overall effective size (Maruyama and Kimura 1980).

These results led Pimm *et al.* (1989), Gilpin (1991) and Hedrick (1996) to question the accepted theory that the lack of genetic variation found in the cheetah was due to an historical bottleneck, as proposed by O'Brien *et al.* (1983 and 1985). If the result that the cheetah has lost 75% of its variation (using other members of the cat family as a baseline) is accurate, then the bottleneck that occurred must have been equivalent to 27 generations of only ten individuals or five generations of two individuals (Gilpin 1991). Pimm *et al.* (1989) and Gilpin (1991) felt that under such extreme circumstances, the cheetah would have gone extinct.

Even when bottlenecks are known to have occurred, they have been found to have had negligible effects on the variation present in the populations. African buffalo are known to have experienced a bottleneck at the start of the 20th century when a morbillivirus rinderpest epidemic swept the continent after being introduced into Ethiopia in 1889 by an Italian military expedition (Simonsen *et al.* 1998; Wenink *et al.* 1998). The buffalos were more affected than any other African mammal and were reduced from being the most numerous large herbivore to very small numbers by the turn of the century. It was reported that only 20 survived in one herd in the Kruger National Park of South Africa. However, their numbers quickly recovered again and the species was again numerous by 1920 (Simonsen *et al.*

1998). Studies of the mitochondrial and nuclear (microsatellites (Simonsen *et al.* 1998) and major histocompatibility complex loci (Wenink *et al.* 1998)) variation showed remarkably high levels of variation despite this dramatic bottleneck. It seems that the numbers surviving the population crash were sufficient, despite the reported near extinction of the species, so that the quick recovery of the herd sizes meant that little variation, if any, was lost.

When a bottleneck is severe, variation is lost. A natural example of this is the northern elephant seal, known to have experienced a severe bottleneck at the end of the 19th century (section 1.3.1). This species is virtually devoid of allozyme variation and shows very little mitochondrial variation. Simulation studies have shown that this lack of variation is consistent with a bottleneck that is at least as severe as less than 30 seals for 20 years or less than 20 seals for 1 year (Hoelzel *et al.* 1993).

These results all illustrate the point made by Pimm *et al.* (1989) and Gilpin (1991) that bottlenecks must be very severe, in terms of number of surviving individuals or duration, to reduce the variation in populations in a noticeable way. They felt that such bottlenecks would be likely to result in extinction and therefore a bottleneck alone could not be a viable explanation for reduced variation in as many species as has been suggested. O'Brien (1989) replied to these arguments by saying that it is unknown how many species and populations have indeed gone extinct as a result of extreme bottlenecks; these species, including the cheetah and the northern elephant seal, may just be the lucky few that survived.

The debate over the relative importance of the demographic processes of bottlenecks and metapopulation structuring in determining the levels of variation in natural populations has centred on the case of the cheetah in southern Africa. Several species that show reduced variability have experienced documented population size reductions, including the northern elephant seal (Bonnell and Selander 1974) and the African lions in the Ngorongoro crater in East Africa which were reduced to 15 animals in 1962 (O'Brien and Evermann 1988), and there is little room for debate as to the causes of their reduced variation. However, there is no evidence for a bottleneck in the cheetah other than the reduced variation at the allozyme and major histocompatibility complex loci. It is perhaps because this example of a theoretical bottleneck quickly became a famous case study in conservation genetics that the conclusions drawn by O'Brien *et al.* have attracted so much attention. Debate continues as to whether the lack of variation in this species is having or will have an effect on its survival and even whether the levels of variation in the cheetah are actually really low for big cats at all (O'Brien and Evermann 1988; Pimm *et al.* 1989; O'Brien 1989; Merola 1994; O'Brien 1994).

Of relevance to this study is the discussion as to the causes of the reduction in variation (assuming it is genuine). O'Brien *et al.* (1983, 1985 and 1988) proposed that the genetic reduction was the result of one or several population bottlenecks. A comparison of the levels of variation at allozyme, mitochondrial and minisatellite loci, along with calculations as to the relative rate of accumulation of new diversity at these loci after a bottleneck, led to the proposal that the bottlenecking occurred at the end of the Pleistocene, about 10,000 years ago, when a major extinction of large vertebrates occurred (Menotti-Raymond and O'Brien 1993). Although, this is a simple and appealing conclusion, it rests on largely circumstantial evidence and simplistic assumptions. Their estimates of possible times since a bottleneck (assuming that only one occurred) range from about 3,500 to 12,750 years ago, this is consistent with their idea that the bottleneck occurred at the end of the Pleistocene, but it in no way proves it. These conclusions do not consider the possible role of demographic pressures other than a bottleneck in reducing the cheetah's genetic variation.

Although O'Brien *et. al.* quickly assumed a bottleneck was responsible for the loss of genetic variation in the species, it is not the only explanation. If the species experienced a persistent metapopulation structure, this would result in a similar amount of genetic loss (Gilpin 1991; Hedrick 1996). In contrast to the allozyme loci, the mitochondrial and minisatellite loci in cheetahs show "appreciable" levels of variation comparable with the more variable of the feline species (Menotti-Raymond and O'Brien 1993). Menotti-Raymond and O'Brien (1993) felt that this reflected the time elapsed since the bottleneck, the more mutable loci show a greater recovery of variation than the less mutable allozyme loci but Hedrick (1996) suggested that these differences could also be explained by the metapopulation theory, only requiring that the effective metapopulation size of the cheetahs is small enough to reduce the allozyme heterozygosity and that the mutation rate of the minisatellite loci is high enough to compensate for the reduced population size, with mitochondrial loci falling between the two and showing some diversity reduction.

As with the cheetah, the lack of mitochondrial variability in the red squirrels of northern Belgium makes it tempting to jump to the conclusion that they have experienced a severe bottleneck in their recent history. The variation found at the microsatellite loci is close to supposed normal levels (assuming the German squirrels show "normal" levels of variation) so if a bottleneck occurred it was long enough ago to allow recovery at the microsatellite loci but recent enough that no recovery is seen in the mitochondrial control region. However, there are many reasons to think that this is too simplistic an explanation.

For a start, it seems unlikely that the mitochondrial locus would remain totally devoid of variation in a large healthy population if enough time has elapsed for the microsatellite loci to almost fully recover. The populations of red squirrels in this area now form a metapopulation and this will be having a continued effect on the variation levels in the mitochondrial genome which, due to a smaller effective population size, will be feeling the effects of drift within the populations more strongly than the nuclear microsatellites. Therefore it seems likely that the low levels of mitochondrial variation are, at least in part, due to metapopulation structuring. It is possible that a severe bottleneck did originally eliminate all the variation at this locus and the metapopulation structure is simply maintaining that situation whilst allowing the microsatellite loci to recover. Thus, the occurrence of a severe bottleneck cannot be ruled out by these data.

The possibility that the populations in this area have never experienced a bottleneck and that the patterns of variability are solely the result of metapopulation structuring must also be considered. Hedrick (1996) stated that the effective size of a metapopulation can be particularly small when there is frequent extinction and recolonisation of subpopulations, when the number of recolonisation founders is small, and when there is little gene flow other than recolonisation. The recent observations of the red squirrel populations in this area do not fulfil the requirements of low levels of migration and small founding populations. For example, during 1995 and 1996 the population of Tallaarthof was colonised by 10 new adult squirrels, which represents a relatively large founding population for such a small woodlot. However, the red squirrel numbers in this region are rapidly increasing and the current high migration rate may reflect these recent demographic changes rather than historical patterns of demography. The populations may have previously experienced much less migration. With population sizes so small (less than 20 individuals), the frequent extinction of populations due to demographic stochasticity certainly seems likely.

The effective size of populations is further reduced by a polygynous mating system, such as that practised by the red squirrels, where one male may be responsible for a disproportionate number of matings. Only the dominant females in a population of red squirrels actually reproduce, further reducing the effective population size. It may therefore be possible that mitochondrial variability has been eliminated by extreme population structuring limiting the effective population size and exposing the populations to high levels of drift. The higher levels of variation at the microsatellite loci would then have to be explained solely by differences in the effective population sizes experienced by the two genomes and higher mutation rate replacing lost variation. The metapopulation would have had to have been small enough to eliminate the mitochondrial variation but big enough to allow the nuclear

variation to remain. The available "window" of possible metapopulation sizes satisfying these requirements is probably not very large. Whilst not being totally implausible, the idea that metapopulation structuring alone is responsible for the patterns of genetic variation seen in these populations seems improbable.

The distribution of mitochondrial alleles in the Belgian populations is an additional reason for thinking that metapopulation structuring is not a sufficient explanation for the patterns seen. All the populations are (or will soon be) fixed for the same allele; this can easily be explained by invoking a bottleneck to at least reduce the variation so that allele A was left dominating the remaining population or populations. If the bottlenecked population then became the source of all the current populations in the fragmented habitat, either through division so that it is broken into several separate populations or through recolonisation of empty areas, all the populations may then have become fixed for allele A. Even the distant and large population of Peerdsbos appears to only contain allele A, if there had been no bottleneck it would perhaps be more likely that, over such a distance, more than the one allele would have remained. A bottleneck also explains the lack of variation in the large populations of Peerdsbos and Merodese Bossen, if their sizes had not been reduced at some time it would be expected that they would show more variability.

To explain the lack of mitochondrial variability by bottlenecking alone would require the bottleneck to be severe and squirrel numbers to have reached extremely low levels, which is unlikely given the secretive nature of red squirrels and rapid rate of reproduction seen in rodents. However, it is also unlikely that metapopulation structuring would leave the squirrel populations over such a large area all fixed for the same mitochondrial control region allele. Therefore, it seems likely that both the demographic processes of a bottleneck, only severe enough to leave one allele dominating all the squirrel populations, combined with the effects of reduced effective numbers in a metapopulation, have contributed to the reduction in mitochondrial variability. If the bottleneck was not very severe then much of the nuclear variation would have remained and if the metapopulation was large enough, the faster rate of evolution seen at these loci, along with the greater effective population size experienced by the nuclear genome, could have feasibly left the microsatellite variation virtually intact.

This may be the most likely explanation given the information available at this stage, but further studies investigating the variation found at other slower mutating nuclear loci and simulation studies to test the various possibilities would be required to establish which processes have actually been involved in determining the genetics of these populations.

6.4 The genetic effects of habitat fragmentation

The microsatellite loci provide a sufficient level of variability to examine the more recent effects of habitat fragmentation on these populations. A tendency towards gene diversity excess and a deficiency of low frequency alleles was detected in three of the small fragment populations, two of which are known to have expanded dramatically within the years before and during the sampling period and are likely to have only recently been founded. A gene diversity excess, due to fewer rare or low frequency alleles being present, is expected in a population that has recently experienced a bottleneck or founder event. It is not surprising to find founder events occurring at this time as red squirrel numbers are rapidly increasing in this area of northern Belgium, probably due to the ban on hunting that was introduced in the early 1990s.

The large population of Peerdsbos also shows evidence of a recent bottleneck or founder event. It seems unlikely, given the size of the area occupied by this population, that this population has actually recently gone extinct and been recolonised. Therefore, it is perhaps more likely that it has recently suffered a bottleneck from which it is recovering. This area is deciduous, whereas Merodese Bossen is a coniferous area. Coniferous woodland is a higher quality habitat for red squirrels, so if numbers of red squirrels in this region have been limited in recent years then the population occupying the lower quality area of Peerdsbos may have been more affected.

In both the microsatellite and mitochondrial studies, the populations of Brede Zijpe and Gasthuisbos were found to be the most variable. These two populations are the largest of the small fragment populations considered in this study. Gasthuisbos occupies a much smaller area than Brede Zijpe (figure 1.6) but the populations they contain are the same size, perhaps indicating that the Gasthuisbos woodlot is of a higher quality. The Brede Zijpe population is the most variable population overall and carries the most low frequency alleles. These facts provide circumstantial evidence to suggest that these two populations are central to the metapopulation, perhaps acting as stable source populations for less stable sink populations.

All the populations contained more allelic diversity at the microsatellite loci than would be expected if they were at mutation-drift equilibrium and they have similar levels of heterozygosity as the large German population. Given the evidence for frequent migration between the populations and the fact that there is little genetic differentiation among them, it seems probable that allelic diversity is being maintained by migration and the maintenance of

allelic diversity is in turn facilitating the maintenance of heterozygosity. As drift removes alleles from the small populations, immigrants bring replacements from other populations within the metapopulation or from the surrounding area. In this way, the combination of population subdivision and migration is maintaining what variation exists within the metapopulation.

Despite the extensive fragmentation of the habitat of the red squirrels in this area of Europe, their nuclear variation appears to be retained. Genetic diversity at the more variable microsatellite loci has been virtually unaffected by habitat fragmentation. However, variation at other less variable nuclear loci may have been lost previously, under more extreme demographic pressures and, if any mitochondrial variation had remained, it would probably have continued to be lost under the current population structure due to the much smaller effective population size experienced by this genome.

The fragment populations do vary in allele content, therefore the levels of migration experienced by the populations are low enough to allow some between population differences to be maintained. Yet, there is enough migration to prevent significant losses of diversity by drift from these extremely small populations. The idea that one migrant per generation is sufficient to maintain variation within the populations of a fragmented habitat has proved controversial in conservation biology (Mills and Allendorf 1996; Varvio *et al.* 1986). The populations in northern Belgium appear to be experiencing a suitable amount of migration to maintain nuclear variation, as was thought possible by the supporters of the 'one migrant per generation' rule. For these populations, it seems that a balance between subdivision and migration has been obtained, at least at the current time, and is effectively maintaining genetic diversity.

However, it is not possible to determine the exact rate of migration between the squirrel populations, therefore it is also not possible to judge the applicability of the 'one migrant per generation' rule to these populations. Mills and Allendorf (1996) showed that substantial divergence in allele frequencies across populations would still be expected with one migrant per generation, but it may represent a sufficient minimum amount of migration. They also pointed out that the theoretical calculation of this rule makes many assumptions that may not hold for natural populations and higher levels of migration may be required in some situations. Varvio *et al.* (1986) found that the level of differentiation between populations was not only dependent on the rate of migration, but on the pattern of population subdivision and number of populations. They concluded that "selecting any single guideline like the 'one migrant per generation' rule is not theoretically well justified".

The genetics of the populations of red squirrels in northern Belgium seem to support the contention that the right amount of migration between the populations will maintain the variation, both allelic diversity and heterozygosity, within the system. These results do not, however, support the 'one migrant per generation' rule specifically as it is quite feasible that the populations are experiencing much higher rates of migration, although they do clearly illustrate the important role that migration plays in the genetics of fragmented populations.

6.5 Further work

Analysis of some other less variable nuclear markers may shed more light on which demographic processes have been involved in determining the patterns of diversity now found in the populations. If a severe bottleneck has occurred the diversity at low mutating nuclear loci may still be reduced if the time elapsed has only been sufficient to allow the restoration of variation at the rapidly mutating satellite loci.

Simulation studies which mathematically model the effects of bottlenecks of varying severity and the effects of metapopulation structuring may be able to determine the parameters of the different processes required to explain the patterns seen. A simulation study of this sort was used by Hoelzel *et al.* (1993) to determine the range in size of possible bottlenecks suffered by the northern elephant seal. Results for the red squirrel populations would be less specific as much less is actually known about what happened to the squirrels in this area, nonetheless the models may help to distinguish which combinations of processes are feasible.

This project was by no means an exhaustive study of red squirrel populations in this region. One of the more interesting results is that all the populations are dominated by the same mitochondrial allele, despite a large geographic distance between them. Extension of the study to include more squirrel populations would reveal the geographical extent of the mitochondrial uniformity. Examination of a larger sample from the large populations of Peerdsbos and Merodese Bossen would enable more reliable conclusions to be drawn about the genetics of these populations and may help explain why they are amongst the least variable in this area despite their larger size.

6.6 Summary

The differences in mutation rate and the effective population size experienced by the two types of markers used in this study mean that they show very distinctive responses to the demographic processes the populations have been exposed to. The mitochondrial genome is almost devoid of variation in the populations of northern Belgium examined in this study. This could be the result of a selective sweep but, given the history of European mammals, it is perhaps more likely to be the result of demographic pressures. Both the occurrence of a bottleneck and the extreme population structuring experienced by a metapopulation could result in an effective population reduced enough in size for the mitochondrial variation to be totally lost. The nuclear variation would be less affected by these processes and the high rate of mutation experienced by microsatellite loci would mean that these loci would recover faster than the mitochondrial control region. The relative importance of bottlenecks and metapopulation structuring in determining the genetics of these populations is unclear and hard to distinguish using the results of these analyses, but it seems likely that both types of demographic process have been involved.

The recent effects of habitat fragmentation and population expansion in this area can be seen in the microsatellite data for the small fragment populations analysed. Three populations have a tendency towards gene diversity excess and show a deficiency of low frequency alleles which indicates a recent bottleneck or founder event. Two of the populations are known to have recently experienced founder events and it seems likely that the third population should also have recently experienced high levels of immigration due to the increase in red squirrel numbers in this region. All the populations within the metapopulation carry more alleles than would be expected in populations of their size at mutation drift equilibrium. Migration between the populations appears to be counteracting the effects of random genetic drift as immigrants carry new alleles to replace those removed by drift. A combination of subdivision and migration is maintaining nuclear variation, both as allelic diversity and heterozygosity, in these populations.

The effects of habitat fragmentation are of great concern to conservation biologists as increasing areas of natural habitats are reduced and fragmented by human activities. The consequential subdivision of populations greatly reduces the effective population size of the populations occupying a fragmented habitat. This has the potential to significantly erode the genetic variation in the remaining isolated populations. The results of this study confirm the theoretical observation that migration can ameliorate these effects and maintain variation within the system. The importance of migration amongst populations of threatened species should not be understated and every effort must be made to allow it to occur.

REFERENCES

AARS, J., R. A. IMS, H-P. LIU, M. MULVEY and M. H. SMITH, **1998** Bank voles in linear habitats show restricted gene flow as revealed by mitochondrial DNA (mtDNA). *Molecular Ecology* 7: 1383-1389.

ALDRICH, P. R., J. L. HAMRICK, P. CHAVARRIAGA and G. KOCHERT, **1998** Microsatellite analysis of demographic genetic structure in fragmented populations of the tropical tree *Symphonia globulifera*. *Molecular Ecology* 7: 933-944.

ALFORD, R. L., H. A. HAMMOND, I. COTO and C. T. CASKEY, **1994** Rapid and efficient resolution of parentage by amplification of short tandem repeats. *American Journal of Human Genetics* 55: 190-195.

ALLENDORF, F. W., **1986** Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology* 5: 181-190.

AMOS, W., S. J. SAWCER, R. W. FEAKES and D. C. RUBINSZTEIN, **1996** Microsatellites show mutational bias and heterozygote instability. *Nature Genetics* 13: 390-391.

AMOS, W. and J. HARWOOD, **1998** Factors affecting levels of genetic diversity in natural populations. *Philosophical Transactions of the Royal Society, London, Series B* 353: 177-186.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. de BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, B. A. ROE, F. SANGER, P. H. SCHREIER, A. J. H. SMITH, R. STADEN and I. G. YOUNG, **1981** Sequence and organisation of the human mitochondrial genome. *Nature* 290: 457-465.

ANDRÉN, H. and A. DELIN, **1994** Habitat selection in the Eurasian red squirrel, *Sciurus vulgaris*, in relation to forest fragmentation. *Oikos* 70: 43-48.

ANDRÉN, H., **1994** Effects of habitat fragmentation on birds and mammals in landscapes with different proportions of suitable habitat: a review. *Oikos* 71: 355-366.

ANKEL-SIMONS, F. and J. M. CUMMINS, **1996** Misconceptions about mitochondria and mammalian fertilisation: implications for theories on human evolution. *Proceedings of the National Academy of Sciences of the USA* 93: 13859-13863.

van APELDOORN, R. C., W. T. OOSTENBRINK, A. van WINDEN and F. F. van der ZEE, **1992** Effects of habitat fragmentation on the bank vole, *Clethrionomys glareolus*, in an agricultural landscape. *Oikos* 65: 265-274.

van APELDOORN, R. C., C. CELADA and W. NIEUWENHUIZEN, **1994** Distribution and dynamics of the red squirrel (*Sciurus vulgaris* L.) in a landscape with fragmented habitat. *Landscape Ecology* 9: 227-235.

AQUADRO, C. F., N. KAPLAN and K. J. RISKO, **1984** An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. *Molecular Biology and Evolution* 1: 423-434.

ARCTANDER, P., P. W. KAT, R. A. AMAN and H. R. SIEGISMUND, **1996** Extreme genetic differences among populations of *Gazella granti*, Grant's gazelle, in Kenya. *Heredity* 76: 465-475.

ARDERN, S. L., D. M. LAMBERT, A. G. RODRIGO and I. G. MCLEAN, **1997** The effects of population bottlenecks on multilocus DNA variation in robins. *Journal of Heredity* 88: 179-186.

ARMOUR, J. A. L., R. NEUMANN, S. GOBERT and A. J. JEFFREYS, **1994** Isolation of human simple repeat loci by hybridization selection. *Human Molecular Genetics* 3: 599-605.

ASHLEY, M. V. and B. D. DOW, **1994** The use of microsatellite analysis in population biology: background, methods and potential applications, in *Molecular ecology and evolution: approaches and applications.* edited by B. SCHIERWATER. Birkhäuser Verlag, Basel.

AVISE, J. C., R. A. LANSMAN and R. O. SHADE, **1979** The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. *Genetics* 92: 279-295.

AVISE, J. C., J. E. NEIGEL and J. ARNOLD, **1984** Demographic influences on mitochondrial DNA lineage survivorship in animal populations. *Journal of Molecular Evolution* 20: 99-105.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB and N. C. SAUNDERS, **1987** Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* 18: 489-522.

AVISE, J. C., **1991** Matriarchal liberation. *Nature* 352: 192.

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

AYALA, F., **1982** *Population and evolutionary genetics.* The Benjamin/Cummings Publishing Company, Inc., California, USA.

BAILEY, N. T. J., **1981** *Statistical methods in biology.* Edward Arnold, London.

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

BECHER, S. A. and R. GRIFFITHS, **1998** Genetic differentiation among local populations of the European hedgehog (*Erinaceus europaeus*) in mosaic habitats. *Molecular Ecology* 7: 1599-1604.

BIBB, M. J., R. A. van ETTEN, C. T. WRIGHT, M. W. WALBERG and D. A. CLAYTON, **1981** Sequence and gene organisation of mouse mitochondrial DNA. *Cell* 26: 167-180.

BONNELL, M. L. and R. K. SELANDER, **1974** Elephant seals: genetic variation and near extinction. *Science* 184: 908-909.

BOSSART, J. L. and D. PASHLEY PROWELL, **1998** Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology and Evolution* 13: 202-206.

BOUZAT, J. L., H. A. LEWIN and K. N. PAIGE, **1998** The ghost of genetic diversity past: historical DNA analysis of the greater prairie chicken. *The American Naturalist* 152: 1-6.

BROOKES, M. I., Y. A. GRANEAU, P. KING, O. C. ROSE, C. D. THOMAS and J. L. B. MALLET, **1997** Genetic analysis of founder population bottlenecks in the rare British butterfly *Plebejus argus*. *Conservation Biology* 11: 648-661.

BROOKFIELD, J. F. Y., **1996** A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology* 5: 453-455.

BROWN, G. G., G. GADALETA, G. PEPE, C. SACCONE and E. SBISA, **1986** Structural conservation and variation in the D-loop-containing region of vertebrate mitochondrial DNA. *Journal of Molecular Biology* 192: 503-511.

BROWN, W. M., M. J. GEORGE and A. C. WILSON, **1979** Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the USA* 76: 1967-1971.

BRUFORD, M. W. and R. K. WAYNE, **1993** Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development* 3: 939-943.

CALLEN, D. F., A. D. THOMPSON, Y. SHEN, H. A. PHILIPS, R. I. RICHARDS, J. C. MULLEY and G. R. SUTHERLAND, **1993** Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. *American Journal of Human Genetics* 52: 922-927.

CASANE, D., N. DENNEBOUY, H. de ROCHAMBEAU, J. C. MOUNOLOU and M. MONNEROT, **1997** Nonneutral evolution of tandem repeats in the mitochondrial DNA control region of lagomorphs. *Molecular Biology and Evolution* 14: 779-789.

CHAKRABORTY, R., P. A. FUERST and M. NEI, **1980** Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. *Genetics* 94: 1039-1063.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON and R. DEKA, **1997** Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the USA* 94: 1041-1046.

CLAYTON, D. A., **1982** Replication of animal mitochondrial DNA. *Cell* 28: 693-705.

COBB, B. D. and J. M. CLARKSON, **1994** A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Research* 22**:** 3801-3805.

COMAS, D., F. CALAFELL, E. MATEU, A. PEREZ-LEZAUN and J. BERTRANPETIT, **1996** Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Molecular Biology and Evolution* 13**:** 1067-1077.

CORNUET, J.-M. and G. LUIKART, **1996** Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144**:** 2001-2014.

CRAWFORD, A. M., S. M. KAPPES, K. A. PATERSON, M. J. de GOTARI, K. G. DODDS, B. A. FREKING, R. T. STONE and C. W. BEATTIE, **1998** Microsatellite evolution: testing the ascertainment bias. *Journal of Molecular Evolution* 46**:** 256-260.

DALLAS, J. F., B. DOD, P. BOURSOT, E. M. PRAGER and F. BONHOMME, **1995** Population subdivision and gene flow in Danish house mice. *Molecular Ecology* 4**:** 311-320.

DAVID, V. A. and M. MENOTTI-RAYMOND, **1998** Automated DNA detection with fluorescence-based technologies, pp. 337-370 in *Molecular genetic analysis of populations: a practical approach*, edited by A. R. HOELZEL. Oxford University Press, Oxford.

DAWKINS, R., **1986** *The blind watchmaker.* Penguin Books, London.

DAWKINS, R., **1995** *River out of Eden.* Weidenfeld and Nicolson, London.

DEAN, M. and B. G. MILLIGAN, **1998** Detection of genetic variation by DNA conformational and denaturing gradient methods, pp. 263-286 in *Molecular genetic analysis of populations: a practical approach*, edited by A. R. HOELZEL. Oxford University Press, Oxford.

DELIN, A., **1996** *Habitat selection, movements and distribution of Eurasian red squirrel (Sciurus vulgaris) in boreal landscapes in relation to habitat fragmentation.* Department of Wildlife Ecology, Swedish University of Agricultural Sciences, Uppsala.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN and N. B. FREIMER, **1994** Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the USA* 91**:** 3166-3170.

DIAMOND, J. M. and R. M. MAY, **1981** Island biogeography and the design of natural reserves, pp. 228-252 in *Theoretical ecology: principles and applications*, edited by R. M. MAY. Blackwell, Oxford.

DIAS, P. C., **1996** Sources and sinks in population biology. *Trends in Ecology and Evolution* 11**:** 326-333.

DODA, J. N., C. T. WRIGHT and D. A. CLAYTON, **1981** Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. *Proceedings of the National Academy of Sciences of the USA* 78**:** 6116-6120.

van DONGEN, S., T. BACKELJAU, E. MATTHYSEN and A. A. DHONT, **1998** Genetic population structure of the winter moth (*Operophtera brumata* L.) (Lepidoptera, Geometridae) in a fragmented landscape. *Heredity* 80**:** 92-100.

DUFRESNE, C., F. MIGNOTTE and M. GUERIDE, **1996** The presence of tandem repeats and the initiation of replication in rabbit mitochondrial DNA. *European Joumal of Biochemistry* 235**:** 593-600.

ELLEGREN, H., **1995** Mutation rates at porcine microsatellite loci. *Mammalian Genome* 6**:** 376-377.

ELLEGREN, H., C. R. PRIMMER and B. C. SHELDON, **1995** Microsatellite 'evolution': directionality or bias? *Nature genetics* 11**:** 360-361.

ELLEGREN, H., S. MOORE, N. ROBINSON, K. BYRNE, W. WARD and B. C. SHELDON, **1997** Microsatellite evolution - a reciprocal study of repeat lengths at homologous loci in cattle and sheep. *Molecular Biology and Evolution* 14**:** 854-860.

ENCALADA, S. E., P. N. LAHANAS, K. A. BJORNDAL, A. B. BOLTEN, M. M. MIYAMOTO and B. W. BOWENS, **1996** Phylogeography and population structure of the Atlantic and Mediterranean green turtle *Chelonia mydas*: a mitochondrial DNA control region sequence assessment. *Molecular Ecology* 5**:** 473-483.

ESTOUP, A., L. GARNERY, M. SOLIGNAC and J.-M. CORNUET, **1995** Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* 140**:** 679-695.

ESTOUP, A., F. ROUSSET, Y. MICHALAKIS, J.-M. CORNUET, M. ADRIAMANGA and R. GUYOMARD, **1998** Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology* 7**:** 339-353.

EWENS, W. J., **1972** The sampling theory of selectively neutral alleles. *Theoretical Population Biology* 3**:** 87-112.

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

FALCONER, D. S., **1981** *Introduction to quantitative genetics.* Longman Group Ltd., London.

FALUSH, D. and Y. IWASA, **1999** Size-dependent mutability and microsatellite constraints. *Molecular Biology and Evolution* 16**:** 960-966.

FAN, E., D. B. LEVIN, B. W. GLICKMAN and D. M. LOGAN, **1993** Limitations in the use of SSCP analysis. *Mutation Research* 288**:** 85-92.

FLEISCHER, R. C., S. CONANT and M. P. MORIN, **1991** Genetic variation in native and translocated populations of the Laysan finch (*Telespiza cantans*). *Heredity* 66: 125-130.

FUERST, P. A. and T. MARUYAMA, **1986** Considerations on the conservation of alleles and of genic heterozygosity in small managed populations. *Zoo Biology* 5: 171-179.

FUMAGALLI, L., P. TABERLET, L. FAVRE and J. HAUSSER, **1996** Origin and evolution of homologous repeated sequences in the mitochondrial DNA control region of shrews. *Molecular Biology and Evolution* 13: 31-46.

GAGGIOTTI, O. E. and P. E. SMOUSE, **1996** Stochastic migration and maintenance of genetic variation in sink populations. *The American Naturalist* 147: 919-945.

GAINES, M. S., J. E. DIFFENDORFER, R. H. TAMARIN and T. S. WHITTAM, **1997** The effects of habitat fragmentation on the genetic structure of small mammal populations. *Journal of Heredity* 88: 294-304.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, **1995** Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. *Molecular Biology and Evolution* 12: 594-603.

GEMMELL, N. J., P. S. WESTERN, J. M. WATSON and J. A. MARSHALL GRAVES, **1996** Evolution of the mammalian mitochondrial control region - comparisons of control region sequences between monotreme and therian mammals. *Molecular Biology and Evolution* 13: 798-808.

GILPIN, M. E., **1987** Spatial structure and population vulnerability., pp. 125-139 in *Viable populations for conservation.*, edited by M. E. SOULÉ. Cambridge University Press, Cambridge.

GILPIN, M., **1991** The genetic effective size of a metapopulation. *Biological Journal of the Linnean Society* 42: 165-175.

GLAVAC, D. and M. DEAN, **1993** Optimisation of the single-strand conformation polymorphism (SSCP) technique for detection of point mutations. *Human Mutation* 2: 404-414.

GONZALEZ, A., J. H. LAWTON, F. S. GILBERT, T. M. BLACKBURN and I. EVANS-FREKE, **1998a** Metapopulation dynamics, abundance, and distribution in a microsystem. *Science* 281: 2045-2047.

GONZALEZ, S., J. E. MALDONADO, J. A. LEONARD, C. VILA, J. M. BARBANTI DUARTE, M. MERINO, N. BRUM-ZORRILLA and R. K. WAYNE, **1998b** Conservation genetics of the endangered Pampas deer (*Ozotoceros bezoarticus*). *Molecular Ecology* 7: 47-56.

GOODMAN, S. J., **1997** Rst Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* 6: 881-885.

GOODMAN, S. J., **1998** Patterns of extensive genetic differentiation and variation among European harbor seals (*Phoca vitulina vitulina*) revealed using microsatellite DNA polymorphisms. *Molecular Biology and Evolution* 15: 104-118.

GOUDET, J., **1999** FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.8). Institut d'Ecologie, Université de Lausanne, Switzerland. http://www.unil.ch/izea/softwares/fstat.html

GRAY, M. W., **1989** Origin and evolution of mitochondrial DNA. *Annual Review of Cell Biology* 5: 25-50.

GUO, S. W. and E. A. THOMPSON, **1992** Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48: 361-372.

GURNELL, J., **1987** *The natural history of squirrels.* Christopher Helm, London.

GYLLENSTEN, U., D. WHARTON, A. JOSEFSSON and A. C. WILSON, **1991** Paternal inheritance of mitochondrial DNA in mice. *Nature* 352: 255-257.

HAGELBERG, D., N. GOLDMAN, P. LIÓ, S. WHELAN, W. SCHIEFENHÖVEL, J. B. CLEGG and D. K. BOWDEN, **1999** Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proceedings of the Royal Society, London, Series B* 266: 485-492.

HAMADA, H., M. G. PETRINO and T. KAKUNAGA, **1982** A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proceedings of the National Academy of Sciences of the USA* 79: 6465-6469.

HANSKI, I. and M. GILPIN, **1991** Metapopulation dynamics: brief history and conceptual domain. *Biological Journal of the Linnean Society* 42: 3-16.

HARRISON, R. G., **1989** Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends in Ecology and Evolution* 4: 6-11.

HAUGE, X. Y. and M. LITT, **1993** A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics* 2: 411-415.

HAYASHI, K. and D. W. YANDELL, **1993** How sensitive is PCR-SSCP? *Human Mutation* 2: 338-346.

HEDRICK, P. W., **1994** Purging inbreeding depression and the probability of extinction: full-sib mating. *Heredity* 73: 363-372.

HEDRICK, P. W., **1995** Elephant seals and the estimation of a population bottleneck. *Journal of Heredity* 86: 232-235.

HEDRICK, P. W., **1996** Bottleneck(s) or metapopulation in cheetahs. *Conservation Biology* 10: 897-899.

HITCHINGS, S. P. and T. J. C. BEEBEE, **1998** Loss of genetic diversity and fitness in Common Toad (*Bufo bufo*) populations isolated by inimical habitat. *Journal of Evolutionary Biology* 11: 269-283.

HOELZEL, A. R. and D. R. BANCROFT, **1992** Statistical analysis of variation. pp. 399-407 in *Molecular genetic analysis of populations: a practical approach*, edited by A. R. HOELZEL. Oxford University Press, Oxford.

HOELZEL, A. R., J. HALLEY, S. J. O'BRIEN, C. CAMPAGNA, T. ARNBOM, B. LE BOEUF, K. RALLS and G. A.DOVER, **1993** Elephant seal genetic variation and the use of simulation models to investigate historical population bottlenecks. *Journal of Heredity* 84: 443-449.

HOELZEL, A. R. and A. GREEN, **1998** PCR protocols and population analysis by direct DNA sequencing and PCR-based DNA fingerprinting, pp. 201-235 in *Molecular genetic analysis of populations: a practical approach*, edited by A. R. HOELZEL. Oxford University Press, Oxford.

HUTCHISON III, C. A., J. E. NEWBOLD, S. S. POTTER and M. H. EDGELL, **1974** Maternal inheritance of mammalian mitochondrial DNA. *Nature* 251: 536-538.

HUTTER, C. M., M. D. SCHUG and C. F. AQUADRO, **1998** Microsatellite variation in *Drosophila melanogaster* and *Drosophila simulans*: A reciprocal test of the ascertainment bias hypothesis. *Molecular Biology and Evolution* 15: 1620-1636.

ISHIBASHI, Y., T. SAITOH, S. ABE and M. C. YOSHIDA, **1995** Polymorphic microsatellite DNA markers in the grey red-backed vole *Clethrionomys rufocanus bedfordiae*. *Molecular Ecology* 4: 127-128.

ISHIBASHI, Y., T. SAITOH, S. ABE and M. C. YOSHIDA, **1997** Sex-related spatial kin structure in a spring population of grey-sided voles *Clethrionomys rufocanus* as revealed by mitochondrial and microsatellite DNA analysis. *Molecular Ecology* 6: 63-71.

ISHIDA, N., T. HASEGAWA, K. TAKEDA, M. SAKAGAMI, A. ONISHI, S. INUMARU, M. KOMATSU and H. MUKOYAMA, **1994** Polymorphic sequence in the D- loop region of equine mitochondrial DNA. *Animal Genetics* 25: 215-221.

JARNE, P. and P. J. L. LAGODA, **1996** Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution* 11: 424-429.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, **1985** Hypervariable "minisatellite" regions in human DNA. *Nature* 314: 67-73.

JIMÉNEZ, J. A., K. A. HUGHES, G. ALAKS, L. GRAHAM and R. C. LACY, **1994** An experimental study of inbreeding depression in a natural habitat. *Science* 266: 271-273.

KANEDA, H., J.-I. HAYASHI, S. TAKAHAMA, C. TAYA, K. F. LINDAHL and H. YONEKAWA, **1995** Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. *Proceedings of the National Academy of Sciences of the USA* 92: 4542-4546.

KELLER, L. F., P. ARCESE, J. N. M. SMITH, W. M. HOCHACHKA and S. C. STEARNS, **1994** Selection against inbred song sparrows during a natural population bottleneck. *Nature* 372: 356-357.

KIMURA, M. and J. F. CROW, **1964** The number of alleles that can be maintained in a finite population. *Genetics* 49: 725-738.

KIMURA, M. and T. OHTA, **1978** Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the USA* 75: 2868-2872.

KINZLER, K. W. and B. VOGELSTEIN, **1989** Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Research* 17: 3645-3653.

KIRBY, K., **1995** *Rebuilding the English countryside: habitat fragmentation and wildlife corridors as issues in practical conservation.* English Nature, Northminster House, Peterborough.

KOHN, M. H. and R. K. WAYNE, **1997** Facts from feces revisited. *Trends in Ecology and Evolution* 12: 223-227.

KONDO, R., Y. SATTA, E. T. MATSUURA, H. ISHIWA, N. TAKAHATA and S. I. CHIGUSA, **1990** Incomplete maternal transmission of mitochondrial DNA in drosophila. *Genetics* 126: 657-663.

KONDO, Y., M. MORI, T. MURAMOTO, J. YAMADA, J. S. BECKMANN, D.SIMON-CHAZOTTES, X. MONTAGUTELLI, J.-L. GUÉNET and T SERIKAWA, **1993** DNA segments mapped by reciprocal use of microsatellite primers between mouse and rat. *Mammalian Genome* 4: 571-576.

KRETTEK, A., A. GULLBERG and U. ARNASON, **1995** Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus* and the phylogenetic position of the Lipotyphla. *Journal of Molecular Evolution* 41: 952-957.

LACY, R. C., **1987** Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection and population subdivision. *Conservation Biology* 1: 143-158.

LACY, R. C., **1995** Clarification of genetic terms and their use in the management of captive populations. *Zoo Biology* 14: 565-578.

LACY, R. C. and D. B. LINDENMAYER, **1995** A simulation study of the impacts of population subdivision on the mountain brushtail possum *Trichosurus caninus* Ogilby (phalangeridae: marsupialia) in south-eastern Australia. II. Loss of genetic variation within and between subpopulations. *Biological conservation* 73: 131-142.

LACY, R. M., **1997** Importance of genetic variation to the viability of mammalian populations. *Journal of Mammology* 78: 320-335.

LANDE, R., **1988** Genetics and demography in biological conservation. *Science* 241: 1455-1460.

LEBERG, P. L., **1992** Effects of population bottlenecks on genetic diversity as measured by allozyme electrophoresis. *Evolution* 46: 477-494.

LESSA, E. P. and G. APPLEBAUM, **1993** Screening techniques for detection allelic variation in DNA sequences. *Molecular Ecology* 2: 119-129.

LEVINSON, G. and G. A. GUTMAN, **1987** Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203-221.

LI, W.-H., **1979** Effect of changes in population size on the correlation between mutation rate and heterozygosity. *Journal of Molecular Evolution* 12: 319-329.

LI, W.-H. and D. GRAUR, **1991** *Fundamentals of molecular evolution.* Sinauer Associates, Inc., Sunderland, Massachusetts.

LINDENMAYER, D. B. and R. C. LACY, **1995** A simulation study of the impacts of population subdivision on the mountain brushtail possum *Trichosurus caninus* Ogilby (phalangeridae: marsupialia) in south-eastern Australia. I. Demographic stability and population persistence. *Biological conservation* 73: 119-129.

LITT, M. and J. A. LUTY, **1989** A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics* 44: 397-401.

LIU, Q. and S. S. SOMMER, **1994** Parameters affecting the sensitivities of dideoxy fingerprinting and SSCP. *PCR Methods and Applications* 4: 97-108.

LOPEZ, J. V., N. YUHKI, R. MASUDA, W. MODI and S. J. O'BRIEN, **1994** *Numt,* a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution* 39: 174-190.

LOPEZ, J. V., M. CULVER, J. C. STEPHENS, W. E. JOHNSON and S. J. O'BRIEN, **1997** Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Molecular Biology and Evolution* 14: 277-286.

LUIKART, G., W. B. SHERWIN, B. M. STEELE and F. W. ALLENDORF, **1998** Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Molecular Ecology* 7: 963-974.

LYRHOLM, T., O. LEIMAR and U. GYLLENSTEN, **1996** Low diversity and biased substitution patterns in the mitochondrial DNA control region of sperm whales: implications for estimates of time since common ancestry. *Molecular Biology and Evolution* 13**:** 1318-1326.

MACKENZIE, A., A. S. BALL and S. R. VIRDEE, **1998** *Instant notes in ecology.* BIOS Scientific Publishers Ltd., Oxford.

MAKOVA, K. D., J. C. PATTON, E. Y. KRYSANOV, R. K. CHESSER and R. J. BAKER, **1998** Microsatellite markers in wood mouse and striped field mouse (genus *Apodemus*). *Molecular Ecology* 7**:** 247-249.

MANLY, B. J. F., **1997** Distance matrices and spatial data, pp. 172-201 in *Randomisation, bootstrap and Monte Carlo methods in biology.* Chapman and Hall, London.

MARKLUND, S., R. CHAUDHARY, L. MARKLUND, K. SANDBERG and L. ANDERSON, **1995** Extensive mtDNA diversity in horses revealed by PCR-SSCP analysis. *Animal Genetics* 26**:** 193-196.

MARUYAMA, T. and M. KIMURA, **1980** Genetic variability and effective population size when local extinction and recolonisation of subpopulations are frequent. *Proceedings of the National Academy of Sciences of the USA* 77**:** 6710-6714.

MARUYAMA, T. and P. A. FUERST, **1984** Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* 108**:** 745-763.

MARUYAMA, T. and P. A. FUERST, **1985** Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* 111**:** 675-689.

MARUYAMA, T. and C. W. J. BIRKY, **1991** Effects of periodic selection on gene diversity in organelle genomes and other systems without recombination. *Genetics* 127**:** 449-451.

MATTHYSEN, E., L. LENS, S. van DONGEN, G. VERHEYEN, L. WAUTERS, F. ADRIAENSEN and A. A. DHONT, **1995** Diverse effects of forest fragmentation on a number of animal species. *Belgian Journal of Zoology* 125**:** 175-183.

MAY, B., T. A. GAVIN, P. W. SHERMAN and T. M. KORVES, **1997** Characterisation of microsatellite loci in the northern Idaho ground squirrel *Spermophilus brunneus brunneus. Molecular Ecology* 6**:** 399-400.

MCDONALD, D. B. and W. K. POTTS, **1997** DNA microsatellites as genetic markers at several scales, pp. 29-48 in *Avian molecular evolution and systematics*, edited by D. P. MINDELL. Academic Press, San Diego.

MENOTTI-RAYMOND, M. and S. J. O'BRIEN, **1993** Dating the genetic bottleneck of the African cheetah. *Proceedings of the National Academy of Sciences of the USA* 90**:** 3172-3176.

MERILA, J., M. BJÖRKLUND and A. J. BAKER, **1997** Historical demography and present day population structure of the greenfinch, *Carduelis chloris* - an analysis of mtDNA control-region sequences. *Evolution* 51: 946-956.

MEROLA, M., **1994** A reassessment of homozygosity and the case for inbreeding depression in the cheetah, *Acinonyx jubatus*: implications for conservation. *Conservation Biology* 8: 961-971.

MESSLER, W., S.-H. LI and C.-B. STEWART, **1996** The birth of microsatellites. *Nature* 381: 483.

MESTROVIC, N., M. PLOHL, B. MRAVINAC and D. UGARKOVIC, **1998** Evolution of satellite DNAs from the genus *Palorus* - experimental evidence for the "library" hypothesis. *Molecular Biology and Evolution* 15: 1062-1068.

MILLS, L. S. and F. W. ALLENDORF, **1996** The one-migrant-per-generation rule in conservation and management. *Conservation Biology* 10: 1509-1518.

MORITZ, C., T. E. DOWLING and W. M. BROWN, **1987** Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* 18: 269-292.

MORITZ, C., **1994** Applications of mitochondrial DNA analysis in conservation: a critical review. *Molecular Ecology* 3: 401-411.

MUNDY, N. I., C. S. WINCHELL and D. S. WOODRUFF, **1996** Tandem repeats and heteroplasmy in the mitochondrial DNA control region of the loggerhead shrike (*Lanius ludovicianus*). *Journal of Heredity* 87: 21-26.

NEI, M., T. MARUYAMA and R. CHAKRABORTY, **1975** The bottleneck effect and genetic variability in populations. *Evolution* 29: 1-10.

NEI, M., **1987** *Molecular Evolutionary Genetics.* Columbia University Press, New York.

NEUMANN, K., **1996** *Isolation and characterisation of microsatellite markers in the house sparrow (*Passer domesticus L.*),* Department of Genetics, University of Nottingham, Nottingham.

NEUMANN, K. and J. H. WETTON, **1996** Highly polymorphic microsatellites in the house sparrow *Passer domesticus. Molecular Ecology* 5: 307-309.

NEWTON, C. R. and A. GRAHAM, **1997** *PCR.* BIOS Scientific Publishers Ltd., Oxford.

NIELSEN, R. and P. J. PALSBØLL, **1999** Single-locus tests of microsatellite evolution: multi-step mutations and constraints on allele size. *Molecular Phylogenetics and Evolution* 11: 477-484.

NIH/CEPH COLLABORATIVE MAPPING GROUP, **1992** A comprehensive genetic linkage map of the human genome. *Science* 258: 67-86.

NISHIKAWA, N., M. OISHI and R. KIYAMA, **1995** Construction of a human genomic library of clones containing poly(dG-dA) poly(dT-dC) tracts by Mg^{2+} -dependent triplex affinity capture. *Journal of Biological Chemistry* 270: 9258-9268.

O'BRIEN, S. J., D. E. WILDT, D. GOLDMAN, C. R. MERRIL and M. BUSH, **1983** The cheetah is depauperate in genetic variation. *Science* 221: 459-462.

O'BRIEN, S., M. ROELKE, L. MARKER, A. NEWMAN, C. WINKLER, D. MELTZER, L. COLLY, J.F. EVERMAN, M. BUSH and D.E. WILDT, **1985** Genetic basis for species vulnerability in the cheetah. *Science* 227: 1428-1434.

O'BRIEN, S. J. and J. F. EVERMANN, **1988** Interactive influence of infectious disease and genetic diversity in natural populations. *Trends in Ecology and Evolution* 3: 254-259.

O'BRIEN, S. J., **1989** Reply from S.J. O'Brien. *Trends in Ecology and Evolution* 4: 178.

O'BRIEN, S. J., **1994** The cheetah's conservation controversy. *Conservation Biology* 8: 1153-1155.

OHTA, T. and M. KIMURA, **1973** A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* 22: 201-204.

ORITA, M., H. IWAHANA, H. KANAZAWA, K. HAYASHI and T. SEKIYA, **1989** Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences of the USA* 86: 2766-2770.

OSTRANDER, E. A., P. M. JONG, J. RINE and G. DUYK, **1992** Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. *Proceedings of the National Academy of Sciences of the USA* 89: 3419-3423.

PAETKAU, D. and C. STROBECK, **1994** Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology* 3: 489-495.

PAETKAU, D. and C. STROBECK, **1995** The molecular basis and evolutionary history of a microsatellite null allele in bears. *Molecular Ecology* 4: 519-520.

PEMBERTON, J. M., J. SLATE, D. R. BANCROFT and J. A. BARRETT, **1995** Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Molecular Ecology* 4: 249-252.

PERNA, N. T. and T. D. KOCHER, **1996** Mitochondrial DNA: molecular fossils in the nucleus. *Current Biology* 6: 128-129.

PIMM, S. L., J. L. GITTLEMAN, G. F. McCRACKEN and M. GILPIN, **1989** Plausible alternatives to bottlenecks to explain reduced genetic diversity. *Trends in Ecology and Evolution* 4: 176-178.

POPE, T. R., **1996** Sociology, population fragmentation and patterns of genetics loss in endangered primates., pp. 119-159 in *Conservation Genetics; case histories from nature.*, edited by J. C. AVISE and J. L. HAMRICK. Chapman and Hall, London.

PRIMMER, C. R., H. ELLEGREN, N. SAINO and A. P. MØLLER, **1996a** Directional evolution in germline microsatellite mutations. *Nature Genetics* 13: 391-393.

PRIMMER, C. R., A. P. MÖLLER and H. ELLEGREN, **1996b** A wide -range survey of cross-species microsatellite amplification in birds. *Molecular Ecology* 5: 365-378.

PRIMMER, C. R. and H. ELLEGREN, **1998** Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution* 15: 997-1008.

PRIMMER, C. R., N. SAINO, A. P. MØLLER and H. ELLEGREN, **1998** Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows *Hirundo rustica*. *Molecular Biology and Evolution* 15: 1047-1054.

PROCHAZKA, M., **1996** Microsatellite hybrid capture technique for simultaneous isolation of various STR markers. *Genome Research* 6: 646-649.

PUSEY, A. and M. WOLF, **1996** Inbreeding avoidance in animals. *Trends in Ecology and Evolution* 11: 201-206.

QUINN, T. W. and A. C. WILSON, **1993** Sequence evolution in and around the mitochondrial control region in birds. *Journal of Molecular Evolution* 37: 417-425.

QUINN, T. W., **1997** Molecular evolution of the mitochondrial genome, pp. 3-28 in *Avian Molecular Evolution and Systematics*, edited by D. P. MINDELL. Academic Press, San Diego.

RASSMANN, K., C. SCHLÖTTERER and D. TAUTZ, **1991** Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. *Electrophoresis* 12: 113-118.

RAYMOND, M. and F. ROUSSET, **1995a** GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *The Journal of Heredity* 86: 248-249.

RAYMOND, M. and F. ROUSSET, **1995b** An exact test for population differentiation. *Evolution* 49: 1280-1283.

ROLSTAD, J., **1991** Consequences of forest fragmentation for the dynamics of bird populations: conceptual issues and the evidence. *Biological Journal of the Linnean Society* 42: 149-163.

ROSE, O. and D. FALUSH, **1998** A threshold size for microsatellite expansion. *Molecular Biology and Evolution* 15: 613-615.

ROUSSET, F., **1997** Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* 145: 1219-1228.

ROUSSET, F. and M. RAYMOND, **1997** Statistical analyses of population genetic data: new tools, old concepts. *Trends in Ecology and Evolution* 12: 313-317.

ROYLE, N. J., M. C. HILL and A. J. JEFFREYS, **1992** Isolation of telomere junction fragments by anchored polymerase chain reaction. *Proceedings of the Royal Society, London, Series B* 247: 57-61.

RUBINSZTEIN, D. C., W. AMOS, J. LEGGO, S. GOODBURN, S. JAIN, S.-H. LI, R. L. MARGOLIS, C. A.ROSS and M. A. FERGUSON-SMITH, **1995** Microsatellite evolution - evidence for directionality and variation in rate between species. *Nature Genetics* 10: 337-343.

SACCONE, C., M. ATTIMONELLI and E. SBISÀ, **1987** Structural elements highly preserved during the evolution of the D-loop-containing region in vertebrate mitochondrial DNA. *Journal of Molecular Evolution* 26: 205-211.

SACCONE, C., G. PESOLE and E. SBISÀ, **1991** The main regulatory region of the mammalian mitochondrial DNA: structure-function model and evolutionary pattern. *Journal of Molecular Evolution* 33: 83-91.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS and H. A. ERLICH, **1988** Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487-491.

SAJANTILA, A., A.-H. SALEM, P. SAVOLAINEN, K. BAUER, C. GIERIG and S PÄÄBO, **1996** Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proceedings of the National Academy of Sciences of the USA* 93: 12035-12039.

SAMBROOK, J., E. FITSCH and T. MANIATIS, **1989** *Molecular cloning: a laboratory manual.* Cold Spring Harbor Laboratory Press.

SANGER, F., S. NICKLEN and A. R. COULSON, **1977** DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* 74: 5463-5467.

SAVILLE, B. J., Y. KOHLI and J. B. ANDERSON, **1998** MtDNA recombination in a natural population. *Proceedings of the National Academy of Sciences of the USA* 95: 1331-1335.

SBISÀ, E., F. TANSARIELLO, A. REYES, G. PESOLE and C. SACCONE, **1997** Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. *Gene* 205: 125-140.

SBISÀ, E., F. TANSARIELLO, A. REYES, G. PESOLE and C. SACCONE, **1998** Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. http://www.ba.cnr.it/dloop.html

SCHLÖTTERER, C. and D. TAUTZ, **1992** Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* 20: 211-215.

SCHLÖTTERER, C. and J. PEMBERTON, **1994** The use of microsatellites for genetic analysis of natural populations, in *Molecular ecology and evolution: approaches and applications.* edited by B. SCHIERWATER. Birkhäuser Verlag, Basel.

SCHLÖTTERER, C., **1998a** Are microsatellites really simple sequences? *Current Biology* 8: R132-R134.

SCHLÖTTERER, C., **1998b** Microsatellites, pp. 237-261 in *Molecular genetic analysis of populations: a practical approach*, edited by A. R. HOELZEL. Oxford University Press, Oxford.

SCHLÖTTERER, C., R. RITTER, B. HARR and G. BREM, **1998** High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution* 15: 1269-1274.

SCHNEIDER, S., J.-M. KUEFFER, D. ROESSLI and L. EXCOFFIER, **1997** Arlequin ver. 1.1: A software for population genetic data analysis., Genetics and Biometry Laboratory, University of Geneva, Switzerland. http://anthropologie.unige.ch/arlequin

SEPPÄ, P. and A. LAURILA, **1999** Genetic structure of island populations of the anurans *Rana temporaria* and *Bufo bufo. Heredity* 82: 309-317.

SHITARA, H., J.-I. HAYASHI, S. TAKAHAMA, H. KANEDA and H. YONEKAWA, **1998** Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. *Genetics* 148: 851-857.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, **1993** VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* 134: 983-993.

SIMONSEN, B. T., H. R. SIEGISMUND and P. ARCTANDER, **1998** Population structure of African buffalo inferred from mtDNA sequences and microsatellite loci: high variation but low differentiation. *Molecular Ecology* 7: 225-237.

SLATKIN, M., **1993** Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47: 264-279.

SLATKIN, M., **1995** A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.

SOKAL, R. R. and F. J. ROHLF, **1981** *Biometry.* W.H. Freeman and Company, New York.

SPINARDI, L., R. MAZARS and C. THEILLET, **1991** Protocols for an improved detection of point mutations by SSCP. *Nucleic Acids Research* 19: 4009.

STACY, J. E., P. E. JORDE, H. STEEN, R. A. IMS, A. PURVIS and K. S. JOKOBSEN, **1997** Lack of concordance between mtDNA gene flow and population density fluctuations in the bank vole. *Molecular ecology* 6: 751-759.

STEVENS, S., J. COFFIN and C. STROBECK, **1997** Microsatellite loci in Columbian ground squirrels *Spermophilus columbianus*. *Molecular Ecology* 6: 493-495.

STEWART, D. T. and A. J. BAKER, **1994** Evolution of mtDNA D-loop sequences and their use in phylogenetic studies of shrews in the subgenus *Otisorex* (*Sorex:* Soricidae: Insectivora). *Molecular Phylogenetics and Evolution* 3: 38-46.

STRASSMANN, J. E., C. R. SOLÍS, J. M. PETERS and D. C. QUELLER, **1996** Strategies for finding and using highly polymorphic DNA microsatellite loci for studies of genetic relatedness and pedigrees, pp. 163-180 in *Molecular Zoology: Advances, Strategies, and Protocols*, edited by J. D. FERRARIS and S. R. PALUMBI. Wiley-Liss, Inc., USA

TALBOT, J., J. HAIGH and Y. PLANTE, **1996** A parentage evaluation test in North American elk (Wapiti) using microsatellites of ovine and bovine origin. *Animal Genetics* 27: 117-119.

TAUTZ, D., M. TRICK and G. A. DOVER, **1986** Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322: 652-656.

TAUTZ, D., **1989** Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acid Research* 17: 6463-6471.

TAUTZ, D., **1993** Notes on the definition and nomenclature of tandemly repetitive DNA sequences, pp. 21-28 in *DNA fingerprinting: state of the science*, edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhäuser Verlag Basel.

TAYLOR, A. C., W. B. SHERWIN and R. K. WAYNE, **1994** Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology* 3: 277-290.

TAYLOR, A. C., J. MARSHALL GRAVES, N. D. MURRAY, S. J. O'BRIEN, N. YUHKI and B. SHERWIN, **1997** Conservation genetics of the koala (*Phascolarctos cinereus*): low mitochondrial DNA variation amongst southern Australian populations. *Genetical Research* 69: 25-33.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, **1994** Clustal-W – improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673-4690.

van TREUREN, R., **1998** Estimating null allele frequencies at a microsatellite locus in the oystercatcher (*Haematopus ostralegus*). *Molecular Ecology* 7: 1413-1417.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, **1993** Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737-749.

VARVIO, S.-L., R. CHAKRABORTY and M. NEI, **1986** Genetic variation in subdivided populations and conservation genetics. *Heredity* 57: 189-198.

VERBOOM, B. and R. van APELDOORN, **1990** Effects of habitat fragmentation on the red squirrel, *Sciurus vulgaris* L.. *Landscape Ecology* 4: 171-176.

WALBERG, M. W. and D. A. CLAYTON, **1981** Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. *Nucleic Acids Research* 9: 5411-5421.

WANG, J. and A. CABALLERO, **1999** Developments in predicting the effective size of subdivided populations. *Heredity* 82: 212-226.

WAUTERS, L. and A. A. DHONT, **1992** Spacing behaviour of red squirrels, *Sciurus vulgaris*, variation between habitats and the sexes. *Animal Behaviour* 43: 297-311.

WAUTERS, L., P. ÇASALE and A. A. DHONT, **1994a** Space use and dispersal of red squirrels in fragmented habitats. *Oikos* 69: 140-146.

WAUTERS, L. A., Y. HUTCHINSON, D. T. PARKIN and A. A. DHONT, **1994b** The effects of habitat fragmentation on demography and on the loss of genetic variation in the red squirrel. *Proceedings of the Royal Society, London, Series B* 255: 197-111.

WAUTERS, L. A., A. A. DHONT, H. KNOTHE and D. T. PARKIN, **1996** Fluctuating asymmetry and body size as indicators of stress in red squirrel populations in woodland fragments. *Journal of Applied Ecology* 33: 735-740.

WEBER, J. L. and P. E. MAY, **1989** Abundant class of human DNA polymorphism which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* 44: 388-396.

WEBER, J. L., **1990** Informativeness of human $(dC\text{-}dA)_n|(dG\text{-}dT)_n$ polymorphisms. *Genomics* 7: 524-530.

WEBER, J. L. and C. WONG, **1993** Mutation of human short tandem repeats. *Human Molecular Genetics* 2: 1123-1128.

WEIR, B. S. and C. C. COCKERHAM, **1984** Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.

WENINK, P. W., A. F. GROEN, M. E. ROELKE-PARKER and H. H. T. PRINS, **1998** African buffalo maintain high genetic diversity in the major histocompatibility complex in spite of historically known population bottlenecks. *Molecular Ecology* 7: 1315-1322.

WIERDL, M., M. DOMINSKA and T. D. PETES, **1997** Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146: 769-779.

WILSON, A. C., R. L. CANN, S. M. CARR, M. GEORGE, U. B. GYLLENSTEN, K. M. HELM-BYCHOWSKI, R. G. HIGUCHI, S. R. PALUMBI, E. M. PRAGER, R. D. SAGE and M. STONEKING, **1985** Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society* 26: 375-400.

WOOTEN, M. C., K. T. SCRIBNER and J. T. KREHLING, **1999** Isolation and characterisation of microsatellite loci from the endangered beach mouse *Peromyscus polionotus*. *Molecular Ecology* 8: 167-168.

WORTHINGTON WILMER, J., C. MORITZ, L. HALL and J. TOOP, **1994** Extreme population structuring in the threatened ghost bat, *Macroderma gigas*: evidence from mitochondrial DNA. *Proceedings of the Royal Society, London, Series B* 257: 193-198.

WRIGHT, S., **1931** Evolution in Mendelian populations. *Genetics* 16: 97-159.

WRIGHT, S., **1943** Isolation by distance. *Genetics* 28: 114-138.

WRIGHT, S., **1950** The genetical structure of populations. *Annuals of Eugenics* 15: 323-354.

WRIGHT, S., **1973** The origin of the *F*-statistics for describing the genetic aspects of population structure, in *Genetic Structure of Populations.*, edited by N. E. MORDEN. University of Hawaii.

WU, C. and W. LI, **1985** Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the USA* 82: 1741-1745.

YOUNG, A., T. BOYLE and T. BROWN, **1996** The population genetic consequences of habitat fragmentation for plants. *Trends in Ecology and Evolution* 11: 413-418.

YOUNG, A. G., A. H. D. BROWN and F. A. ZICH, **1999** Genetic structure of fragmented populations of the endangered daisy *Rutidosis leptorrhynchoides*. *Conservation biology* 13: 256-265.

ZHANG, D.-X. and G. M. HEWITT, **1996** Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in Ecology and Evolution* 11: 247-251.

ZHIVOTOVSKY, L. A., M. W. FELDMAN and S. A. GRISHECHKIN, **1997** Biased mutations and microsatellite variation. *Molecular Biology and Evolution* 14: 926-933.

ZIEGLE, J. S., Y. SU, K. P. CORCORAN, L. NIE, P. E. MAYRAND, L. B. HOFF, L. J. MCBRIDE, M. N. KRONICK and S. R. DIEHL, **1992** Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics* 14: 1026-1031.

APPENDIX A: THE RESULTS OF THE MITOCHONDRIAL CONTROL REGION ANALYSIS

sample number	year	SSCP haplotype	sequence allele
BZ 555	95&96	A	
BZ 658	95&96	A	
BZ 670	95&96	A	i
BZ 31B	95&96	A	
BZ 955	95&96	A	
BZ 635	95&96	A	
BZ 660	95	A	
BZ A24	96	A	i
BZ 696	96	A	
BZ 968	96	A	
BZ E39	96	A	
BZ 26A	96	A	
BZ E27	96	A	
BZ 967	96	A	i
BZ 912	95&96	A	
BZ 907	95	A	
BZ 905	95&96	B	ii
BZ 903	95	A	
BZ 908	95&96	A	
BZ 970	95&96	A	
GH D69	95&96	A	
GH 530	95	B	ii
GH 808	95&96	A	
GH 47B	95&96	B	ii
GH 651	95&96	A	
GH F3E	95	C	iii
GH 652	95	A	
GH 81D	95&96	A	
GH 951	95&96	A	
GH 533	95		
GH 605	95&96	A	
GH 649	95&96	A	i
GH 752	95&96	A	
GH 953	95&96	A	
GH 954	95	A	
GH 906	95&96	A	i
GH 139	96	A	i
GH 971	96	A	
GH 95A	96	A	
KE 699	96	A	i
KE 969	96	A	
KE 509	96	A	i
KE F7F	96	A	

sample number	year	SSCP haplotype	sequence allele
L 917	96	A	i
L A04	96	A	
L B40	96	A	i
T 921	96	A	
T 956	96	A	
T F34	96	A	i
T 45A	96	A	
T 603	96	A	
T 55D	96	A	
T 36E	96	A	
T 13D	96	A	
T 276	96	A	i
T A0E4E	96	A	
AWW1 D16	96	A	
AWW1 F2C	96	A	i
AWW1 A1C	96	A	
AWW1 A54	96	A	
AWW1 63A	96	A	
AWW1 C6C	96	A	
AWW1 E7E	96	A	
AWW1 F05	96	A	
AWW1 130	96	A	i
AWW1 67F	96	A	
AWW1 C65	96	A	
AWW1 67B	96		
AWW1 93D	96	A	
AWW2 30E	96	A	
AWW2 O77	96	A	
AWW2 206	96	A	
AWW2 E24	96	A	i
AWW2 431	96	A	
AWW2 953	96	A	
AWW2 C30	96	A	
AWW2 825	96	A	
AWW2 17A	96	A	i
AWW2 D21	96	A	
AWW2 60B	96	A	
AWW2 554	96		
AWW2 F2E	96	A	

sample number	year	SSCP haplotype	sequence allele
AWW3 173	96	A	
AWW3 750	96	A	
AWW3 127	96	A	
AWW3 B59	96	A	i
AWW3 74A	96	A	
AWW3 24F	96	A	
AWW3 06C	96	A	
AWW3 604	96	A	
AWW3 D23	96	A	
AWW3 357	96	A	
AWW3 D02	96	A	i
MB 10		A	
MB 12		A	
MB 21		A	
MB 27		A	i
MB 93		A	
MB 120		A	
MB 167		A	
MB 202		A	i
MB 322		A	
MB 325		A	
MB 326		A	
MB 327		A	i
MB 396		A	
MB 932		A	
MB 979		A	i
P O7		A	
P 20		A	
P 22		A	
P 25		A	
P 26		A	i
P 35		A	i
P 36		A	i
P 40		A	i
P 43		A	
P 44		A	i
P 48		A	
P 72		A	
P 88		A	

sample number	year	SSCP haplotype	sequence allele
P 106		A	
P 222		A	
P 238		A	
P 479		A	
P 489		A	
P 491		A	
P 499		A	
P 844		A	i
P 919		A	
P 920		A	
P 924		A	i
WH 12		e	5
WH 21		f	6
WH 39		g	7
WH 43		d	2
WH 61		h	8
WH 62		i	9
WH 63		j	10
WH 64		k	11
WH 65		d	2
WH 66		l	12
WH 67		a	13*
WH 68		m	14
WH 69		b	15*
WH 70		a	1
WH 71		n	16
WH 72		c	3
WH 73		c	3
WH 74		a	1
WH 75		b	4
WH 76		o	17
WH 77		p	18
WH 78		q	19
WH 80		r	20
WH 81		a	1
WH 83		b	4

* these samples had a different sequence allele than that indicated by the SSCP haplotype (see text: chapter 3, section 3.3.1)

The sample number indicates the population from which the sample was taken with the following codes: BZ= Brede Zijpe, GH= Gasthuisbos, KE= Kegelslei, L= Luisbos, T= Tallaarthof, AWW 1-3= Antwerp water works areas 1,2 and 3, MB= Merodese Bossen, P= Peerdsbos, WH= Waldhäuser.

Where the year the individual was present in the population is known it is shown.

The sequence allele number is given for the samples that were sequenced, the German population of Waldhäuser did not share any SSCP haplotypes or sequence alleles with the Belgian populations.

APPENDIX B: THE RESULTS OF THE MICROSATELLITE ANALYSIS

sample number	year	RSμ1		RSμ3		RSμ4		RSμ5		RSμ6	
BZ 555	95&96	180	188	171		264		139	141	122	128
BZ 658	95&96										
BZ 670	95&96	176	180	165	169	260		139		128	
BZ 31B	95&96	188		171		260	264	139	141	122	
BZ 955	95&96	180	196	165		264	280	139		122	128
BZ 635	95&96	188		165		260	264	139		128	
BZ 660	95	180	196	169	171	260	264	139		122	
BZ A24	96	180	196	171		260	264	139		128	
BZ 696	96	188		165		264	280	139		122	128
BZ 968	96	192		165		264		139		122	128
BZ E39	96	188	196	165	169	260	264	139	141	128	
BZ 26A	96	180	192	165	171	260	264	139		128	
BZ E27	96	176	180	165		260	264	141		122	128
BZ 967	96	180		171		260	264	139	141	128	
BZ 912	95&96	180	184	161	163	264		139	141	122	128
BZ 907	95	180	188	165		264		139		128	
BZ 905	95&96	180	188	171		264	268	139		128	
BZ 903	95	180	188	165	171	260	264	139	141	128	
BZ 908	95&96	180	196	165		260	264	139		122	128
BZ 970	95&96	180	196	167	171	264	268	139	141	128	
GH D69	95&96			165		260	268	139			
GH 530	95	176	180	165	171	260	268	141		128	
GH 808	95&96	180	196	165	167	260	264	139	141	128	
GH 47B	95&96	180		165	167	260	268	139		128	
GH 651	95&96			165				139		122	128
GH F3E	95	180	188	165		264	268	141		122	128
GH 652	95	188	192	165	167	264	280	127	139	128	
GH 81D	95&96	180	196	171		264	268	139	141	128	
GH 951	95&96	180		165	167	264		139		128	
GH 533	95										
GH 605	95&96	188		165	171	260	264	139	141	122	128
GH 649	95&96	180	192	163	169	260	264	139		128	
GH 752	95&96	180	196	165		264		139		122	
GH 953	95&96	180	188	165		260	280	127	141	128	
GH 954	95	180	192	165		264		139		122	128
GH 906	95&96	180	196	165	171	260	264	139		128	
GH 139	96	192	196	165	169	260	264	139		122	128
GH 971	96	180		165		260	264	139			
GH 95A	96	180	188	165	167	264	280	139		122	128

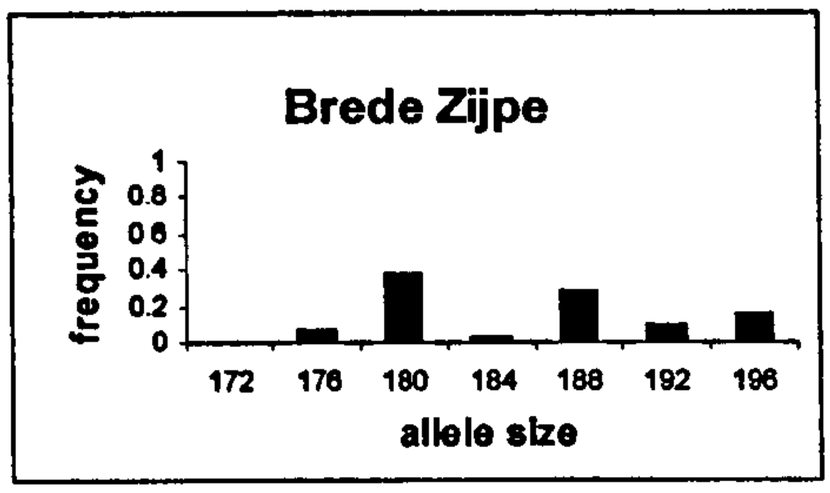
sample number	year	RSμ1		RSμ3		RSμ4		RSμ5		RSμ6	
KE 699	96	180	188	163	165	264	268	139		122	128
KE 969	96	180	188	165		264		139		128	
KE 509	96	188	192	167	169	268		139	141	128	
KE F7F	96	180	184	165		264	268	139		122	128
L 917	96	196		165	169	264	268	139		122	128
L A04	96	188		165	169	264	268	139		128	
L B40	96	180	192	165	169	264	268	139		122	
T 921	96	180		165		264	268	139		128	
T 956	96	180		165	169	260	264	139		122	
T F34	96	184	192	165	169	260		139	141	122	128
T 45A	96	180	196	165	169	264	268	139	141	122	128
T 603	96	180	196	165	169	260	280	139		128	
T 55D	96	180	192	165	169	256	260	139	141	128	
T 36E	96			165	169	264		139		128	
T 13D	96	180	192	165		260	264	139		122	128
T 276	96	184		165	169	276		139		122	128
T A0E4E	96	180	184	169		264	268	139	141	122	128
AWW1 D16	96	180		167		260	264	141		128	
AWW1 F2C	96	180		165		264		127	139	128	
AWW1 A1C	96	180	192	165		264		139		128	
AWW1 A54	96	188	192	165		264		141		128	
AWW1 63A	96	180	188	165		264	268	127	139	128	
AWW1 C6C	96	180		165		260		127	139	128	
AWW1 E7E	96	180		165	167	260	264	127	141	128	
AWW1 F05	96	180	192	167	169	264	280	127	141	128	
AWW1 130	96	180	192	165	171	268	280	127	139	128	
AWW1 67F	96	180		165	167	264	268	123	127	128	
AWW1 C65	96	180	192	165		268	280	127	139	128	
AWW1 67B	96										
AWW1 93D	96	192	196	165	167	264	268	139		128	
AWW2 30E	96	180		165	169	260		127		128	
AWW2 O77	96	180	188	165		260	280	139	141	128	
AWW2 206	96	180		165		260	264	139		122	128
AWW2 E24	96	180	192	165		264		139		128	
AWW2 431	96	180	192	169		264	280	123	141	128	
AWW2 953	96	196		165	169	260	264	139		128	
AWW2 C30	96	180	192	165	167	260	264	127		128	
AWW2 825	96	180		165	169	260	264	139	141	128	
AWW2 17A	96	180	188	165		260	264	127	141	128	
AWW2 D21	96	180	192	165		256	264	127	141	128	
AWW2 60B	96	180	192	169		260	264	139		122	128
AWW2 554	96										
AWW2 F2E	96	192		167		260	264	141		128	

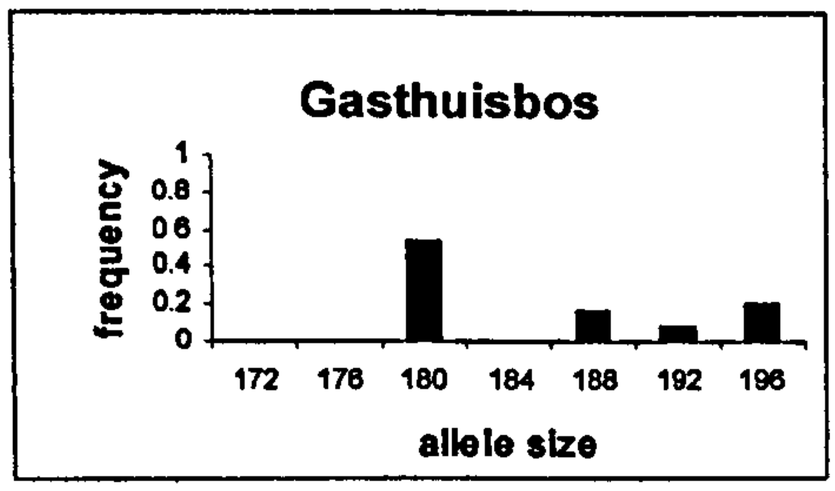
sample number	year	RSμ1		RSμ3		RSμ4		RSμ5		RSμ6	
AWW3 173	96	188	192	165	169	264	280	139	141	122	
AWW3 750	96	180	192	165		264	280	141		122	128
AWW3 127	96	180	192	165	169	264	280	139	141	128	
AWW3 B59	96	180	192	165		264	280	139	141	122	128
AWW3 74A	96	180	192	165	169	260	264	139	141	122	128
AWW3 24F	96	180	188	165	169	260	264	139		122	128
AWW3 06C	96	180	192	165	169	260	264	139		122	128
AWW3 604	96	180	188	165	169	260	264	127		128	
AWW3 D23	96	180	188	165		260	280	127	141	128	
AWW3 357	96	180	188	165		260	264	139		122	
AWW3 D02	96	192		165	169	260	264	139	141	122	128
MB O4		188		165		264	280	139		128	
MB O8		188	196	165	169	260		127	139	128	
MB 10		180	188	165		260	272	139		128	
MB 13		180	188	165	169	260		139		128	
MB 14		188	196	165		260		139		128	
MB 17		188		165		260	264	139		128	
MB 27		180	188	165	167	268	280	139	141	128	
MB 31		180		171		264	268	139		128	
MB 38		180		165	167	260	268	127	139	128	
MB 43		180	188	165	171	260	264	139		128	
MB 70		180	196	165	169	260	264	139		128	
MB 78		180	188	165		260	268	139		128	
MB 102		188	196	165	167	268		127	139	128	
MB 104		180	188	165	169	260		139		128	
MB 120		188	196	165	167	264	268	139		128	
MB 167		180		165	167	260	264	139	141	128	
MB 202		180	188	165		260		139		128	
MB 295		180	192	165		264	268	139		128	
MB 932		180		165		260	264	139		128	
MB 979		180	196	165		264	272	139	141	128	
P O1		180	196	165		268		139		128	
P O7		180	196	165	167	260	264	139		128	
P 20		180		165	167	264		139		128	
P 22		180	188	165	167	264		139		128	
P 25		196		165	167	260	280	139		128	
P 26		196		167		260		139	141	128	
P 35		180	188	165	167	264		139		128	
P 40		192		165		260	268	139		128	
P 43		188	196	165	167	264		139	141	128	
P 44		188	196	165		268	280	139		128	
P 48		180	196	165	167	260	264	139		128	
P 72		188	196	165				139	141	128	
P 82		188	196			260	264	139		128	
P 85		180	196	165	167	260	264	139	141	128	
P 88		180	196	165		260	268	139	141	128	

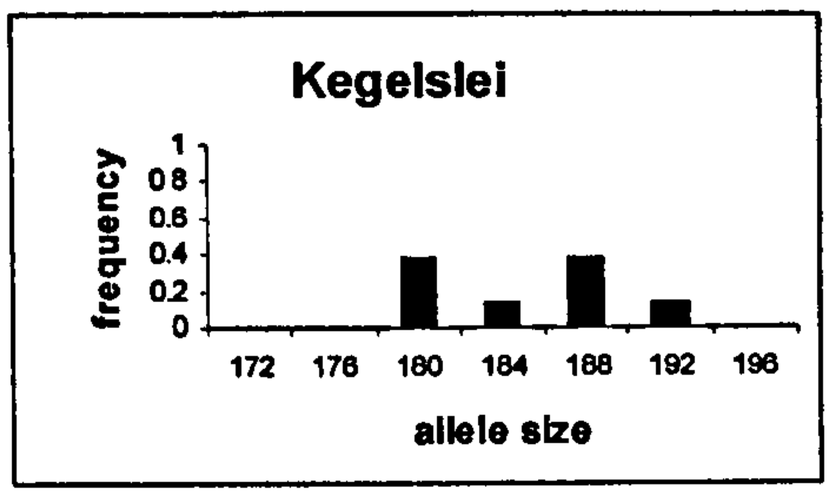
sample number	year	RSμ1		RSμ3		RSμ4		RSμ5		RSμ6	
P 106		180		165	167	264		139		128	
P 222		180	196	165		260	268	141		128	
P 238		180	196	165	167	260	264	139	141	128	
P 279		188	196	165	167	264	268	139		128	
P 479		180	196	165	167	264		139		128	
P 489		188	192	165	167	260	264	139	141	128	
P 491		188	196	165	167	260	268	139		128	
P 499		180	196			260	280	139		128	
P 844		188	192			280		127	139	128	
P 919		192	196			264	268	139	141	128	
P 920		192	196			260	264	139		128	
P 924		180						139		128	
WH 2		180		165	171	264	276	139	141	122	128
WH 12		180		165		276	280	139		128	
WH 21		180	188	165	171	264	276	135	139	122	
WH 39		180	192	171	173	260	272	139	141	122	
WH 43		180	184	165	171	280		139		122	128
WH 61		176	196	173		264	268	139	141	122	128
WH 62		188		165	171	260	280	139		122	131
WH 63		172	196	165		260	280	139		122	128
WH 64		188		165	171	260	264	139		128	
WH 65		180		165		260	276	139	141	122	
WH 66		172	180	165		260	276	139		122	131
WH 67		184	196	165		264	276	139		128	
WH 68		180	188	165		260	280	139		122	128
WH 69		188		165		260	280	137	139	122	
WH 70		180	192	165		260	276	139		122	128
WH 71		192		165		264	268	139		131	
WH 72		188		165	171	260	264	139	141	122	128
WH 73		192	196	165	171	260	264	139	141	122	128
WH 74		192	196	165	171	276	280	139		128	
WH 75		180	188	165	171	260	268	135	139	122	128
WH 76		188	192	165	171	268	276	139	141	122	125
WH 77		180	188	165		260	280	139		122	128
WH 78		180	188	165	171	272	284	139		122	128
WH 80		184	188	165	171	264	276	139		122	
WH 81		188	196	171		264	272	137	139	128	
WH 82		184	188	165		264	276	141	143	122	128
WH 83		180	188	165		264	268	137	139	122	

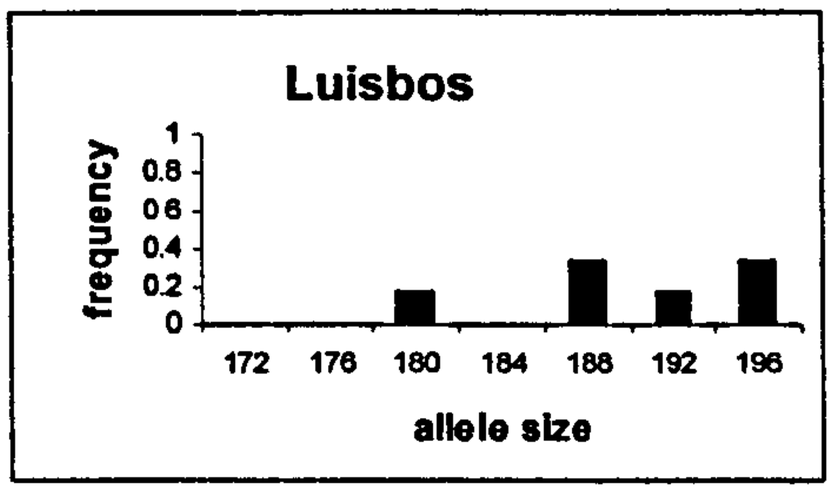
The microsatellite allele sizes in base pairs are shown for each locus and sample. Where a sample was missing or the locus was not amplified the entry is left blank. The sample number indicates the population from which the sample was taken with the following codes: BZ= Brede Zijpe, GH= Gasthuisbos, KE= Kegelslei, L= Luisbos, T= Tallaarthof, AWW 1-3= Antwerp water works areas 1,2 and 3, MB= Merodese Bossen, P= Peerdsbos, WH= Waldhäuser. The year the individual was present in the population in shown where it is known.

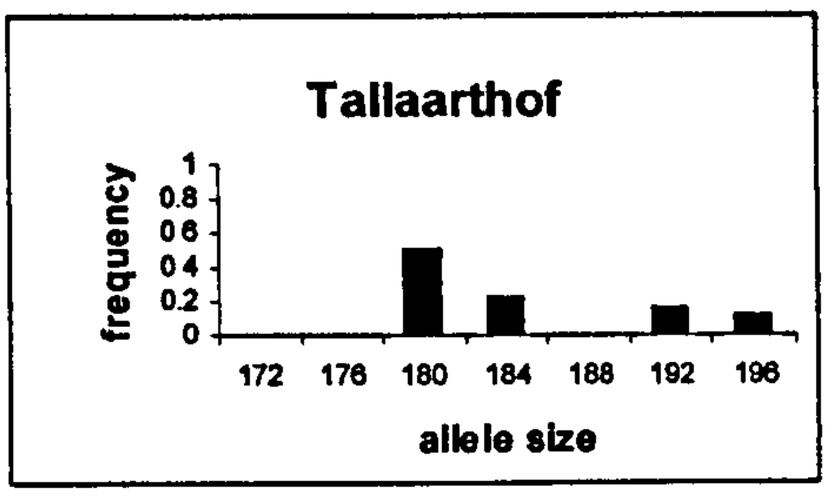
APPENDIX C: THE MICROSATELLITE ALLELE FREQUENCY DISTRIBUTIONS

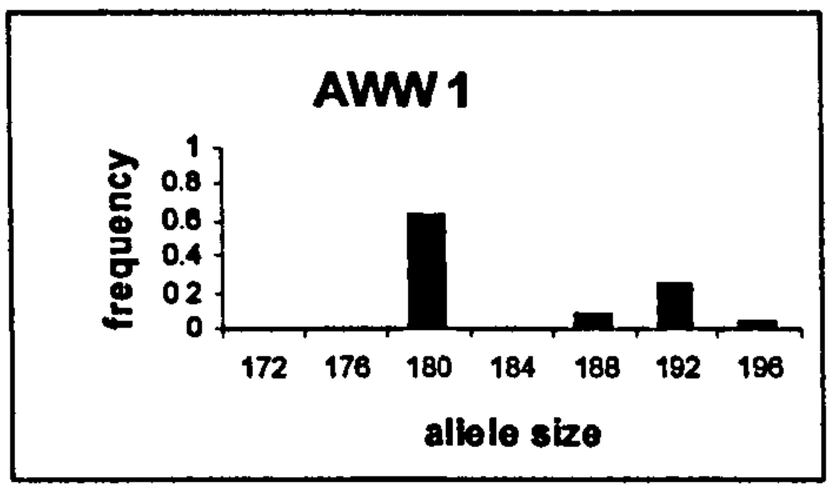


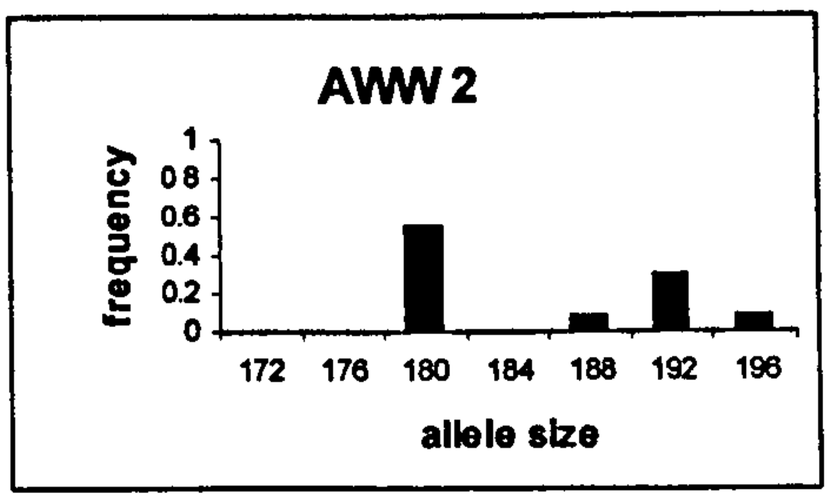


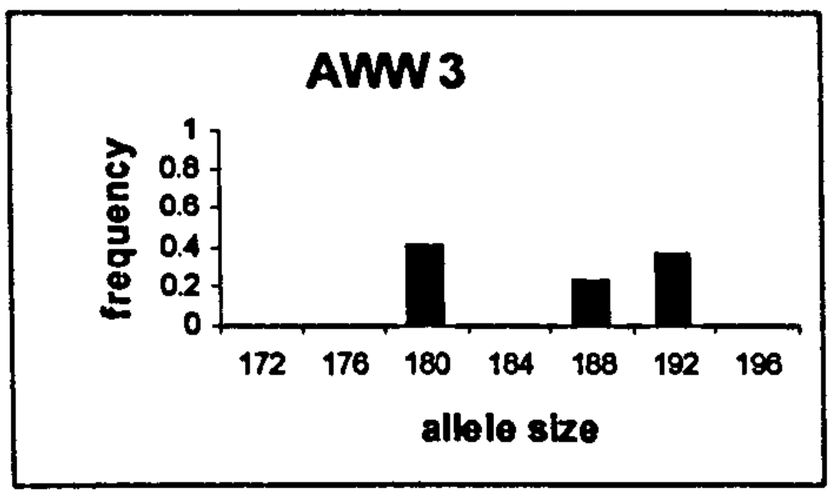


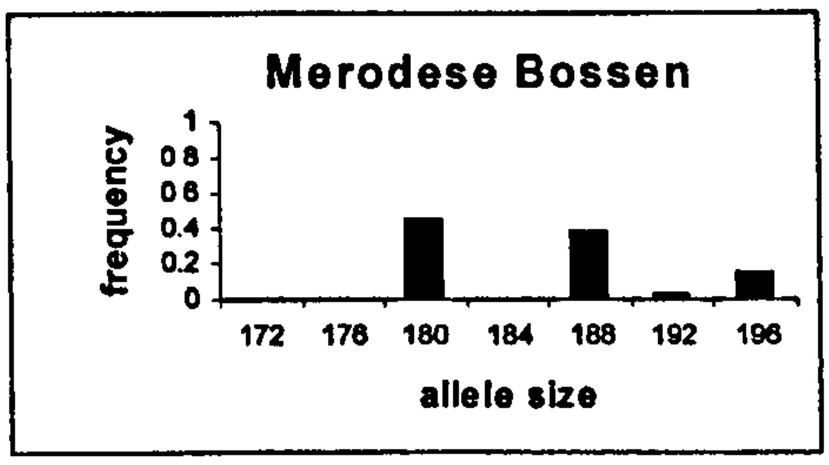


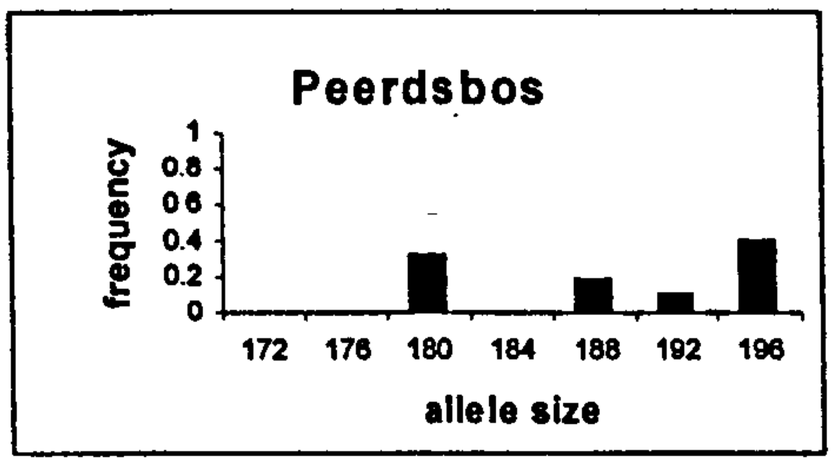


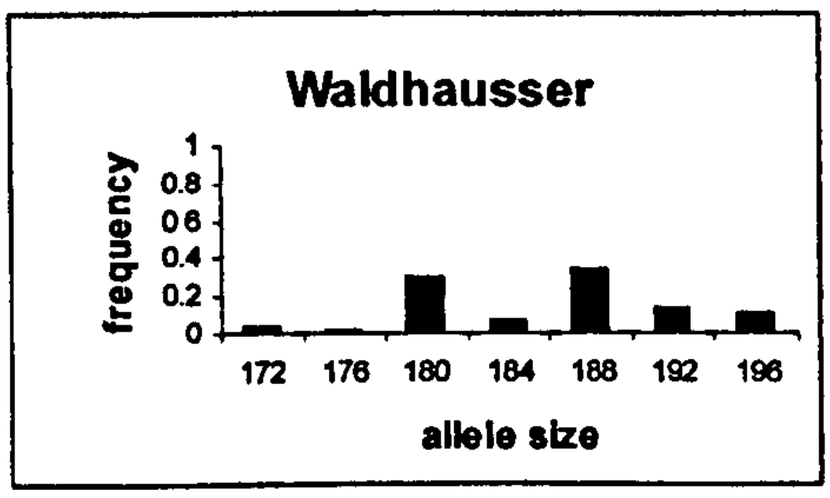




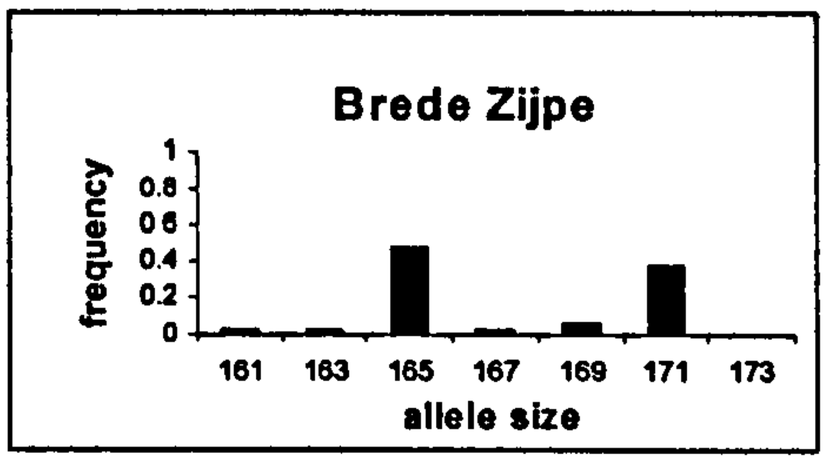


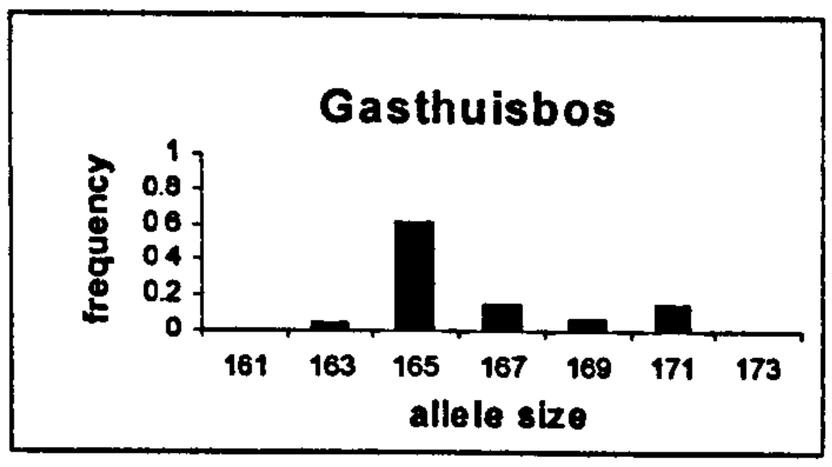


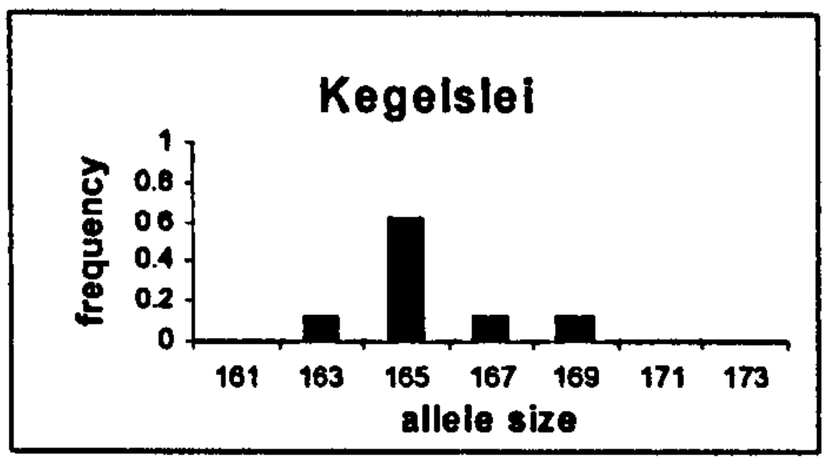


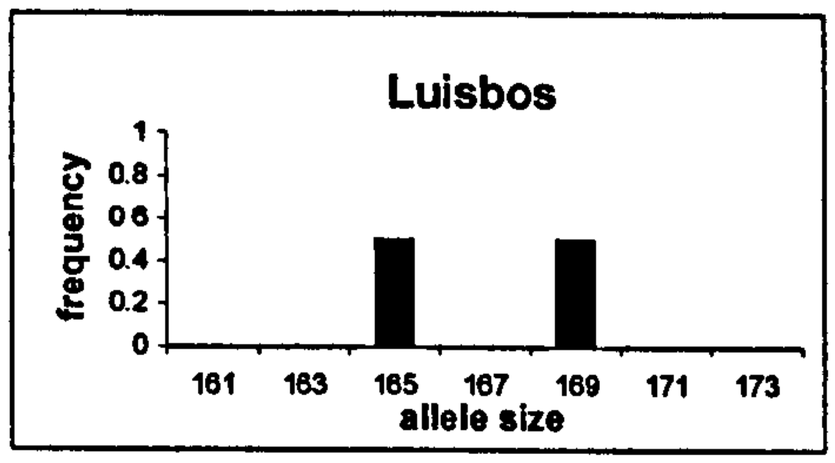


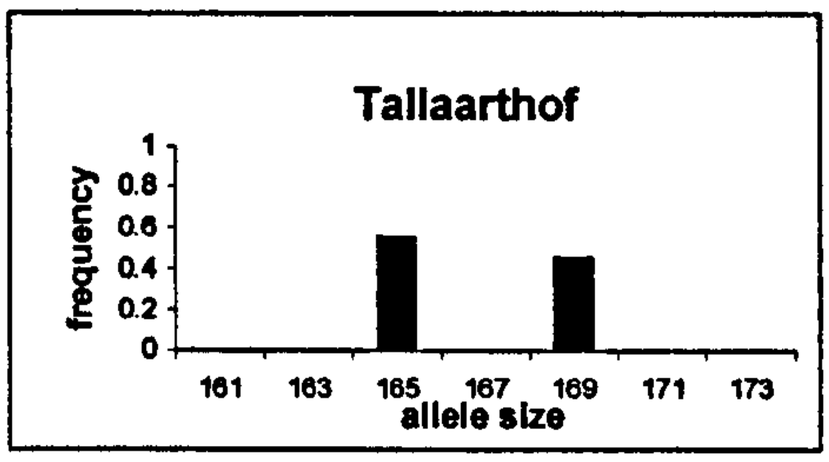
Appendix C.1: The allele frequency distributions of alleles found at locus RSμ1 in each population.

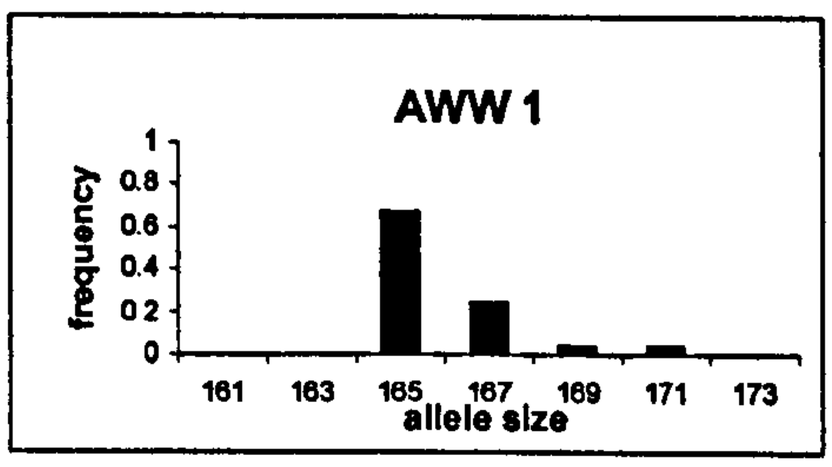


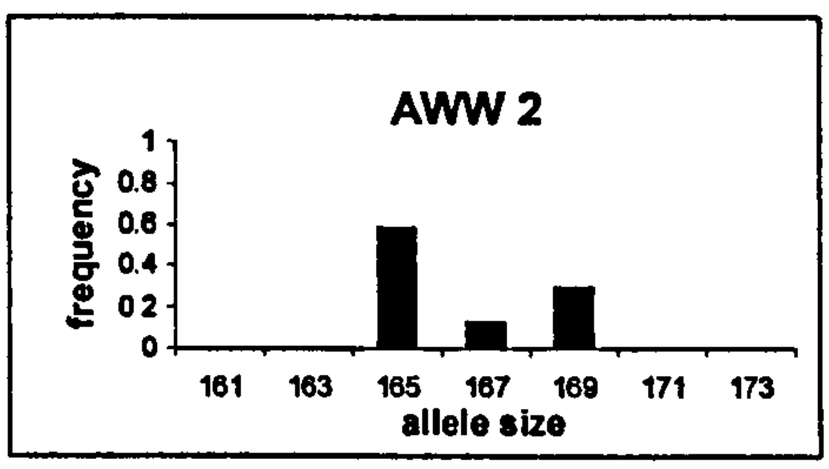


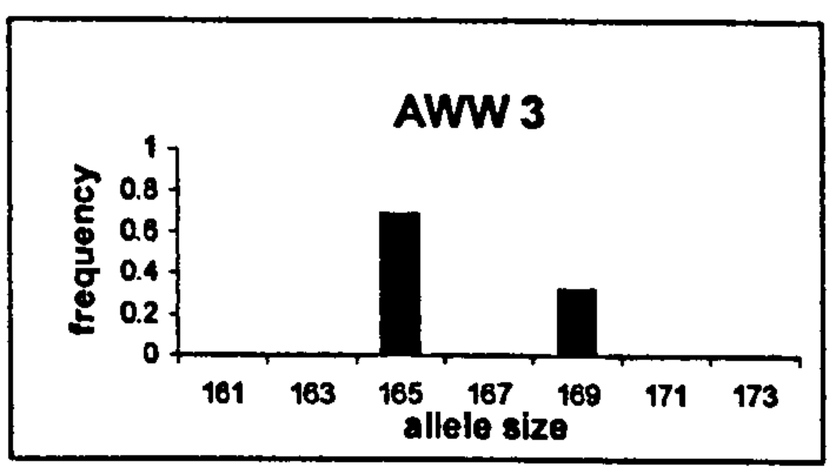


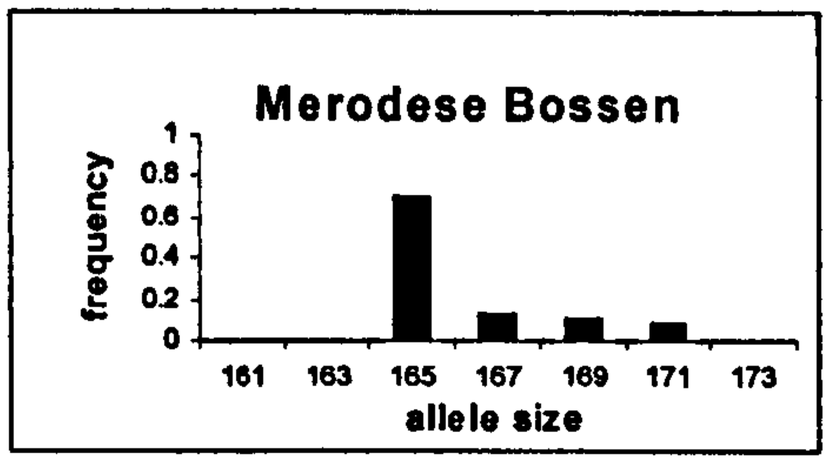


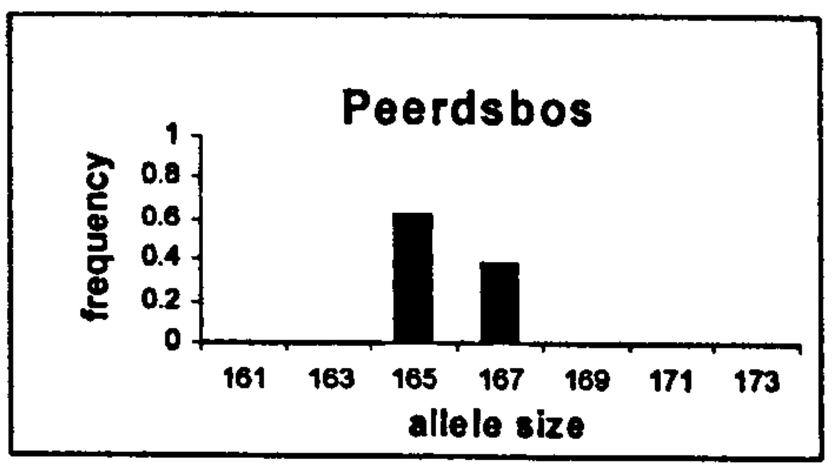


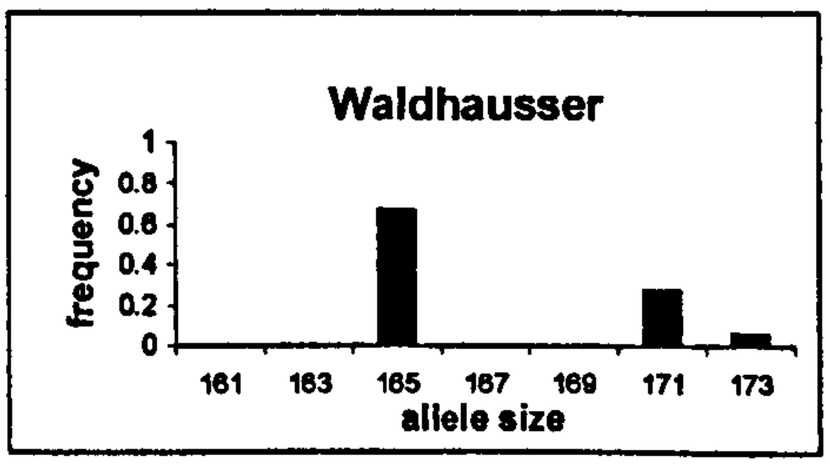




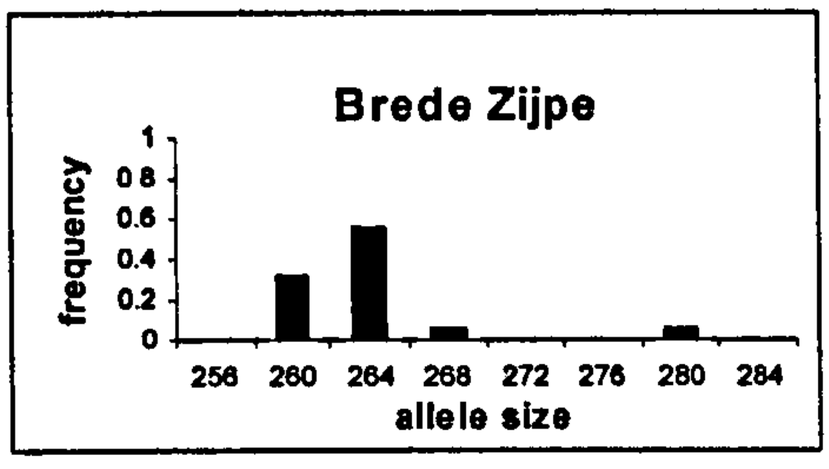


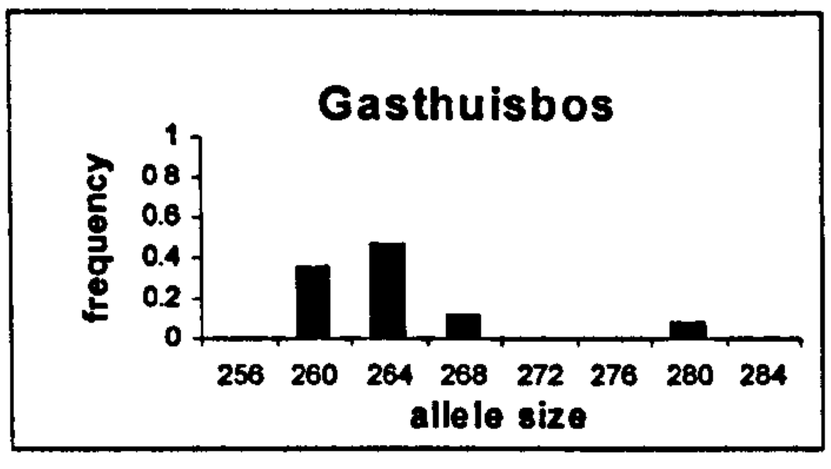


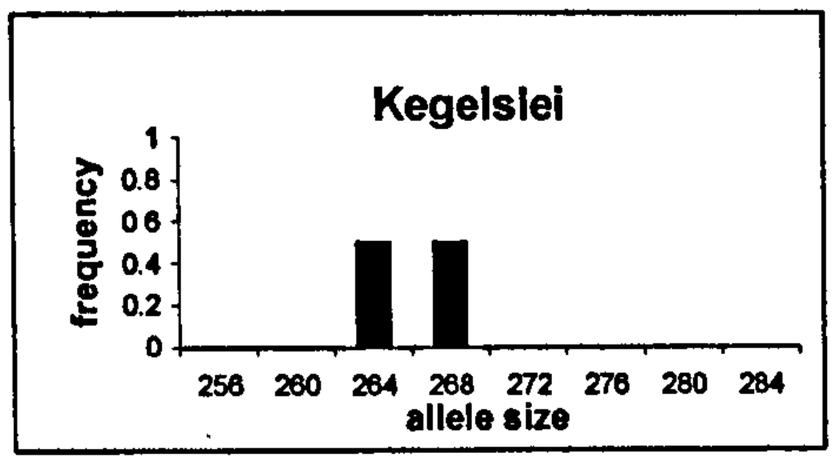


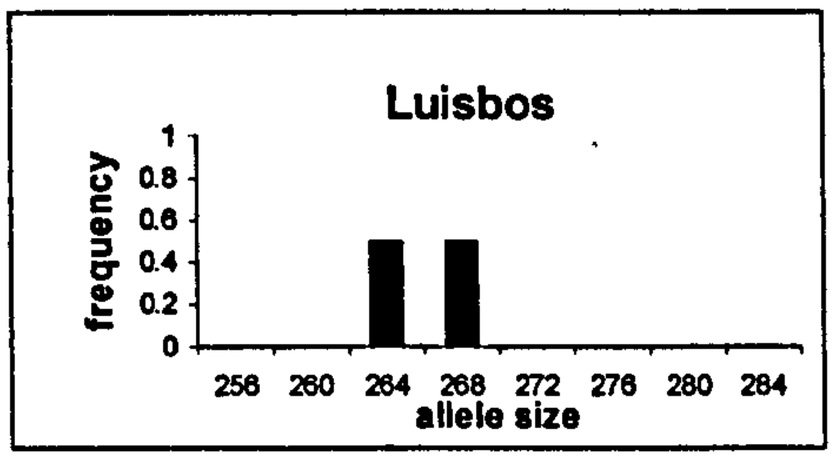


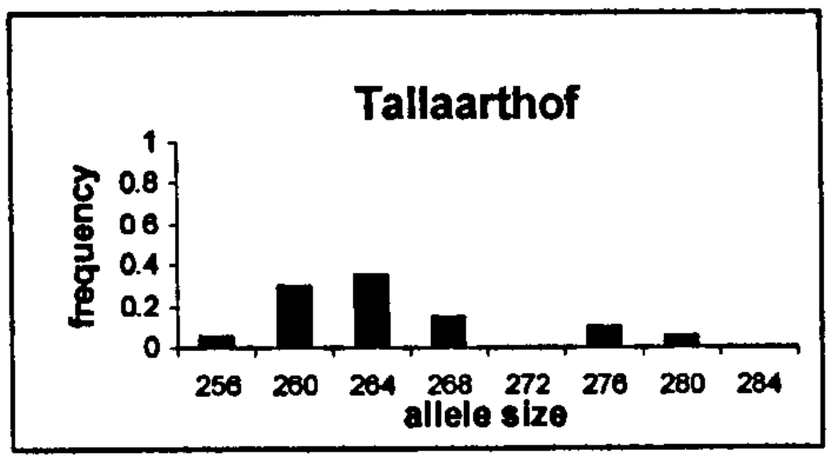
Appendix C.2: The allele frequency distributions of alleles found at locus RSμ3 in each population.

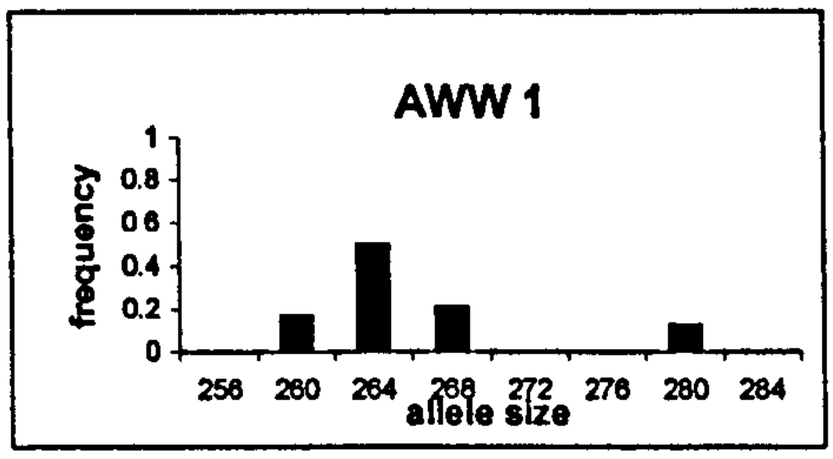


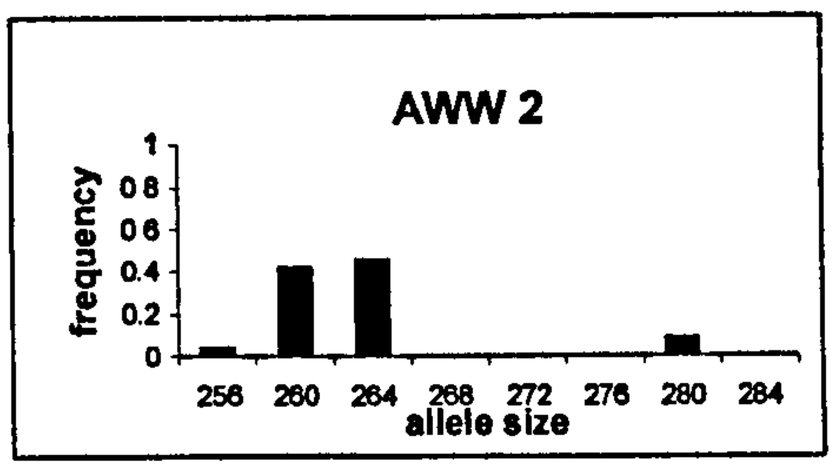


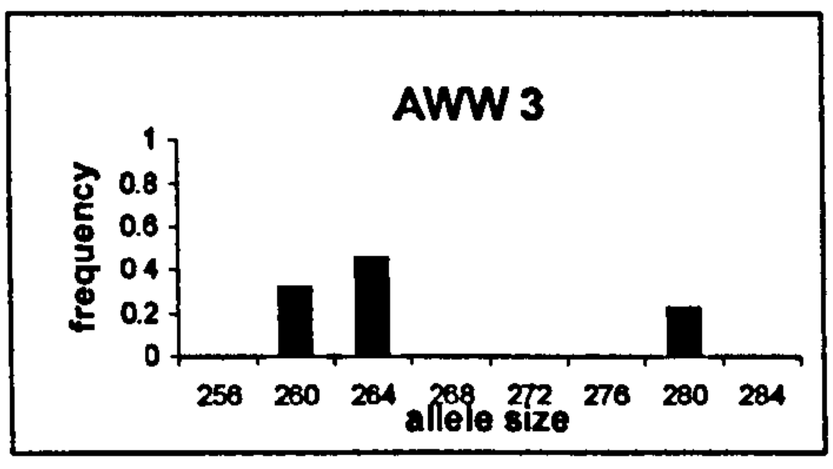


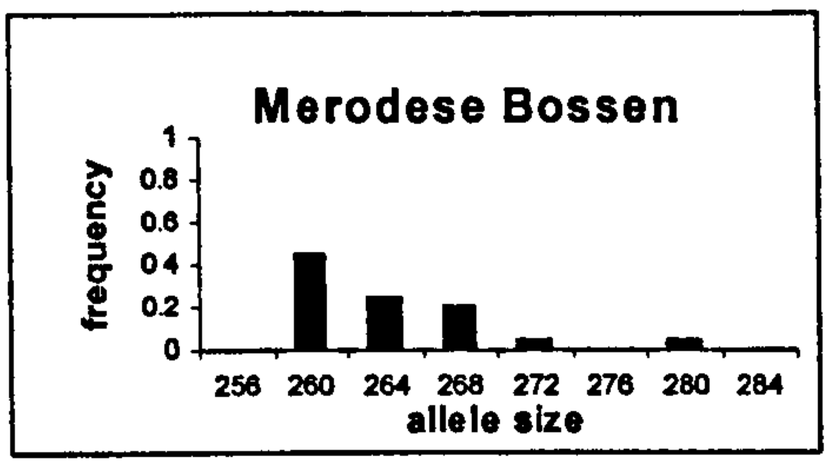


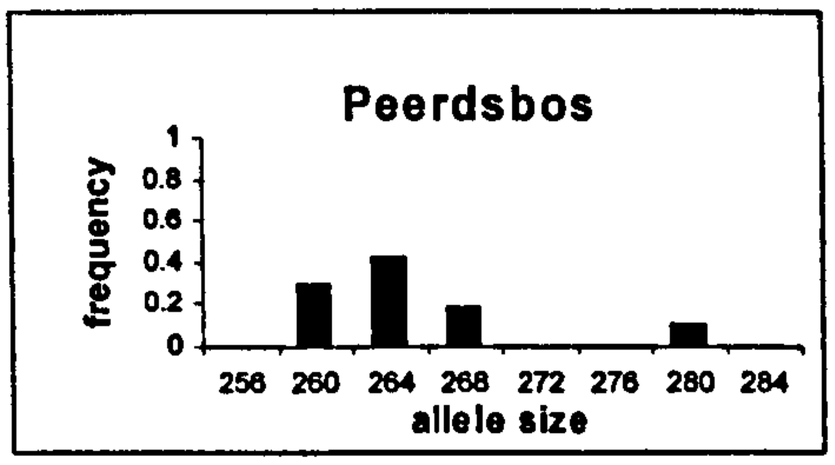


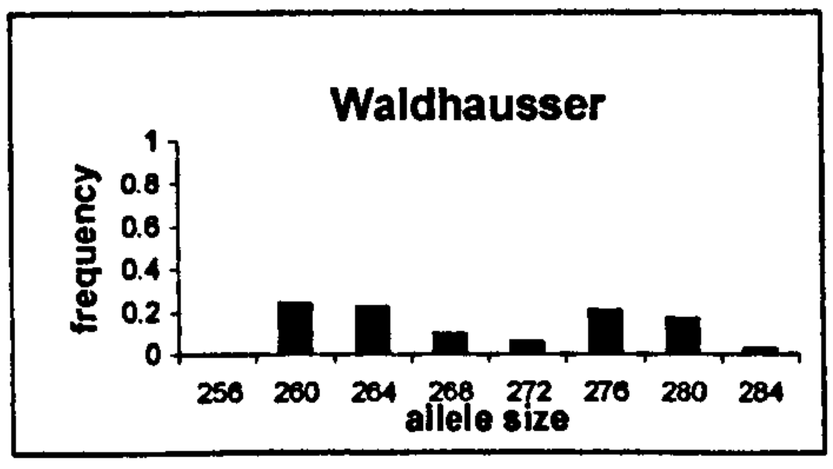




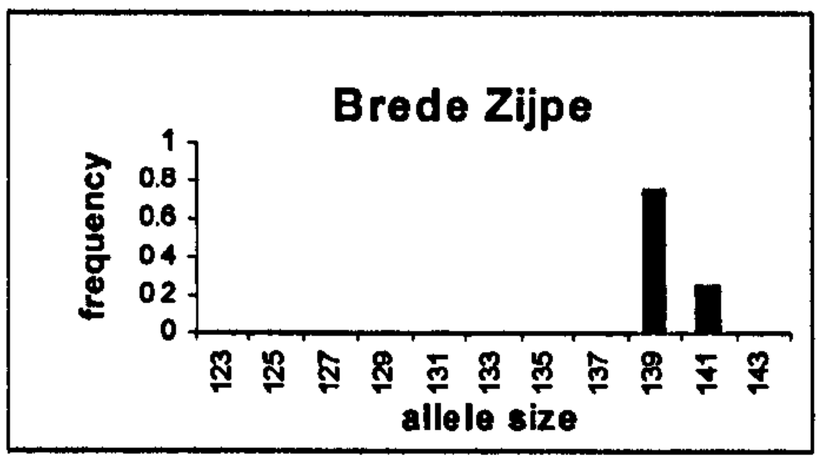


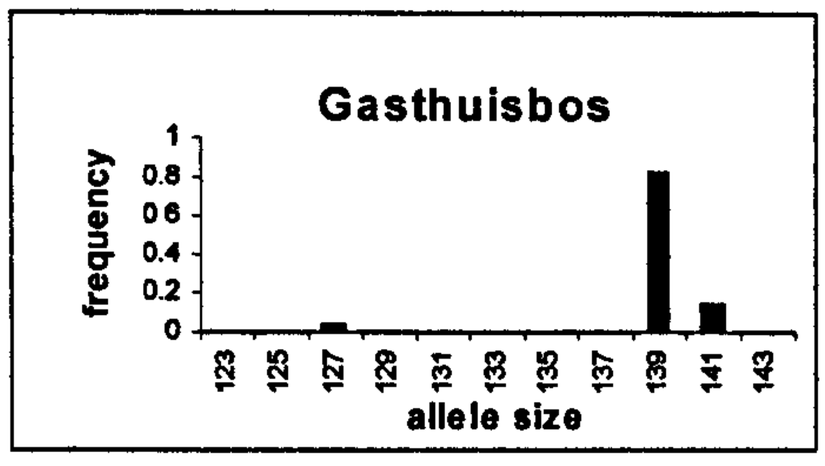


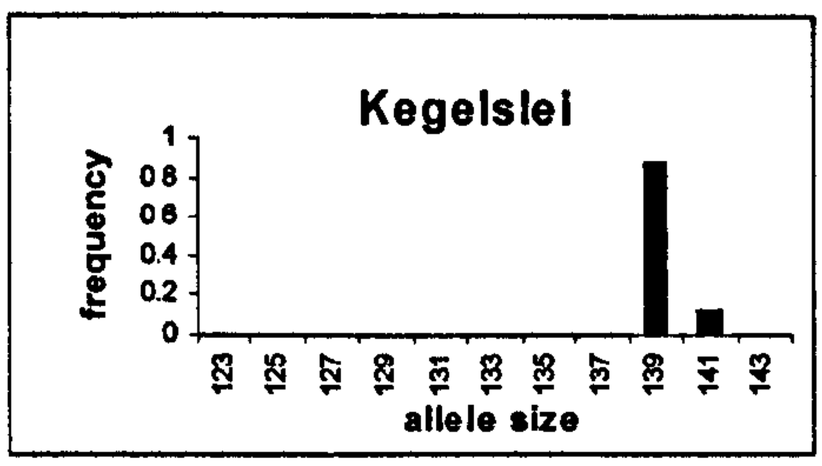


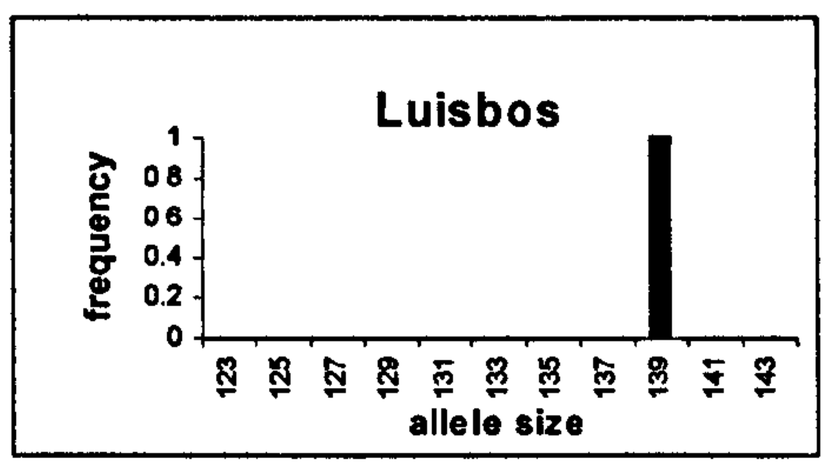


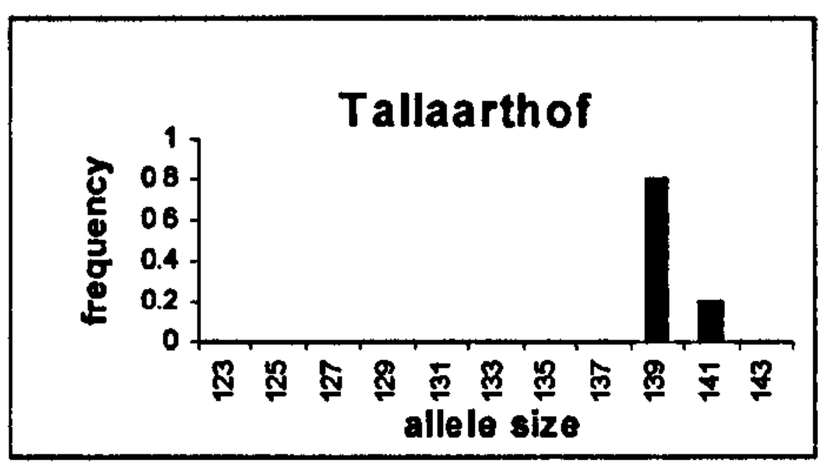
Appendix C.3: The allele frequency distributions of alleles found at locus RSμ4 in each population.

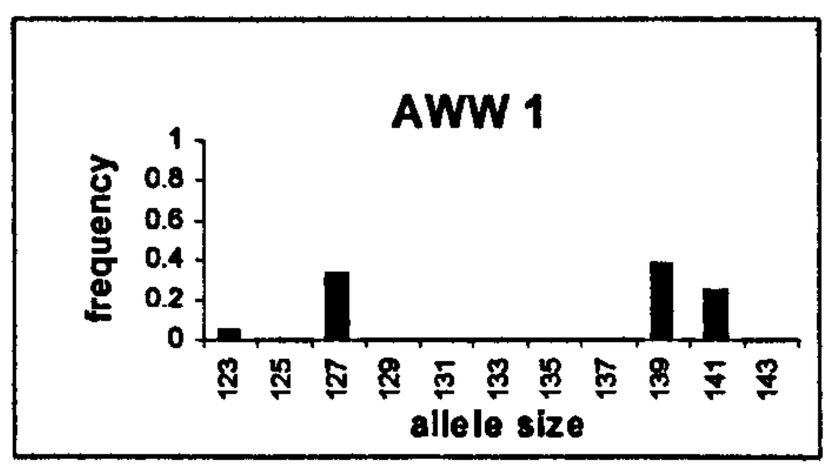


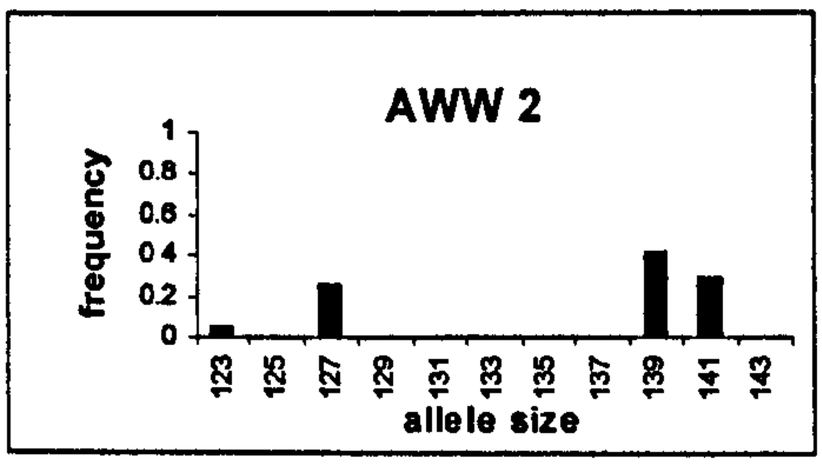


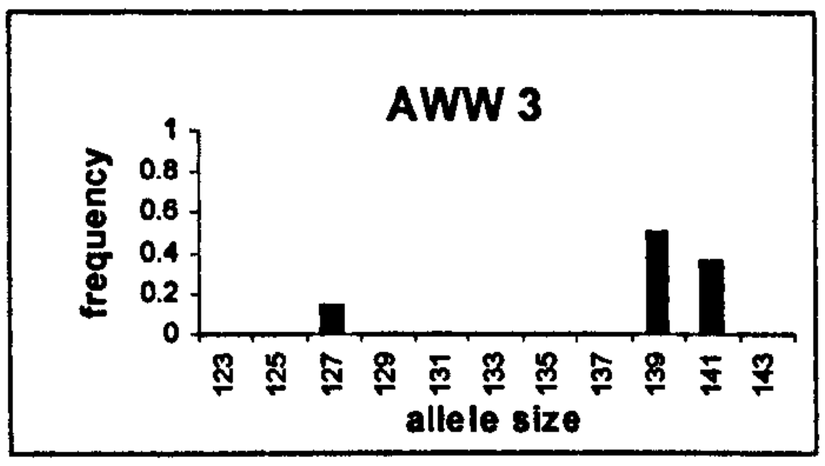


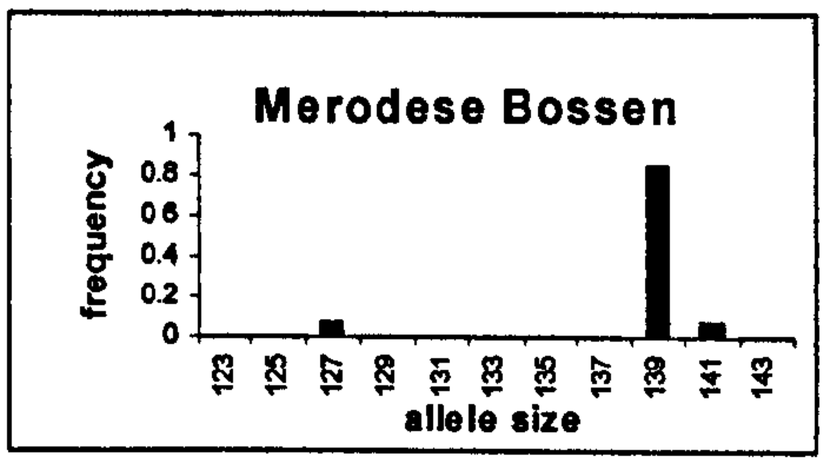


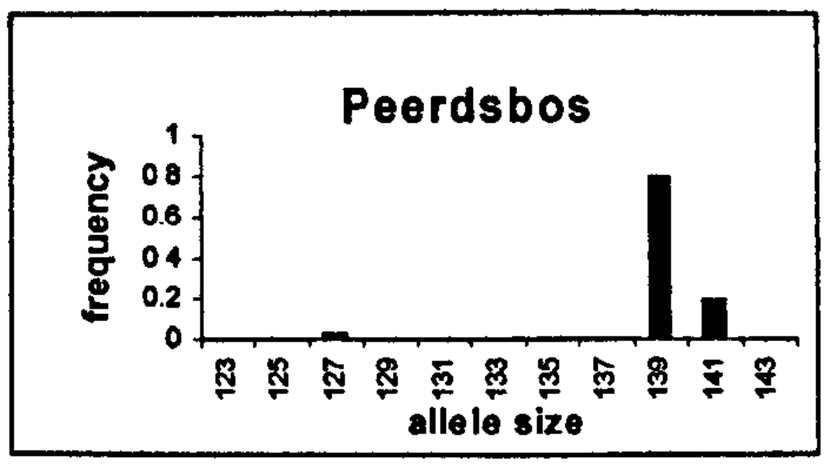


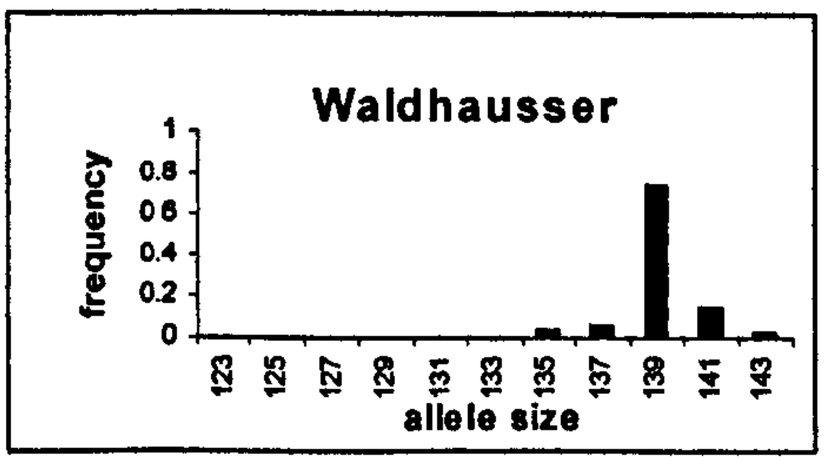




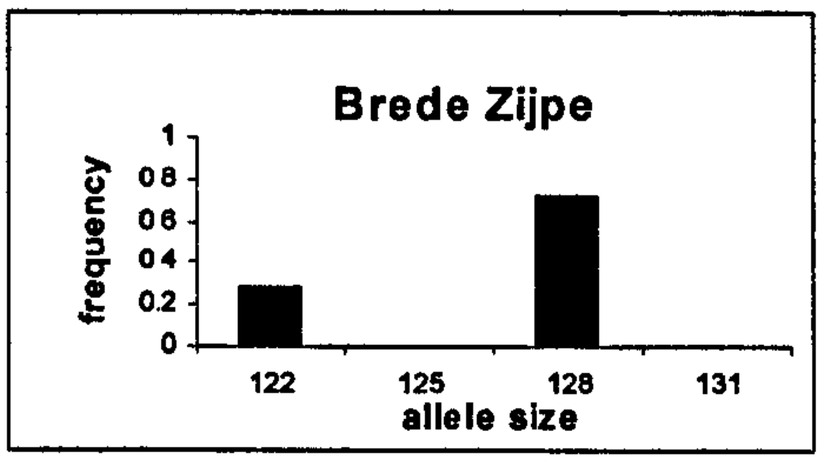


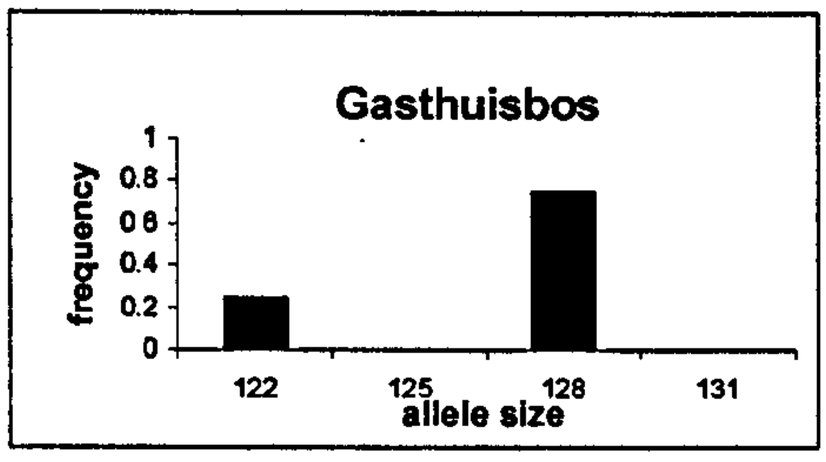


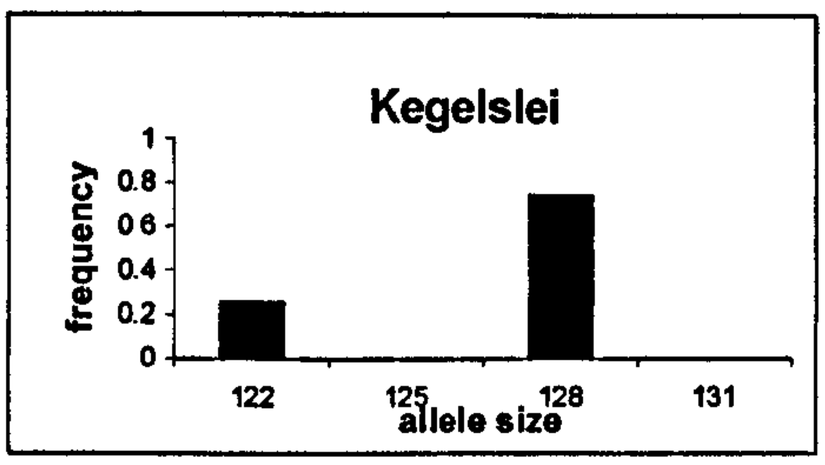


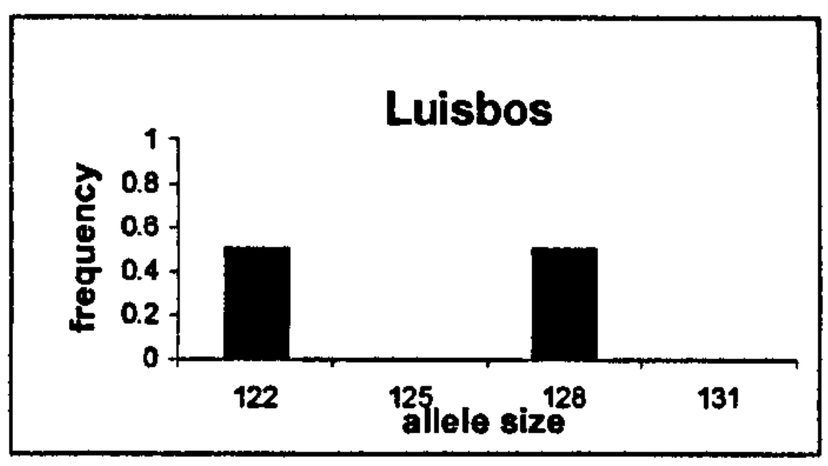


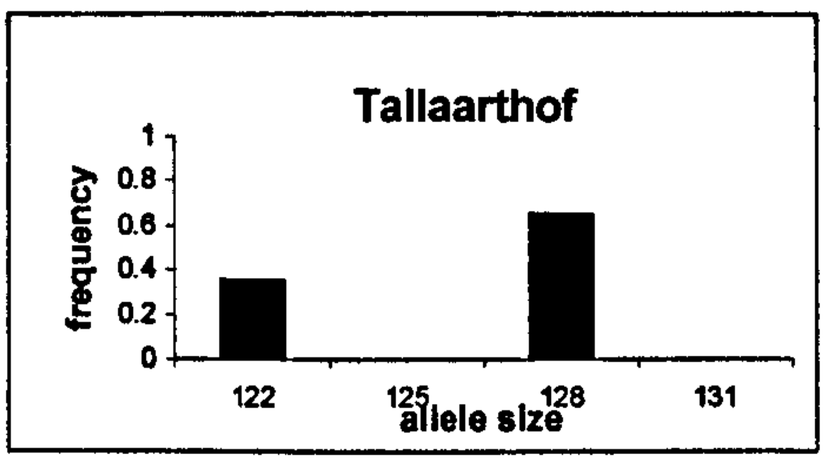
Appendix C.4 : The allele frequency distributions of alleles found at locus RSμ5 in each population.

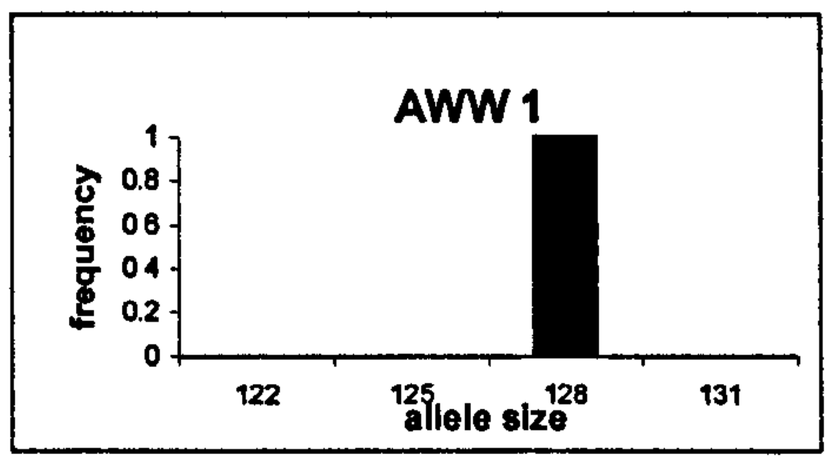


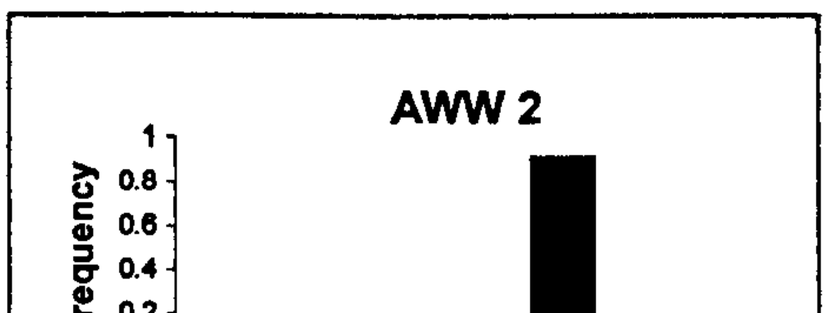


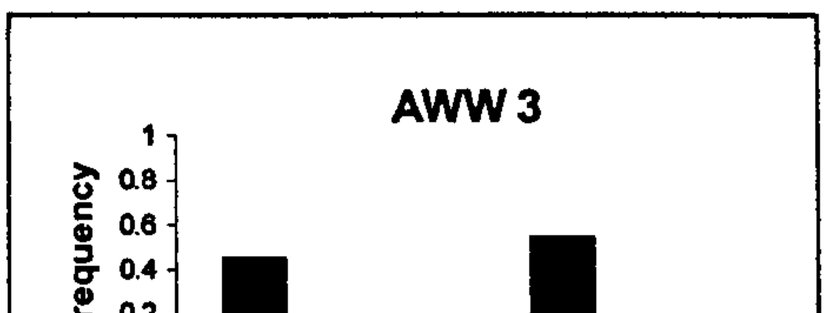


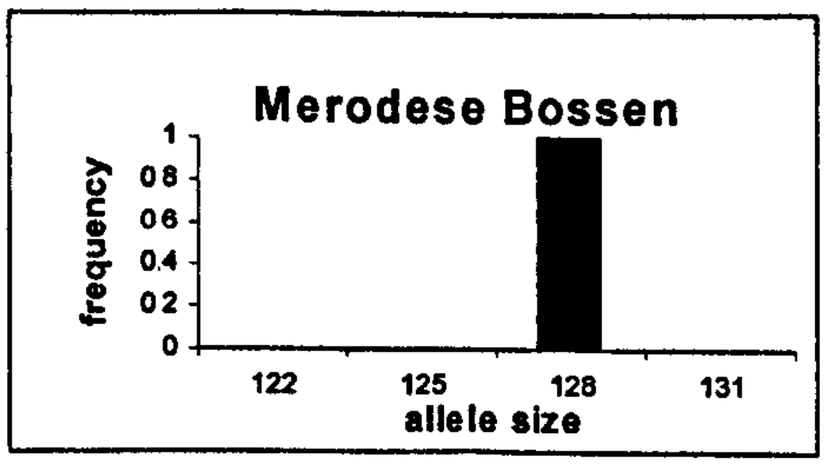


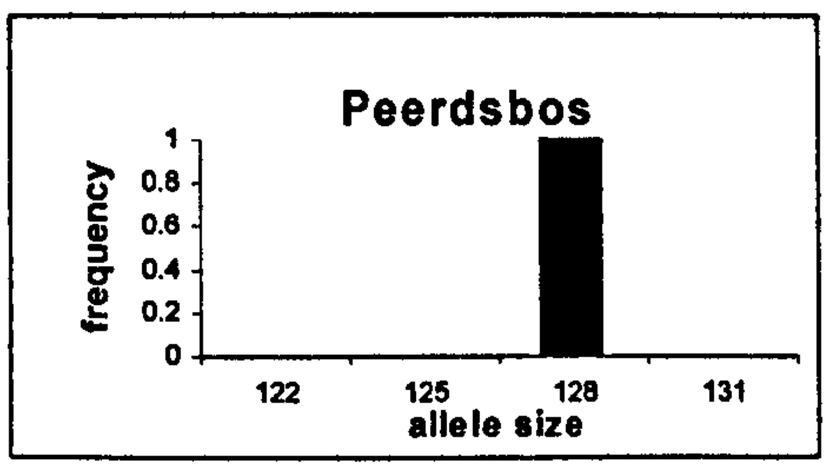


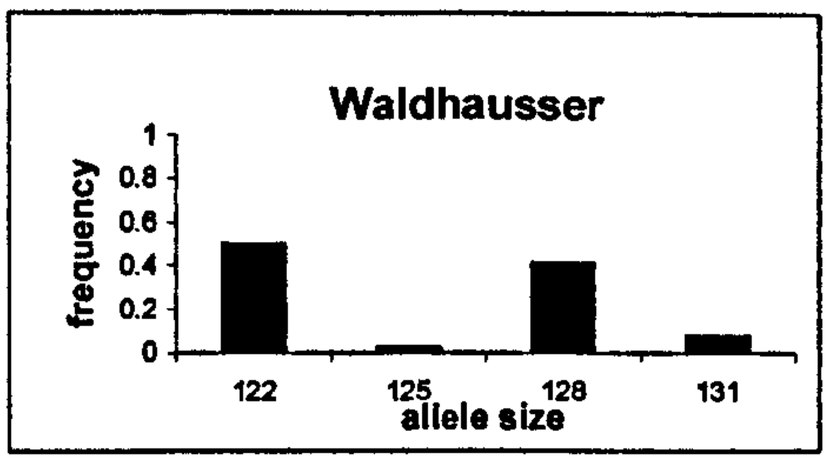












Appendix C.5: The allele frequency distributions of alleles found at locus RSμ6 in each population.